UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS
División de Estudios de Posgrado

Evolución Molecular Temprana de Rutas Biosintéticas

T E S I S

Que para obtener el Grado Académico de
DOCTOR EN CIENCIAS (BIOLOGIA)
P r e s e n t a

ANTONIO EUSEBIO LAZCANO ARAUJO

DIRECTORA DE TESIS:
DRA. MARIA DEL CARMEN GOMEZ EICHELMANN

Este trabajo está dedicado con afecto, gratitud, y profundo respeto intelectual, a mis amigos Carmen Gómez Eichelmann, Jesús Manuel León Cáceres, y Jorge Soberón Mainero

A. Lazcano-Araujo
Abril 8, 1996

RESUMEN

Aunque no conocemos como surgió la vida, la síntesis bajo condiciones prebióticas de moléculas orgánicas, la presencia de muchas de estas en meteoritos, y las propiedades catalíticas de diversas ribozimas, apoyan la idea de que los primeros organismos se formaron a partir de compuestos presentes en el medio ambiente primitivo, y de que durante etapas tempranas de la evolución biológica tanto la reproducción como el metabolismo de los seres vivos dependía de las propiedades catalíticas y replicativas del RNA. Debido a las limitaciones para la extrapolación de filogenias moleculares a épocas anteriores a la síntesis de proteínas, no se conocen ni los mecanismos que hayan podido conducir al llamado mundo del RNA, ni las rutas metabólicas que lo hayan podido sustentar. La situación es distinta cuando se pretende estudiar la evolución molecular de sistemas primitivos ya dotados de enzimas. En 1945 N. H. Horowitz propuso la llamada hipótesis retrógada, según la cual las vías biosintéticas se habían formado en sentido inverso al utilizar en forma gradual una serie de intermediarios metabólicos de origen abiótico presentes en los mares primitivos. Sin embargo, el análisis filogenético de las secuencias de las enzimas que participan en la biosíntesis de la histidina, por una parte, y de las que catalizan la formación de de la valina, la leucina, y la isoleucina, apoya no tanto la hipótesis retrógrada, sino la idea de que las rutas metabólicas originalmente estaban mediadas por unas cuantas enzimas de especificidad relativamente baja. En particular, la comparación de las secuencias de los genes que codifican para diversas enzimas de la biosíntesis de la histidina sugieren que esta ruta se estableció antes de la separación evolutiva de los tres linajes celulares, y que los fenómenos de duplicación y elongación génica jugaron un papel central en su ensamblaje. La hipótesis de que los fenómenos de amplificación génica jugaron un papel central en la evolución de los genomas celulares permite explicar, al menos en parte, la rapidez con la que parece haber tenido lugar la diversificación de los procariontes durante el Arqueano temprano.

NOTA PRELIMINAR

La parte central de esta tesis está compuesta por una serie de artículos, la mayoría de los cuales están ya publicados, y que han sido escritos en los tres últimos años como parte de un esfuerzo por comprender etapas muy tempranas de la evolución celular, anteriores a la diverrgencia de los tres linajes contemporáneos: arqueobacterias, eubacterias, y eucariontes. Esta es una fase aún por definir en la historia de la biósfera, y comienza, según algunos, con el origen del mundo del RNA, en el cual supuestamente los organismos dependían para su perpetuación y reproducción de de las propiedades catalíticas y replicativas de ribozimas de origen prebiotico. Sin embargo, como se discute en varios de los textos que se incluyen aquí, hoy se duda de que el mundo del RNA haya podido surgir directamente de la sopa primitiva, y es perfectamente posible que haya sido precedido a su vez de sistemas aún mas sencillos que dependían de polímeros cuya naturaleza se desconoce del todo. Nada es posible, por lo tanto, decir de los procesos metabólicos que pueden haber caracterizado estas etapas aún hipotéticas.

En cambio, y como se demuestra en el resto del material que se ha incluído en este trabajo, el análisis filogenético de las secuencias de las enzimas que participan en una serie de rutas metabólicas, especialmente las que llevan a la histidina y a los aminácidos hidrofóbicos alifáticos (valina, isoleucina, y leucina), ha permitido asomarse a etapas previas a la separación evolutiva de los tres linajes celulares. Por ello, este trabajo se orientó primero hacia la definición de un marco teórico dentro del cual se pudiera analizar la evolución de las rutas biosintéticas. Como se discute en el resto de los textos incluídos aquí, este análisis ha permitido no solo validar la idea de que las rutas anabólicas se ensamblaron merced al reclutamiento de enzimas ancestrales poco específicas, sino que también ha permitido ofrecer una explicación preliminar a la rapidez aparente con la que se establecieron los procesos metabólicos durante el Arqueano temprano. Esta segunda parte del trabajo llevado a cabo ha sido posible gracias a la colaboración con Stanley L. Miller (University of California, San Diego, EEUU), Joan Oró y George E. Fox (University of Houston, Houston, EEUU), Renato Fani (Università degli Studi di Firenze, Florencia, Italia).

En este tipo de estudios no se puede ser excesivamente ambicioso: a lo mas a lo que se puede aspirar es a la elaboración de una narrativa de caracter histórico lo mas coherente posible, y que sea congruente con las características de las rutas biosintéticas y las propiedades de los compuestos de importancia biquímica en los seres vivos contemporáneos. Ello explica el porqué algunas de las premisas e hipótesis discutidas aquí no han sido sometidas al análisis empirico, en su sentido ortodoxo, sino unicamente a las pruebas que se pueden derivar del análisis cladístico molecular.

Los textos incluídos en esta tesis son los siguientes:

1. Lazcano, A. (1994a) The RNA world, its predecessors and descendants. In S. Bengtson (ed), Early Life on Earth: Nobel Symposium No. 84 (Columbia University Press, New York), pp. 70-80

2 Lazcano, A. (1994b) The transition from non-living to living. In S. Bengtson (ed), Early Life on Earth: Nobel Symposium No. 84 (Columbia University Press, New York), pp. 60-69

3. Lazcano, A. (1993) The significance of ancient paralogous genes in the study of the early stages of microbial evolution. In R. Guerrero and C. Pedrós-Alió (eds), Trends in Microbial Society (Spanish Society for Microbiology, Barcelona), 559-562

4. Lazcano, A., Díaz-Villagómez, E., Mills, T., and Oró, J. (1995) On the levels of enzymatic substrate specificity: implications for the early evolution of metabolic pathways. Adv. Space Res. **15**: 345-356

5. Lazcano, A. (1995c) Cellular evolution during the early Archean: what happened between the progenote and the cenancestor? Microbiologia SEM **11**: 185-198

6. Keefe, A. D., Lazcano, A., and Miller, S. L. (1995) Evolution of the biosynthesis of the branched-chain amino acids. Origins of Life and Evol. Biosph. **25**: 99-110

7. Fani, R., Liò, P., and Lazcano, A. (1995) Molecular evolution of the histidine biosynthetic pathway. J. Mol. Evol. **41**: 760-774

8. Alifano, P., Fani, R., Liò, P., Lazcano, A., Bazzicalupo, M., Carlomagno, M. S., and Bruni, C. B. (1996) Histidine biosynthetic pathway and genes: structure, regulation, and evolution. Microbiol Rev. **(en prensa)**

9. Fani, R., Barberio, C., Casalone, E., Cavalieri, D., Lazcano, A., Lió, P., Mori, E., Perito, B., and Polsinelli, M. (1996) Paralogous histidine biosynthetic genes: evolutionary analysis of the Saccharomyces cerevisiae HIS6 and HIS7 genes. GENE **(enviado al Comité Editorial)**.

10. Mills, T., Fox, G. E., Fani, R., Leguina, I., Lazcano, A., and Oró, J. (1996) Molecular evolution of glutamine amidotransferases: implications for the origin of metabolic pathways. **(título tentativo; en preparación)**

11. Lazcano, A. and Miller, S. L. (1994) How long did it take for life to appear and evolve to cyanobacteria? J. Mol. Evol. **39**: 546-554

1. INTRODUCCION

1.1. La hipótesis retrógrada y el origen del metabolismo

Es fácil reconocer a la preocupación por comprender el origen y la evolución del metabolismo como uno de los principales impulsos intelectuales que llevaron a Oparin (1924) a formular sus ideas sobre la aparición de la vida (Lazcano, 1995a, b). Sin embargo, es igualmente cierto que aunque a partir de la década de los 60s se comenzó a experimentar con poblaciones de microorganismos para estudiar los mecanismos que permiten el desarrollo de rutas catábolicas y el uso de substratos artificiales (Mortlock, 1984; Hall y Hauer, 1993), salvo un número reducido de trabajos publicados durante el periodo comprendido entre 1945 y 1976, muy pocos se interesaron en la aparición misma de las vías anábolicas comunes a todos los organismos. La excepción mas importante la constituyó la llamada hipótesis retrógrada (Horowitz, 1945), según la cual las rutas biosintéticas se habían ensamblado hacia atrás, es decir, el eventual agotamiento por parte de los organismos primitivos de lo que ahora es el producto final de una ruta catabólica dada había llevado a la utilización de una molécula de estructura similar, y así sucesivamente hasta llegar a los precursores con los que se inicia cada proceso biosintético. La idea de Horowitz (1945), que estaba basada en las propuestas de Oparin (1924) sobre la llamada sopa primitiva y el caracter heterótrofo de los primeros seres vivos, suponía que los procesos prebiótica habían permitido la síntesis y acumulación en el medio ambiente primitivo de todas las moléculas que participan como intermediarios en las rutas biosintéticas contemporáneas. De acuerdo con este modelo, la duplicación de los genes que codificaban para las proteínas ancestrales había permitido la aparición de enzimas capaces de utilizar substratos con estructuras comparables hasta llegar al producto final.

Veinte años mas tarde, Horowitz (1965) afinó su idea inicial al proponer que el orden de los genes en un operon era precisamente el resultado evolutivo de la duplicación sucesiva de las secuencias genéticas que codificaban para las distintas enzimas que participan en una ruta biosintética dada. La hipótesis retrógrada se vió rapidamente transformada en la explicación

canónica de la aparición y la evolución temprana del metabolismo. No es difícil reconocer el origen de esta aceptación acrítica de las ideas de Horowitz. En primer lugar, la investigación sobre la aparición de la vida se concentró sobre todo en el estudio del origen de la replicación. En segundo término, la mayoría de los bioquímicos permaneció ajena durante muchos años ajenos a los problemas evolutivos. En tercer lugar, la mayoría de las rutas biosínteticas básicas parecen ser de origen monofilético, y hasta antes de la explosión de la información contenida en los bancos de secuencias, no existían muchas posibilidades de analizar su aparición bajo la óptica de la biología comparada y evolutiva.

Por otra parte, la hipótesis de Horowitz posee, en si misma, una serie de incentivos nada desdeñables: (a) como ya se dijo arriba, desde un punto de vista histórico constituye el primer intento para explicar el origen de las rutas biosintéticas; (b) permite establecer una continuidad entre el medio ambiente prebiótico y los procesos de propiamente biológicos, lo que le da un atractivo intelectual considerable; y (c) la comparación de las propiedades bioquímicas y, en algunos casos, de secuencias, ha permitido demostrar el origen común de algunas parejas de enzimas que catalizan pasos sucesivos en rutas biosintéticas, lo que algunos han interpretado como evidencia favorable a las ideas de Horowitz. Estos casos incluyen la homología entre (i) la ß-cistationasa y la cistation-γ-sintasa, que catalizan reacciones sucesivas en la biosíntesis de metionina en eubacterias (Belfaiza et al., 1986); (ii) la protoclorofila reductasa y la clorin-reductasa, que participan en la formación de la clorofila (Burke et al., 1993); y (iii) la fosforibosil formil-imino-5-amino imidazol-4-carboxamida ribotido reductasa y la amidociclasa, dos enzimas que catalizan pasos sucesivos en la biosíntesis de la histidina (Fani et al., 1994).

Sin embargo, es evidente que la demostración del origen monofilético de dos enzimas que participen en pasos sucesivos en una ruta biosintética no es suficiente para probar la validez de las ideas de Horowitz (1945, 1965). Es posible, por ejemplo, que los pasos catalizados sean químicamente equivalentes, y que el origen común de las enzimas en cuestión sea resultado de una duplicación que permitió el aumento de la especificidad de los catalizadores (Fani et al., 1995). De hecho, es posible oponer al modelo de

Horowitz una serie de contraargumentos (Lazcano et al., 1992), que incluyen (a) la inestabilidad intrinsíca de muchos intermediarios metabólicos, cuya síntesis y acumulación en la tierra primitiva es difícil de comprender en términos estrictamente químicos (Cánovas et al., 1967; Ornston, 1971; Jensen, 1976); (b) muchos intermediarios metabólicos son compuestos fosforilados que difícilmente podrían atravesar las membranas primitivas (Ornston, 1971) en ausencia de mecanismos de transporte especializados (Jensen, 1976); (c) la hipótesis de Horowitz supone que todos los cambios químicos sufridos por los intermediarios en una vía metabólica dada son iguales o equivalentes, lo que sabemos que no necesariamente es cierto; (d) en sentido estricto, la aplicación de las ideas de Horowitz al problema del origen del DNA parecería implicar que los desoxirribonucléotidos precedieron a los ribonucleótidos, pero existen muchas evidencias en contra (Lazcano et al., 1988); y (e) la hipótesis retrógrada no explica el origen de los mecanismos de regulación presentes vías catabólicas o anabólicas.

A pesar de estas limitaciones, existe una variante de la hipótesis de Horowitz que apenas comienza a ser explorada. Como se mencionó arriba, la presencia de intermediarios metabólicos en el medio ambiente primitivo parece ser poco probable; sin embargo, la síntesis prebiótica de un compuesto orgánico de importancia biológica permitiría la acumulación de los productos de su descomposición. Un caso evidente lo constituyen, por ejemplo, la acumulación de poliaminas que resultarían de la descarboxilación de aminácidos de origen prebiótico, o la presencia de uracilo como resultado de la desaminación oxidativa de la citosina. A partir de esta premisa, se ha propuesto que la biosíntesis de la valina, la leucina, y la isoleucina, que no puede ser explicada con la aplicación estricta de la hipótesis retrógrada, puede haber tenido su origen en la carboxilación reductiva de ácidos grasos de cadena corta, seguida de una transaminación no enzimática (Keefe et al (1995). Este intento por desarrollar una variante de la hipótesis retrógrada no fue aceptado por Horowitz (comunicación personal), quien prefiere seguir creyendo en la presencia de intermediarios metabólicos en la sopa primitiva y en la aplicación directa de su idea tal como la sugirió en 1945.

1.2 La duplicación génica y la evolución de rutas biosintéticas

Pocos años después de que Horowitz propusiera su esquema, Lewis (1951) sugirió que la presencia de enzimas con propiedades comparables pero que participaran en rutas metabólicas diferentes podría ser explicado como resultado de duplicación y divergencia de secuencias génicas. Sin embargo, ni Lewis discutió su idea en el contexto de la evolución metabólica temprana, ni la comunidad interesada en el estudio del origen de la vida se percató del significado de esta hipótesis. Esta misma idea fue propuesta y desarrollada por mas tarde por Waley (1969), Ycas (1974) y Jensen (1976), quienes sugirieron en forma independiente que las rutas biosintéticas se habían ensamblado a partir de enzimas poco específicas, que catalizaban no tan solo una reacción, sino conjuntos de reacciones similares en las que intervenían substratos de naturaleza química comparable.

El propio Ycas (1974) hizo notar que la suposición de que las enzimas primordiales eran poco específicas permitía obviar el problema de las capacidades codificadoras reducidas que deben haber poseído los genomas primitivos. Es decir, aunque las células ancestrales codificaran tan solo para unas cuantas enzimas, la escasa especificidad de estas les dotaba de un conjunto de habilidades metabólicas no necesariamente desdeñables. De acuerdo al modelo de Waley-Ycas-Jensen, la especificidad enzimática contemporánea es resultado de la duplicación y divergencia de unos cuantos genes ancestrales. Esta hipótesis parece chocar contra la idea generalizada de que las enzimas son extraordinariamente selectivas. Es posible, sin embargo, mostrar lo contrario. A veces basta substituir un cofactor metálico, por ejemplo, para lograr que una DNA polimerasa se comporte como una reverso transcriptasa, o vice versa (Lazcano et al., 1994).

La primera en darse cuente de que este modelo es comparable a las labores de retacería fue Clarke (1983), quien lo bautizó como la "patchwork hypothesis" --equivalente al "bricollage" de los franceses. Es decir, al igual que ocurre con una colcha de retacería, en donde el conjunto final está formado por componentes de distinta procedencia, las distintas enzimas que forman una via metabólica dada tienen orígenes diversos. Ello implica, por supuesto, que la validez de esta hipótesis se puede demostrar con evidencias

del origen común de diversas enzimas que participen en distintas rutas metabólicas, o de enzimas homólogas que catalicen pasos distintos en la misma vía biosintéticas (Fani et al., 1995).

En los últimos diez años se ha acumulado una buena cantidad de evidencias que apoyan de manera directa e indirecta la hipótesis de Waley-Ycas-Jensen. Es evidente, por ejemplo, que el descubrimiento de que parece existir un número relativamente reducido de familias de proteínas, o la posibilidad de alinear las secuencias de deshidrogenasas o de glutaminamidotransferasas, apoya fuertemente la idea de que en epocas ancestrales las diversas reacciones bioquímicas en las que hoy intervienen estas enzimas eran catalizadas por proteínas menos específicas. Esta hipótesis se ve reforzada por la confirmación del papel que la duplicación de genes ha jugado en el origen de cuando menos una tercera parte de los genomas de Escherichia coli (Labedan y Riley, 1995) y de Haemophilus influenza (Brenner et al., 1995) y, por supuesto, con la demostración del papel que la duplicación y elongación de genes han tenido en la evolución molecular de la biosíntesis de la histidina (Fani et al., 1995). Resta, sin embargo, no solo el extender este tipo de análisis a otras rutas anábolicas ancestrales, sino también a la comprensión del origen de secuencias aisladas que no parecen formar parte de familia alguna, bien sea por que en efecto hayan surgido en forma independiente, o bien porque han divergido tanto que no es posible detectar su homológia con otras proteínas.

1.3 El mundo del RNA

Aunque la idea de que el RNA hubiera precedido al DNA como material genético comenzó a ser aceptada desde hace unos cuarenta años (Lazcano et al., 1988), la posibilidad de que los primeros organismos carecieran también de proteínas y que, por lo tanto, dependieran unicamente de moléculas de RNA, no dejó de ser vista como una mera especulación sino hasta el descubrimiento de las ribozimas. La caracterización de las propiedades catalíticas de estas últimas moléculas llevó a la formulación del llamado mundo del RNA, entendido como una fase en la evolución en la que las células primitivas eran poco mas que pequeños liposomas conteniendo cuyo

metabolismo y reproducción dependía de ribozimas capaces de replicarse (Lazcano, 1994a).

A pesar del valor heurístico que tiene la hipótesis del mundo del RNA, estamos lejos de saber cual fue su origen. Ello es debido no solo a la ausencia de ribozimas autoreplicativas, sino también a (a) la falta de modelos capaces de explicar la síntesis y acumulación en la Tierra primitiva de los diversos componentes del RNA, incluyendo la ribosa misma (Larralde et al., 1995); (b) la inhibición que sufren las reacciones no-enzimáticas de polimerización molde-dependientes en mezclas racémicas de nucleótidos activados (Joyce et al., 1987); y (c) la ausencia de mecanismos prebióticos que expliquen la presencia de ésteres de fosfatos (Keefe y Miller, 1995).

Las dificultades que enfrenta la síntesis prebiótica del RNA ha llevado a postular la existencia de sistemas genéticos mas antiguos que el RNA mismo, lo que ahora se conocen como los mundos de pre-RNA (Joyce et al., 1987). Esta hipótesis supone que estos se caracterizaban por la existencia de polímeros genéticos que reunían, al igual que el RNA, las propiedades de fenotipo y genotipo en un solo tipo de molécula. Sin embargo, estas macromoléculas ancestrales pueden haber carecido no solo del esqueleto fosfodiestérico del RNA, sino tal vez hasta de las purinas y pirimidinas que hoy endía forman parte de los ácidos nucleicos (Miller y Lazcano, en prep.). El problema fundamental de la caracterización de estos sistemas biológicos ancestrales radica, por supuesto, en su naturaleza por ahora totalmente hipotética. Postular la existencia de sistemas biológicos mas antiguos que el RNA implica, por supuesto, que el origen de este último se encuentra no en procesos meramente químicos, sino de naturaleza propiamente biológica. Es decir, la suposición de que hubo mundos de pre-RNA hace necesario el proponer la existencia de fuentes de energía y rutas metabólicas ancestrales capaces de síntetizar no solo ribosa (Larralde et al., 1995), sino también nucleósidos y nucleótidos, a partir de los cuales el RNA se formaría y eventualmente pudiera substituir a sus progenitores. Huelga decir que nada sabemos de estos procesos hipotéticos.

1.4 ¿Qué sabemos del último ancestro común que tuvieron los tres linajes celulares?

El análisis filogenético de las secuencias del 16/18S rRNA no solamente permitió a Woese y Fox (1977) caracterizar a las arqueobacterias como una rama monofilética claramente definida, sino que también mostró que todos los organismos conocidos forman parte de un árbol en el cual es posible distinguir otros dos grandes linajes, formados por las eubacterias y el nucleocitoplasma eucarionte. Este árbol evolutivo, que resultó de la comparación de las secuencias de genes ortólogos del rRNA, se trifurca a partir de un ancestro común, al que Woese y Fox (1977) denominaron progenote. Debido a que no se ha descubierto ningún organismo que pueda servir como grupo externo a los tres grandes linajes celulares, el progenote fue definido no solamente como el ancestro común a las eubacterias, las arqueobacterias, y los eucariontes, sino también como una entidad hipotética primitiva en la que la separación de fenotipo y genotipo aún no había tenido lugar (Woese y Fox, 1977). Eventualmente Woese (1983, 1987) supuso que el progenote era un sistema en donde el material hereditario estaba constituído por moléculas fragmentadas de RNA que no estaban integradas en un solo polímero genético (Woese, 1983, 1987).

No todos aceptaron la posibilidad de que el último ancestro común fuese, en efecto, un progenote. A partir del análisis de las secuencias de tRNAs de los tres linajes celulares, Fitch y Upper (1987) sugirieron que el ancestro común a estos ya poseía un código genético equivalente al de las células contemporánea, y propusieron que el árbol del rRNA se trifurcaba no a partir de un progenote sino de un organismo complejo a la que denominaron cenancestro. Por otra parte, la comparación de secuencias homológas en los tres linajes permitió proponer que el llamado progenote era, en realidad, una célula procarionte dotada de los mismos rasgos de la biología molecular y las habilidades metabólicas que cualquier bacteria contemporánea (Lazcano et al., 1992; Lazcano, 1995c). A pesar de que no se puede excluir del todo la posibilidad de que hayan ocurrido fenómenos de transporte horizontal entre las eubacterias, las arqueobacterias y los eucariontes, la caracterización del ancestro común a estos tres linajes como una célula procarionte

compleja sugiere la existencia de una fase de evolución biológica previa a su trifurcación.

Como es sabido, no fue sino hasta que se utilizaron conjuntos de genes parálogos que se habían duplicado antes de la separación de los tres linajes, cuando se pudo comenzar a construir árboles universales con raíz, la cual se ha ubicado en la rama eubacteriana (Gogarten et al., 1989; Iwabe et al., 1989). Aunque la idea de que las eubacterias corresponden al fenotipo mas antiguo de todas las formas actuales de vida ha ido ganando una aceptación creciente (Brown y Doolittle, 1995; Doolittle, 1995), es igualmente cierto que existen anomalías aún no explicadas, entre las que se incluyen las filogenias construídas con secuencias de glutamato deshidrogenasas, glutamino sintetasas (Forterre et al., 1993), carbamoyl-fosfato sintetasas (Lazcano, Puente, y Gogarten, en prep), proteínas de choque térmico, y otras mas. Ello ha llevado a sugerir que en el pasado pudo haber ocurrido un transporte masivo de genes entre los ancestros de las bacterias Gram positivas y las arqueobacterias (Gogarten, 1994).

Algo que pasó desapercibido para muchos, sin embargo, fue el hecho de que las duplicaciones parálogas ancestrales nos permiten asomarnos a épocas muy tempranas de la evolución y describir, así sea en forma parcial, sistemas biológicos mas simples que el cenancestro mismo (Lazcano, 1993, 1994b). Por ejemplo, la presencia de dos conjuntos parálogos de factores de elongación (Iwabe et al., 1989) y del caracter homólogo de las subunidades α y β de las ATPasas tipo F en los tres linajes (Gogarten et al., 1989), por ejemplo, permite reconocer una fase ancestral en la que la síntesis de proteínas requería de un solo factor de elongación, y las ATPasas con habilidades reguladoras limitadas (Lazcano, 1993, 1994b). Este tipo de análisis, que se ha podido ir extendiendo a los conjuntos de genes parálogos que codifican a las DNA polimerasas, las aminoacil-tRNA sintetasas, las glutamin-amido transferasas, y otras enzimas mas, ha permitido reconocer una etapa temprana de la evolución biológica que no había sido caracterizada previamente. Así, aunque aún estamos lejos de comprender los eventos que tuvieron lugar entre el mundo del RNA y el cenancestro, el análisis de los genes parálogos comunes a los tres linajes permite no solamente detectar otros candidatos para enraizar árboles universales, sino también para

describir, así sea de manera parcial, la evolución de la especificidad enzimática y la forma en que se fueron ensamblando las rutas metabólicas antes de la separación de eubacterias, arqueobacterias y eucariontes (Fani et al., 1995)

1.5. Los ritmos y tasas de la evolución en el Arqueano

Aunque existe un número limitado de técnicas, como los valores de fraccionamiento isotópico, que podrían ser utilizadas para datar la aparición de ciertos metabolismos, no es fácil utilizar la información del registro fósil para establecer las fechas de divergencia de los principales grupos procariontes. De cualquier manera, el descubrimiento de diversos morfotipos similares a cianobacterias filamentosas en las rocas sedimentarias de 3.5 x 10^9 años de Warrawoona (Schopf, 1993), ha venido a demostrar que la vida es un fenómeno cuya antiguedad se remonta a las primeras etapas de la formación del planeta. La complejidad estructural de estos fósiles implica, por lo tanto, que la vida surgió y alcanzó un grado considerable de diversificación en un tiempo de tan solo 10^9 años. Esta conclusión es, por supuesto, contraria al prejuicio de que estos son procesos que deben haber requirido de varios miles de millones de años (Oparin, 1924, 1938).

El problema anterior se ha agudizado al reconocer que durante su historia temprana la Tierra, al igual que otros cuerpos del Sistema Solar, sufrió un gran nnúmero de colisiones con meteoritos, cometas, y asteorides que deben haber tenido efectos devastadores. En particular, el estudio de la superficie lunar ha llevado a sugerir que hasta hace unos 3.7 x 10^9 años, todavía tenían lugar colisiones capaces de fundir el material rocoso de la luna (Chyba et al., 1990). Debido a que la Tierra no se pudo haber escapado de este proceso, se cree que hasta esas épocas chocaban contra nuestro planeta cuerpos menores capaces de fundir la superficie terrestre y de provocar la evaporación de los mares primitivos. Ello implica, por lo tanto, que la vida pudo haber surgido y desaparecido varias veces o, al menos, que no pudo haberse establecido definitivamente antes de esos tiempos (Sleep et al., 1990). Por lo tanto, quienes trabajan en origen de la vida se ven obligados a reconocer que el tiempo disponible para el surgimiento y evolución de la

biósfera primitiva debe haber quedado limitado, a lo sumo, a unos 200 o 300 millones de años.

No ha sido fácil aceptar la conclusión anterior. Después de todo, las cianobacterias son uno de los clados mas tardíos de la rama eubacteriana, y su presencia en las rocas de Warrawoona (Schopf, 1993) indicaría que bastaron 300 millones de años o menos para que surgiera la vida celular, se desarrollaran los sistemas de replicación y traducción, los organismos se diversificaran en las dos grandes ramas procariontes, y la evolución metabólica permitiera el uso del agua como donador de electrones en la reducción fotosintética del CO_2.

El reconocimiento de que buena parte del genoma procarionte parece ser el resultado de procesos de duplicación génica permite, al menos de manera aproximada, calcular el tiempo necesario para que el genoma de un heterótrofo anaerobio ancestral similar a un micoplasma alcanzara las dimensiones correspondientes al de una cianobacteria filamentosa y cuyo número de genes (Lazcano y Miller, 1994). Bajo la suposición de que las tasas de duplicación génica se pueden considerar como eventos neutrales, y que durante el Arqueano su frecuencia era comparable a que se observa en microorganismos contemporáneos (10^{-5} a 10^{-3} duplicaciones genicas por gene por generación), se concluyó que la acrección de duplicones sería del orden de 0.001 gene por año, es decir, de unos 7 nucleótidos anuales (Lazcano y Miller, 1994). Por lo tanto, el tiempo requerido para que un genoma creciera hasta alcanzar las dimensiones del de una cianobacteria, estaría entre los 5 y los 7 millones de años.

Aunque estos cálculos dan una cierta idea de los tiempos requeridos para que se incremente el tamaño de un genoma celular, es evidente que atrás de ellos existen una serie de suposiciones y simplificaciones no necesariamente válidas, incluyendo la hipótesis de que no hay pérdida de duplicones. Por otra parte, este modelo se ve limitado por la ausencia de información empírica que permita calcular los tiempos de divergencia y fijación de genes duplicados que codificaran para enzimas homólogas pero con habilidades catalíticas diferentes. Existen otros aspectos igualmente poco explorados. Por ejemplo, se sabe que las bacterias merodiploides

contemporáneas son extraordinariamente inestables, y las poblaciones rapidamente regresan a su condición haploide originales (Anderson y Roth, 19877). Sin embargo, la gran proporción de secuencias duplicadas involucradas en distintas rutas metabólicas (Fani et al., 1995; Mills, Fox, Fani, Leguina, Lazcano y Oró, en prep.) sugiere que cuando estas vias biosintéticas se establecieron, tanto la tasa de mutación como la de fijación de nuevas secuencias ocurrían con mayor eficiencia que hoy en día. En segundo lugar, se sabe que las copias duplicadas sufren fenómenos de conversión génica, lo que tiende a homogeneizar las secuencias y reduce la posibilidad de divergencia hacia nuevas funciones (Walsh, 1987). Es evidente que durante el Arqueano la conversión génica debe haber sido menos eficaz, o de lo contrario no observariamos enzimas homólogas pero involucradas en diferentes rutas biosintéticas. Las explicaciones que subyacen a estos fenómenos no se conocen aún, y representan un área no explorada de genética bacteriana susceptible de ser analizada no solo bajo una perspectiva teórica sino también experimental.

The RNA world, its predecessors and descendants

Antonio Lazcano

Departamento de Biología, Facultad de Ciencias, UNAM, Apdo. Postal 70-407, Cd. Universitaria, México 04511, DF, Mexico

According to the RNA-world hypothesis, primordial cells based on catalytic and replicative polyribonucleotides of abiotic origin were the phylogenetic forerunners of the complex functional relationship between DNA, RNA, and proteins that drives extant life. This is a Russian-doll scheme that assumes that nucleic-acid-directed protein synthesis was selected for in primordial cells because of the enhanced adaptability that protein catalysts brought with them. It is in such an enzyme-rich intracellular environment that DNA genomes eventually evolved, confiscating the genetic role that until then had been performed by RNA molecules. The process of increasing complexification assumed by the RNA world model is consistent with many examples of biological evolution in which new traits are added without the complete loss of previous characteristics. Problems involved with the prebiotic synthesis and accumulation of RNA have led to the suggestion that the RNA-based biosphere was preceded by a pre-RNA world in which genetic macromolecules were nucleic-acid-like polymers of nucleoside analogues. No explanation has been offered for the transition from such a hypothetical archaic system into the RNA world, which remains the best working model for explaining the origin and early evolution of life.

The RNA-world hypothesis states that during the long-forgotten primordial times when protein biosynthesis and DNA genomes had not yet evolved, the reproduction and metabolism of the earliest cells depended on the catalytic and replicative properties of RNA molecules (Fig. 1). This idea, which was sparked by the startling discovery of the catalytic activities of RNA molecules, has opened the possibility of understanding the origin of extant nucleic acid-directed protein biosynthesis by suggesting that both DNA and proteins are the evolutionary outcomes of RNA-based cells (Alberts 1986; Gilbert 1986; Lazcano 1986).

RNA may be unique among biomolecules because of its dual ability to serve as repository of genetic information (as in the case of infamous viruses associated with diseases like polio, AIDS, and others), and to perform catalytic activities – a property that until a few years ago was believed to be found exclusively in proteins. RNA molecules are also conspicuous structural components of all cells, where they play a key role in protein synthesis, DNA replication, and RNA processing. In one known case the polar moeity of an eubacterial membrane component is a ribonucleotide and, at least in some eukaryotes, RNA is also involved in transcription and in protein translocation

Bengtson, S. (ed.) 1993: *Early Life on Earth. Nobel Symposium 84.* Columbia U.P., New York
4 pp. 70 - 80

across membranes. It is unlikely that all these activities are relicts of a vanished RNA world. Nonetheless, they demostrate the multiple roles that polyribonucleotides can play in biological processes and thus provide indirect support to the idea that RNA-based life once existed.

As argued in the accompanying chapter, it is conceivable that early Archean RNA-cells were the ancestors of contemporary cells (Lazcano, this volume, p. ■■■). Did their appearance also mark the beginnings of life? The strong points of this idea have their corresponding weaknesses, including the non-trivial issue of the prebiotic availability of RNA molecules. The purpose of this chapter is to discuss the advantages and limitations of the RNA world hypothesis, as well as to review the evidence indicating that rudimentary protein synthesis once existed and evolved through biological mechanisms and not because of chemical events in the prebiotic environment.

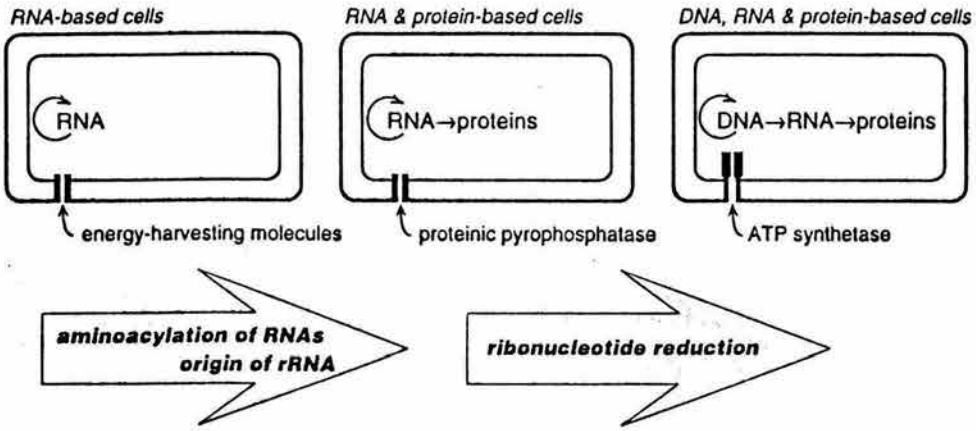


FIGURE 1 The three major steps in cellular evolution beginning from an RNA world (modified from Alberts 1986).

Imagining life in an RNA world

Despite some initial skepticism, it is now generally accepted that the cut-and-trim activities involved in the hydrolysis and transfer of phosphodiester bonds first described in self-splicing introns were the first evidence of the extraordinary versatility of ribozymes, i.e. catalytic RNA molecules (Cech & Bass 1986). Previous claims on the catalytic properties of the RNA moiety of a polysaccharide branching enzyme (Shvedova *et al.* 1987) have not been sustained, but there is a growing tide of experimental results showing that ribozymes are truly astonishing catalysts that are not confined to nucleic acid substrates. Is it possible to construct a reasonable working model of primordial RNA-based cells?

An RNA world requires ribozymes capable of replicating RNA templates. The existence of such molecules is supported by the work of Jennifer A. Doudna and Jack W. Szostak (1989), who have engineered a catalytic RNA with RNA-joining activity that

uses external templates. As noted by the Scripps Research Institute evolutionary biologist Gerald F. Joyce (1991), this ribozyme is not completely equivalent to proteinic RNA-dependent RNA polymerases: it does not move processively along the template, nor can it copy self-structured regions. However, it is reasonable to assume that the discovery of a catalytic RNA overcoming such limitations is only a matter of time.

It is unlikely that a template-dependent ribozymic polymerase can achieve absolute copying fidelity. In such a case, populations of molecules with considerable genetic variation would exist, due both to point mutation and to RNA-mediated rearrangements of polynucleotide sequences. Selection of some strands over others, i.e., darwinian evolution at the molecular level, can thus be expected (Joyce 1989). That such a system can be designed is supported by experiments in which RNA templates are copied by Qβ replicase, a viral RNA-dependent proteinic RNA polymerase (Spiegelman 1971; Biebricher *et al.* 1982). After running approximately 100 replication cycles, several smaller variants of the original templates accumulated, as they were selected by the elimination of segments unnecessary to the Qβ replication process (Mills *et al.* 1973). A similar phenomenon can be predicted in a system with catalytic RNA-mediated RNA replication, but template-shortening would not have been advantageous in an RNA-world. In other words, ribozyme-mediated replication was a necessary but not sufficient condition for the emergence of RNA-based life, whose existence and maintenance would depend on the dynamic equilibrium between replication, and on the functional interplay of additional mechanisms insuring nutrient capture from the external environment and the utilization of high-energy bonds from energy-harvesting molecules (Lazcano, this volume, p. ■■■).

If RNA were present in the primitive environment, it could not have been engaged in chemical soliloquies. Company would have been kept by a wide range of potential cofactors and substrates including metal ions, amino acids, polypeptides, sugars, lipids, and many other molecules of prebiotic origin (Oró *et al.* 1990). The coexistence of ribozymes with other chemical species could have led RNA to the acquisition of additional functional groups. Examples of such molecular interactions are provided by contemporary biochemistry. For instance, the Pb^{++}-dependent self-cleavage of a transfer RNA (Sampson *et al.* 1987) strongly suggests the existence of primordial metalloribozymes (Gilbert 1987). Another particularly interesting case is that of a membrane component of the purple non-sulphur bacterium *Rhodopseudomonas acidophila*, which is formed by a hydrophobic terpenoid covalently linked to a ribonucleotide (Neunlist & Rohmer 1985; Ourisson, this volume). This unusual compound suggests that a direct association between lipids and RNA may have existed, one which could have facilitated both encapsulation and chemical reactions in the lining of primordial liposomes.

The abiotic synthesis of several coenzymes has been achieved in the laboratory (Mar & Oró 1991), but it is also possible that contemporary nucleotide-like coenzymes are molecular remnants of an RNA world (Orgel & Sulston 1971; White 1982). This alternative possibility is strengthened by (a) the highly ubiquitous presence of pyridine nucleotide coenzymes and other ribonucleotide prosthetic groups; and (b) the fact that in the absence of their corresponding protein, many coenzymes can catalyze chemical reactions similar to those in which they take part as mere cofactors. According to the Delaware University biochemist Harold B. White (1982), even histidine, an imidazole-bearing amino acid which forms part of the active center of many enzymes, may be a

ghost of its former self. Although the prebiotic synthesis of histidine has been reported (Oró, this volume), it is the only amino acid whose biosynthesis begins from a phosphorylated sugar and a ribonucleotide. This unusual pathway has led to the idea that histidine may be the molecular descendant of a catalytic ribonucleotide derivative (White 1982). This is not a far-fetched suggestion – as a matter of fact, the disappearance of biological traits that fade away leaving behind only the shadow of their grin is a well documented evolutionary phenomenon (Margulis & Cohen, this volume).

In the RNA world, some biochemical reactions may have been spontaneous, and others could have depended on the catalytic effect of transition metals, but RNA was the major catalyst. It was probably never a very efficient one – otherwise, ribozymes would have been discovered long time ago by biochemists. The RNA world was in constant risk of sailing over the edge, since self-maintenance and reproduction of RNA-cells were probably hindered by the insidious hydrolytic cleavage of the RNA phosphodiester backbone. By assuming that darwinian mechanisms began operating earlier than previously thought, it has been argued that the enhancement of the catalytic activities of RNA-cells was the basic selective pressure underlying the origin and stabilization of the translation apparatus (Alberts 1986; Gilbert 1986; Lazcano 1986). The sequence of events leading to RNA-directed protein synthesis probably began with simple chemical interactions between amino acids and ribozymes, but it would eventually seal the fate of RNA-based cells. How this may have happened is discussed below.

A world without mirrors: the molecular predecessors of RNA

A major assumption underlying the hypothesis that life began with RNA-based cells is that catalytic and replicative RNA molecules were brewed in the prebiotic environment from random abiotic condensation reactions of ribonucleotides. Although this view is mildly supported by the discovery of rather small ribozymes (Uhlenbeck 1987), the initial optimism surrounding the possibility of an RNA-world has been challenged by an increasing awareness that current evidence does not support the abiotic formation of RNA molecules. One case in point, as cogently argued by Joyce (1989), is the prebiotic synthesis and accumulation of pyrimidines, nucleosides, and nucleotides, which face several major obstacles. Moreover, the formose reaction, traditionally invoked to account for the presence of ribose on the primitive Earth, requires unrealistically alkaline conditions, and yields a complex array of many different sugars of which ribose is only a minor and relatively unstable component (Shapiro 1986; Ferris 1987).

Although the existence of a nucleotide-rich primitive broth is debatable, recent experiments have shown that high yields of a ribose derivative are obtained by the formaldehyde-mediated condensation of phosphorylated glycolaldehyde (Müller *et al.* 1990). It is not known if this chemical pathway can lead to the accumulation of biological nucleotides (Fig. 2). However, it does suggest the existence of as-yet-unaccounted prebiotic reactions and geochemical settings that could have favored the formation of D- and L-nucleosides (Gedulin & Arrhenius, this volume). Even if we assume that

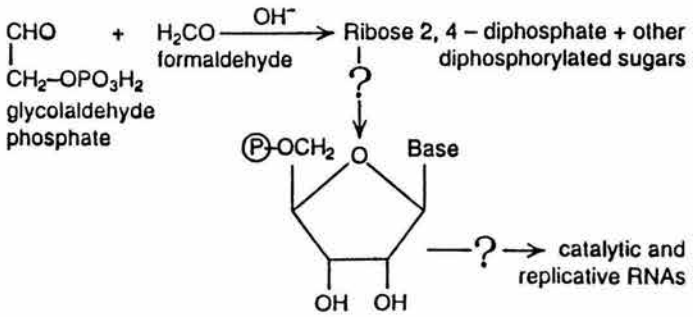


FIGURE 2 A possible prebiotic synthesis of ribonucleosides (based on Müller *et al.* 1990).

ribose was present in the primitive soup, important objections can still be raised against the RNA-world hypothesis. On the one hand, the abiotic condensation of ribose and nitrogen bases leads to a mixture of nucleosides with different configurations, including not only those with the β-glycosidic linkage of contemporary nucleic acids but also the slightly twisted geometry of the non-biological alpha form. On the other hand, while biological systems use only D-ribose, both the formose reaction and the condensation of glycolaldehyde–phosphate produce racemic mixtures, i.e., equal amounts of D- and L-ribose and their derivatives. This is a major obstacle, since under such racemic conditions, non-enzymatic template-dependent polymerization reactions are halted by the so-called enantiomeric cross-inhibition, which rapidly stops chain elongation (Joyce 1991).

As argued forcefully by Joyce *et al.* (1987), some of these problems can be avoided by assuming that the most archaic genetic system was based not on RNA, but on simpler polymers of prochiral open-ring nucleoside analogues such as glycerol, a simple three-carbon chain alcohol (Fig. 3). Since such molecules are prochiral, i.e., lack mirror images, enantiomeric cross-inhibition would not have prevented their replication. Support for this possibility has been provided by the pioneering work of Alan W. Schwartz and Leslie L. Orgel (1985), who have shown that nucleoside analogues can undergo template-dependent polymerization reactions. These experiments have been extended to a large variety of compounds, including phosphoramidates, acetic acid derivatives, deoxynucleoside diphosphates, and many other molecules that yield oligomers with rather unusual backbone structures (Rodriguez & Orgel 1991). We are thus faced with the dazzling possibility of a pre-RNA world based on informational polymers akin neither to RNA or DNA, molecular structures that lacked ribose, phosphates, and perhaps even the conventional nitrogen bases found in contemporary nucleic acids!

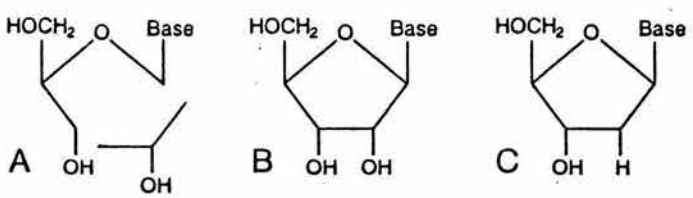


FIGURE 3 A: Glycerol-based nucleoside analog. B: Ribonucleoside. C: Deoxyribonucleoside.

At the very least, the study of proto-nucleic acids suggests that many different molecules may have been involved in replication-like reactions in the primitive environment. Did such reactions form part of a biological prelude to the RNA world? Such a possibility pushes further back the origin of life, and touches other issues like the antiquity of protein synthesis and the genetic code, which may be older than RNA itself (Orgel 1987). As argued in the accompanying chapter, the pre-RNA world hypothesis is not without problems of its own (Lazcano, this volume, p. ■■■). Acceptance of a pre-RNA scenario requires (a) evidence of catalytic activities in nucleic acid-like molecules; (b) a convincing explanation of its evolutionary transition into an RNA world, which may have involved a step-wise process through several intermediates (Orgel 1987); and (c) if ribonucleotides cannot form abiotically, then a primordial ribozyme-based metabolism with the ability to synthesize them is required. Is such an RNA-based system feasible? Answers to this and other equally significant questions lie in the largely untreaded ground of the chemistry of nucleoside analogues and their polymers. Until adequate solutions are offered to overcome these problems, the RNA world may be considered as a reasonable model for understanding the transition from non-living to living (Lazcano, this volume, p. ■■■).

The origin of protein biosynthesis

The possibility that DNA genomes evolved prior to protein synthesis cannot be completely discarded. Nonetheless, according to the RNA-world hypothesis, ribosome-mediated protein synthesis evolved in membrane-bounded systems in which RNA molecules had functioned until then, both as the source of inheritable genetic information and as principal catalysts. It has been difficult, however, to exorcize the looming legacy of academic emphasis on the proteinic nature of biological catalysts. There is still some resistance to the idea that RNA could have played a primordial role in biological catalysis in the absence of proteins. It is frequently argued that nucleic acid replication and genetic coding of proteins coevolved, i.e. that protein synthesis and replicative nucleic acids emerged as a result of the interactions between catalytic peptides and replicative RNA templates (Oró, this volume). Are proteins and nucleic acids molecular siamese twins, as implied by the different coevolution theories on the origin of translation?

I find it difficult to support this possibility. Of course, there is considerable evidence suggesting that amino acids and oligopeptides were present in the primitive Earth (Oró *et al.* 1990). Histidyl–histidine and other small catalytic oligopeptides that can use nucleotides or oligonucleotides as substrates have been synthesized under plausible prebiotic conditions (Brack & Barbier 1989; Shen *et al.* 1990c). It is doubtful, however, that the promiscuous coexistence of replicative oligonucleotides and catalytic peptides would lead to an inheritable liaison between them. Even if ('and oh, what a big if') we assume that huge amounts of oligopeptides with catalytic and structural characteristics necessary for polynucleotide replication were available in the prebiotic environment, sooner or later this supply would have been exhausted or hydrolyzed in the absence of translation. Regardless of how many different peptides were formed on the primitive Earth, nucleic-acid-instructed protein synthesis could never have evolved without a

replicating mechanism insuring the maintenance, stability, and diversification of its basic components.

Protein synthesis is an elaborate exquisitely-tuned process requiring more than a hundred distinct components which include ribosomal RNA (rRNA) and proteins, transfer RNAs (tRNAs), aminoacyl-tRNA synthetases, and proteins such as the initiation and elongation factors. How the actual formation of peptidic bonds takes place inside the ribosome is still a matter of conjecture. Tampering with rRNA has dramatic effects on translation, but there is no evidence that ribosomal proteins have any kind of catalytic activity (Nomura 1987). No ribosomal protein is known to catalyze the formation of a covalent bond between adjacent amino acids, and recent experiments (Noller *et al.* 1992) have confirmed previous speculations that this ability lies in rRNA itself (Nomura 1987; Noller *et al.* 1990). The observations by Noller *et al.* (1992) add considerable credibility to the idea that protoribosomes were devoid of proteins (Woese 1967, 1980; Crick 1968). How did the ancestral rRNA originate? The UC-Santa Cruz molecular biologist Harry F. Noller (1991) has suggested, from observations of von Ahsen *et al.* (1991) on the inhibition of self-splicing ribozymes by antibiotics which are also known to affect translation, that ribosomes may have evolved from an RNA related to the so-called group I catalytic introns. As noted by Peter B. Moore (1988) of Yale University, there is a certain chemical similarity between the trans-esterification reactions in ribozymes (Cech & Bass 1986) and the transpeptidation event that takes place during protein synthesis.

Interactions between amino acids and RNA in the primitive environment were almost unavoidable, and it is generally agreed that the attachment of amino acids to polyribonucleotides was one of the earliest steps of protein synthesis to evolve. According to Yale University molecular biologists Alan Weiner and Nancy Maizels (1987), transfer RNAs may be derived from terminal tRNA-like structures that tagged primordial genomes at their 3' end, marking an initiation site for ribozymic-mediated RNA replication. As argued by Orgel (1989), the linkage of amino acids or dipeptides to RNAs could have provided a transcription initiation site, while according to Wächtershäuser (1988b), aminoacylated RNA molecules could associate to cationic mineral surfaces and undergo further changes. An additional possibility has been raised by the Hong-Kong University biochemist J. Tze-Fei Wong (1991), who has noted that the bonding of amino acids to ribozymes may have increased their catalytic activities by adding more chemically active functional groups. These ideas are not mutually exclusive. Aminoacylation may have originated as a tagging process, and then been maintained and further refined and exploited because of the additional capabilities gained by RNA-cells.

There is no experimental evidence showing that ribozymes can react with activated amino acids and catalyze the formation of peptide bonds, but there are several indications that such reactions are feasible. A statistically significant correlation between the polarity and the hydrophobicity of amino acids and their anticodon nucleotides has been described, hinting to a primordial interaction between RNAs and amino acids that may be related to the origin of the genetic code (Lacey & Mullins 1983). The specific binding of arginine to a catalytic RNA (Yarus 1988b), and of aromatic amino acids to fragments of phenylalanine tRNA (Bujalowski & Porschke 1988), are also well documented. Even more encouraging are the recent results that have ex-

panded the known reperoire of ribozyme chemistry by showing that peptide bond formation depends on the catalytic properties of ribosomal RNA (Noller *et al.* 1992), and that RNA also has the ability to hydrolyze the bonds that join amino acids to RNA – suggesting that the reaction can also proceed in the opposite direction if the equilibrium conditions are changed (Piccirilli *et al.* 1992).

Although it is not known how protein synthesis began, there are several independent but complementary lines of evidence showing that a rudimentary version of this process once existed. Data supporting this possibility includes:

1 *In vitro* experiments in which peptide bond formation is catalyzed solely by a ribosomal RNA moiety (Noller *et al.* 1992), or can take place in systems with complete ribosomes but in the absence of a number of proteinic components of the translation apparatus such as initiation and elongation factors (Spirin 1986). Protein synthesis probably took place in a distant past without these molecules which may be related to the appearance of regulatory or optimization mechanisms (Woese 1980; Spirin 1986; Moore 1988).

2 An extensive comparison of primary and secondary structures of cellular, mitochondrial, and plastid rRNAs has led to the recognition of conserved highly defined cores significantly smaller than the typical eubacterial ribosomal RNAs. Since these minimal rRNAs contain most of the functional sites involved in translation, they may be the oldest recognizable functional portion of extant ribosomes (Gray & Schnare 1990).

3 The genes coding for the two elongation factors in eubacteria and their homologues in the other two major cellular lineages are the result of an ancient gene duplication thought to predate the separation of prokaryotes into eubacteria and archaebacteria (Iwabe *et al.* 1989). This suggests that a simpler process of protein biosynthesis, proceeding with only one elongation factor, may have taken place (Lazcano, this volume, p. ■■■).

As argued persuasively by Carl R. Woese (1967), primitive translation must have been an ambiguous error-ridden process, with triplets coding probably not for individual amino acids but for classes of amino acids with similar physicochemical properties. It is unlikely that the first RNA-directed proteins emerged fully endowed with catalytic properties. Tertiary structure plays a major role in extant enzymatic activity, but the earliest coded proteins must have been rather small oligopeptides. Their properties probably depended more on their primary structure than on the limited secondary and tertiary conformations available to them. The first proteins may have been nucleic acid-binding oligopeptides involved in RNA unwinding or in ribozyme stabilization. A good contemporary example of such interactions is the functional interplay between the components of the RNA-protein complex of ribonuclease P (Lazcano 1986; Westheimer 1986), in which the stabilizing effect of the basic protein over the RNA moiety enhances its catalytic effects (Altman 1984). Such hybrid intermediate stages in which RNA and proteins acted together may have existed, but it is unlikely that they lasted for long: with the appearance of protein synthesis the RNA world ran full stride towards evolutionary oblivion.

Into the DNA world

Protein synthesis can take place without DNA but not in the absence of RNA. It is thus reasonable to assume that DNA cellular genomes are the result of an early molecular takeover that took place well after the origin of life, when protein synthesis was already established and different enzymes were available. In the words of Bruce Alberts (1986), 'all DNA functions must have evolved in an intracellular environment rich in protein catalysts, where RNA catalysis had become largely obsolete.' This conclusion is supported by the fact of deoxyribonucleotide biosynthesis, which removes the highly reactive 2'-OH group from a pool of preexisting ribonucleotides (Lammers & Follmann 1983). It is possible that this enzyme-mediated process may have been acquired in cells with RNA genomes as a final step in nucleotide biosynthesis (Lazcano *et al.* 1988), a mechanism that is consistent with the pathway in which the DNA-specific deoxythymidilate is formed by adding a methyl group to a deoxy-derivative of the RNA-specific base uracil (Kornberg & Baker 1992).

That DNA had displaced RNA genomes long before 3.5 Ga ago is suggested by the morphological complexity of the Warrawoona microfossils (Schopf, this volume). No archaebacterial ribonucleotide reductase gene has been cloned and sequenced, but it is likely that DNA cellular genomes are a monophyletic trait that evolved prior to the divergence of the three main cellular lineages. Sequence similarities shared by ancient proteins found in all three lines of descent suggest that considerable fidelity already existed in the then-operative genetic system (Lazcano *et al.* 1992). Given the chronic high mutation rates of RNA genomes, it is unlikely that such fidelity could have been achieved if the last common ancestor of eubacteria, archaebacteria and eukaryotes lacked DNA and repair mechanisms insuring its genetic integrity.

As shown by its amazing recovery from a magnolia leaf between 17 and 20 millions years old (Golenberg *et al.* 1990), double-stranded DNA is an extremely resistant macromolecule. It is generally agreed that DNA genomes were selected over RNA for a very simple reason: the latter are fragile reactive polymers that undergo many chemical changes, including their almost complete hydrolysis. Genetic information stored in RNA degrades because of the cytosine's strong tendency to deaminate to uracil and the lack of a correcting enzyme. Furthermore, the lack of substantial amounts of free atmospheric oxygen and the consequent lack of an ozone shield would have led to a high ultraviolet flux in the early Archean, leading to deleterious rates of UV-induced mutations in cells with RNA genomes. Since DNA-repair systems depend on the duplication of the genetic information contained in the complementary strands of the duplex DNA molecules (Friedberg 1985), the emergence of double-stranded DNA would have stabilized the earlier irreparable system leading to the selection of mechanisms to correct damage caused by UV-light (Lazcano *et al.* 1988).

The appearance of DNA led to more stable ways of storing genetic information in hitherto highly-mutating RNAs with limited coding abilities. Since DNA genomes are error-correcting because of their double-stranded structure, their presence opened the possibility of increasing genome size by gene duplication. Of course, multiple copies of the same gene may have existed in cells with RNA genomes. However, no RNA polymerase is endowed with the proof-reading activity that DNA polymerases possess. RNA replication is an intrinsically noisy process, one that limits template size,

since the number of accumulated point mutations is proportional to their template length. As shown by some contemporary viruses, this limitation may be overcome in part by segmented genomes (Reanney 1982). It is reasonable to assume that in some intermediate stage of early cellular evolution, genomes were disaggregated and rapidly mutating RNAs, in which several copies of the same gene existed. Because of the difficulties in insuring genetic identity of their offsprings, such cells would be rapidly selected against.

The evolutionary emergence of double-stranded DNA genomes and of DNA polymerases with editing properties allowed the drastic development of large cellular genomes with increased coding potential. The appearance of DNA unleashed the enormous catalytic potential of proteins. Tinkering of exons and of the products of gene duplications made proteins truly malleable commodities, developing their catalytic prowess and enhancing the fitness of primitive cells. The use of new substrates and the regulation of metabolic pathways became possible, leading to intricate webs of reactions involved in basic metabolic networks and well-attuned biochemical processes. The world of modern cells with DNA, RNA and proteins was well on its way.

Conclusions

Catalytic RNA may be a molecular pentimento of a bygone early stage of biological evolution, but were the first forms of life actually based on ribozymes? As summarized by Joyce (1991), there is an unbridgeable gulf between our current descriptions of the primitive environment and the biochemical properties of RNA molecules. In spite of this limitation and of its inherent panselectionist explanations, the RNA-world hypothesis has the advantage of readdressing the problem of the origin of proteins and DNA from an articulate novel perspective amenable to empirical analysis.

From this standpoint, future major insights on the emergence of nucleic acid-directed protein synthesis can be expected to result not from chemical simulation experiments, but from detailed characterizations of ribozymes and the development of *in vitro* evolving RNA systems. Further understanding of the origin of extant DNA genomes will be provided by phylogenetic comparisons of genes from the major cellular lineages coding for ribonucleotide reductases, thymidilate synthesases, DNA polymerases, primases, and other proteins involved in DNA replication. Results from such research can be expected to influence other fields of biological enquiry. For instance, are RNA viruses, retroviruses, and DNA viruses related to the evolutionary transition from RNA to DNA cellular genomes, as argued by Weiner (1987b), or should we seek an explanation to their origin in more recent biological processes?

The study of the origins of life has focused mainly on the appearance of proteins and nucleic acid replication. A wealth of data has accumulated, but a broader approach is needed. More emphasis should be given to the long neglected question of the emergence of basic metabolic pathways. The notion of a retrograde evolution, suggested many years ago by the Caltech biologist Norman H. Horowitz (1945) to explain the appearance of metabolic pathways, is not supported by current evidence. In fact, the alternative idea that metabolic pathways evolved through the 'patchwork' assembly of primitive proteins with broad substrate specificity (Jensen 1976) is consistent with the

properties of early nucleic-acid-coded peptides discussed in this chapter, and it is also supported by the homologous character of different enzymes involved in widely separated biosynthetic processes (Parsot 1987; Lazcano *et al.* 1992).

In the past decade considerable progress has been achieved in our understanding the emergence and early evolution of living systems, but we are still haunted by major uncertainties, the magnitude of which is matched only by our ignorance. Even though the sequence of evolutionary events discussed in this chapter may be correct, continuing enquiries into prebiotic chemistry, paleobiological analysis, and molecular phylogenetic comparisons are required to fully validate it. Empirical evidence is the ultimate bonfire of our theoretical vanities.

Acknowledgements. – I wish to express my gratitude to the organizers of the 1992 Nobel Symposium *Early Life on Earth* for their kind invitation to contribute with this additional chapter. I am indebted to Dr. Stefan Bengtson for his continuous encouragement and patience, and to Drs. Gail R. Fleischaker and Juan Oró, who painstakingly read the manuscript and made many helpful and constructive comments. I thank Drs. Thomas R. Cech, Gustaf Arrhenius, and Joseph A. Piccirilli for kindly sharing with me the results of their work prior to publication. This manuscript was finished during a short leave of absence at the University of Houston, under the auspices of NASA Grant 44-005-002 to Juan Oró. Work reported here has been supported in part by UNAM.IN. 105289.

References: Lazcano (RNA World)

Alberts, B.M. 1986: The function of the hereditary materials: biological catalyses reflect the cell's evolutionary history. *American Zoologist 26*, 781–796. [*63, 70, 71, 73, 78*]

Altman, S. 1984: Aspects of biochemical catalysis. *Cell 36*, 237–239. [*77*]

Biebricher, C.K., Diekmann, S. & Luce, R. 1982: Structural analysis of self-replicating RNA synthesized by Qb replicase. *Journal of Molecular Biology 154*, 629–648. [*72*]

Brack, A. & Barbier, B. 1989: Early peptidic enzymes. *Advances in Space Research 9*, 83–87. [*75*]

Bujalowski, W. & Porschke, D. 1988: Contributions to selective binding of aromatic amino acid residues to $tRNA^{Phe}$. *Biophysical Chemistry 30*, 151–157. [*76*]

Cech, T.R. & Bass, B.L. 1986: Biological catalysis by RNA. *Annual Review of Biochemistry 55*, 599–629. [*57, 62, 71, 76, 132*]

Crick, F.H.C. 1968: The origin of the genetic code. *Journal of Molecular Biology 38*, 367–379. [*63, 76, 138*]

Doudna, J.A. & Szostak, J.W. 1989: RNA-catalyzed synthesis of complementary-strand RNA. *Nature 339*, 519–522. [*64, 71*]

Ferris, J.P. 1987: Prebiotic synthesis: problems and challenges. *Cold Spring Harbor Symposia on Quantitative Biology 52*, 29–35. [*64, 73*]

Friedberg, E.C. 1985: *DNA Repair*. 613 pp. Freeman, San Francisco, Calif. [*78*]

Gilbert, W. 1986: The RNA world. *Nature 319*, 618. [*62, 68, 70, 73, 88, 125, 132*]

Gilbert, W. 1987: The exon theory of genes. *Cold Spring Harbor Symposia on Quantitative Biology 52*, 901–905. [*72, 511, 515*]

Golenberg, E.M., Giannasi, D.E., Clegg, M.T., Smiley, C.J., Durbin, M., Henderson, D. & Zurawski, G. 1990: Chloroplast DNA sequence from a Miocene *Magnolia* species. *Nature 344*, 656–658. [*78*]

Gray, M.W. & Schnare, M.N. 1990: Evolution of the modular structure of rRNA. *In* Hill, W., Dahlberg, A., Garrett, R.A., Moore, P.B., Schelssinger, D. (ed.): *Ribosomes: Structure, Function and Genetics*, 589–597. American Microbiological Society, Washington, D.C. [*77*]

Horowitz, N.H 1945: On the evolution of biochemical synthesis. *Proceedings of the National Academy of Sciences, USA 31*, 153–157. [*79*]

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. & Miyata, T. 1989: Evolutionary relationship of Archaebacteria, Eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proceedings of the National Academy of Sciences, USA 86*, 9355–9359. [*67, 77, 144, 145, 153, 189, 288*]

Jensen, R.A. 1976: Enzyme recruitment in evolution of new function. *Annual Review of Microbiology 30*, 409–425. [*79*]

Joyce, G.F. 1989: RNA evolution and the origins of life. *Nature 338*, 217–224. [*54, 57, 72, 73, 132, 135*]

Joyce, G.F. 1991: The rise and fall of the RNA world. *The New Biologist 3*, 399–407. [*72, 74, 79*]

Joyce, G.F., Schwartz, A.W., Orgel, L.E. & Miller, S.L. 1987: The case for an ancestral genetic system involving simple analogues of the nucleotides. *Proceedings of the National Academy of Sciences, USA 84*, 4398–4402. [*64, 74*]

Kornberg, A. & Baker, T. 1992: *DNA Replication. 2nd Edition*. 932 pp. Freeman, San Francisco, Calif. [*66, 78*]

Lacey, J.C. & Mullins, D.W. 1983: Experimental studies related to the origin of the genetic code and the process of protein synthesis^s– a review. *Origins of Life 13*, 3–42. [*76*]

Lammers, M. & Follmann, H. 1983: The ribonucleotide reductases^s– a unique group of metalloenzymes essential for cell proliferation. *Struct••••• Bonding 54*, 27–91. [*78*]

Lazcano, A. 1986: Prebiotic evolution and the origin of cells. *Treballs de la Societat Catalana de Biologia 39*, 73–103. [*63, 70, 73, 77*]

Lazcano, A., Fox, G.E. & Oró, J. 1992: Life before DNA: the origin and evolution of early Archean cells. *In* Mortlock, R.P. (ed.): *The Evolution of Metabolic Function*, 237–295. CRC–Telford, Caldwell, N.J. [*52, 57, 61, 62, 64, 66, 67, 68, 78, 80*]

Lazcano, A., Guerrero, R., Margulis, L. & Oró, J. 1988: The evolutionary transition from RNA to DNA in early cells. *Journal of Molecular Evolution 27*, 283–290. [*78*]

Mar, A. & Oró, J. 1991: Synthesis of the coenzymes adenosine diphosphate glucose, guanosine diphosphate glucose, and cytidine diphosphoethanolamine under primitive Earth conditions. *Journal of Molecular Evolution 32*, 201–210. [*54, 72*]

The transition from non-living to living

Antonio Lazcano

Departamento de Biología, Facultad de Ciencias, UNAM, Apdo. Postal 70-407, Cd. Universitaria, México 04511, DF, Mexico

Where, when and how did life appear? Although we have no detailed answers for these three equally alluring questions, if current interpretations on the significance of the properties of RNA molecules for the origin and early evolution of life are correct, then a major step towards the appearance of cells was the emergence of a liposome-bounded system in which energy conversion became associated with template-directed ribonucleotide polymerization, producing both catalytic and replicative RNA molecules within its lipidic membranes. The attributes of the first forms of life are unknown, but preliminary insights gained from molecular phylogenetic analysis are providing unequivocal evidence that the organisms that preceded eubacteria, archaebacteria and the eukaryotic nucleocytoplasm component were ancestral prokaryotes with simpler ATP-synthetases and protein-synthesis machinery. This suggests that a large number of (not necessarily slow) uncharacterized evolutionary changes took place between the origin of life itself and the last common ancestor of all extant life.

'All the organic beings which have ever lived on this Earth', wrote Charles Darwin in the *Origin of Species*, 'may be descended from some one primordial form'. But how did this common ancestor come into being? What was its nature? Although Darwin never overcame his reluctance to discuss in public the appearance of life, it was within the framework of his ideas that seventy years later A.I. Oparin and J.B.S. Haldane suggested a possible explanation for the emergence of the first living systems, based on the hypothesis that the earliest organisms were fermentative, obligate anaerobic bacteria that had been preceded by a long period of chemical abiotic synthesis of organic compounds.

Alternative routes to biopoiesis have been suggested (Wächtershäuser, this volume), including the possibility that the origin of life was concomitant with the fortuitous formation of a single replicating ribozyme, i.e., a catalytic RNA molecule self-assembled from unorganized prebiotic raw material in which lifelike properties were completely absent. However, the possibilities of scientific enquiry are much broader if the study of life's emergence is approached by assuming a procession of changes, through stages of gradually increasing complexity, until a system which can be recognized as living is attained (Oró, this volume). If this scheme is valid, then there must have been a turning point in this evolutionary process during which the transition from non-living to living took place. The study of this crucial but largely undefined stage is the subject of this short essay.

4

Bengtson, S. (ed.) 1993. *Early Life on Earth. Nobel Symposium 84.* Columbia U.P., New York

pp. 60-69

What is life?

Music', once said Isaac Stern, 'can be described, but not defined'. Perhaps the same is true of life itself. An all-embracing, generally agreed-upon definition of life has proven to be an elusive intellectual endeavour, but any explanation of the origin of living systems should attempt the definition of a set of minimal criteria for what constitutes a living organism, including the extremely elementary basic characteristics with which the first living beings were endowed. What are these essential attributes? As argued forcefully by Gail R. Fleischaker (1990), there is a categorical distinction between non-living and living, and the latter can be characterized by operational criteria that account not only for the internal structure, organization and operation of organisms, but also their interactions with their environment. Despite the spectacular molecular acrobatics performed by viruses, viroids, catalytic RNAs, and many other subcellular systems, extant life is generally identified at the very minimum with cells, i.e., with dynamic membrane-bounded systems incessantly exchanging matter and energy with their environment, with the common imperative operations involved in basic metabolism, self-maintenance, heredity and reproduction with variation. The history of change and continuity between the earliest forms of life and extant organisms implies, says the Cambridge University philosopher Harmke Kamminga (1992), that 'the first living organisms took part in the evolutionary process – in other words, that they had descendants unlike themselves'.

Unfortunately, the inability to discriminate between traits that may have resulted from truly abiotic processes, and those that are outcomes of biological evolution, has led to the frequent misconception that modern cells are perfect models for the first forms of life. Evolutionary criteria have frequently been absent in the chemical approach to the origins of life. For instance, the non-enzymatic synthesis of deoxyribose, thymine and many different oligodeoxyribonucleotides has been achieved in several laboratories. Do these results imply that wriggling DNA molecules were floating in the waters of the primitive ocean, ready to be used as primordial genes? Of course, this is unlikely. From a biological perspective the presence of DNA in contemporary cells can be explained not in terms of prebiotic chemistry, but rather as the endproduct of an ancient metabolic pathway that evolved in early Archean cells possessing RNA genomes in which translation had already appeared (Lazcano *et al.* 1992).

Given the adequate expertise and experimental conditions, it is possible to synthesize almost any organic molecule. The fact that a number of molecular components of contemporary cells can be formed non-enzymatically in the laboratory does not necessarily mean that they were also essential for the origin of life, or that they were available in the prebiotic environment. The primitive broth must have been a bewildering organic chemical wonderland, but it could not include all the compounds or the molecular structures found today in even the most primitive prokaryotes – nor did the first bacteria spring completely assembled, like Frankenstein's Monster, from simple precursors present in the prebiotic soup.

The RNA world revisited

'RNA and DNA are the dumb blondes of the biomolecular world', wrote Francis Crick in *Life Itself* (1981), 'fit mainly for reproduction (with a little help from proteins) but of little use for much of the really demanding work'. This may still be true for DNA, but neither for blondes nor RNA. The independent discovery of ribozymes by Thomas Cech of the University of Colorado and Sidney Altman of Yale University barely one year after Crick's book was published quickly raised RNA from a humble biochemical position as a molecular handyman and mere go-between, to a central character in the early evolutionary drama (Fig. 1).

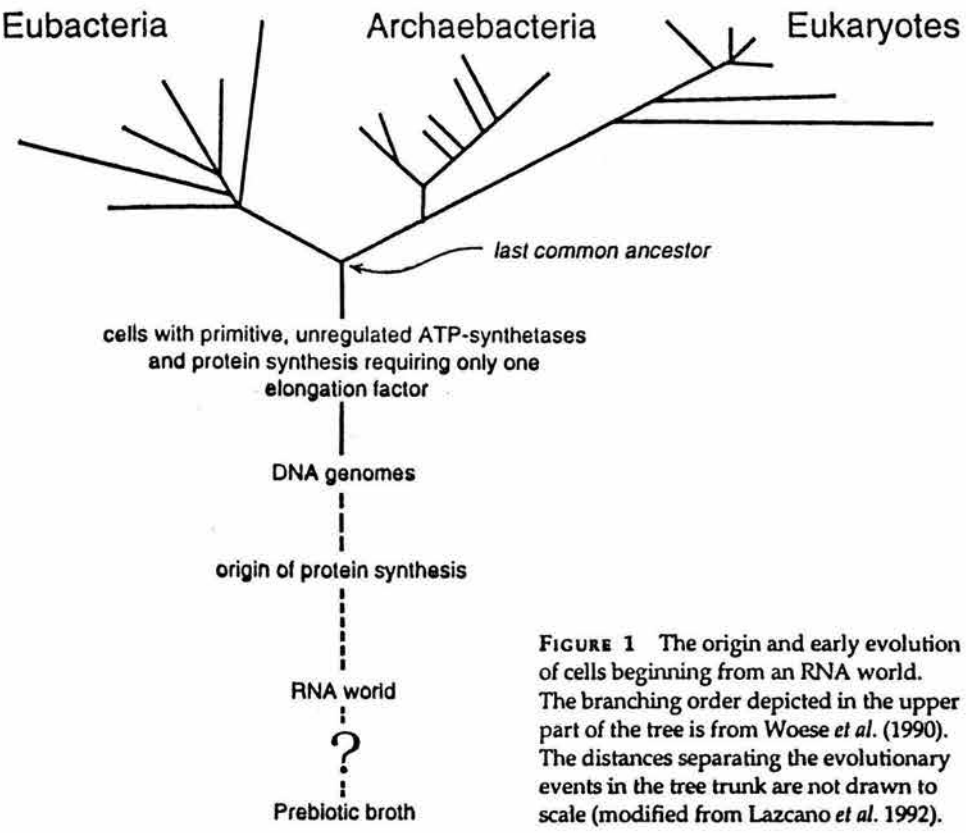


FIGURE 1 The origin and early evolution of cells beginning from an RNA world. The branching order depicted in the upper part of the tree is from Woese *et al.* (1990). The distances separating the evolutionary events in the tree trunk are not drawn to scale (modified from Lazcano *et al.* 1992).

It is unlikely that the intron self-splicing reaction discovered by Cech and the catalytic abilities of the RNA moeity of ribonuclease P described by Altman are truly vestigial activities (Cech & Bass 1986). However, the disclosure of RNA-mediated catalysis led Harvard's molecular biologist Walter Gilbert (1986) to suggest that the starting point for life's history on Earth had been the so-called 'RNA world', an early stage during which alternative life forms based on ribozymes existed prior to the development of protein biosynthesis and DNA genomes. Somewhat similar proposals

have been made independently by a number of authors, but together with Wally Gilbert, Bruce Alberts (1986) and I (Lazcano 1986) specifically argued that the development of the translation machinery had begun in the snug, organic-rich microenvironment defined by the lipidic boundaries of cells whose metabolism and reproduction had been mediated until then by ribozymes. Thus, although the abiotic formation of peptide bonds under possible primitive conditions is well documented (Oró *et al.* 1990), it may have no direct relevance to the origin of protein biosynthesis. According to this rather liberal definition of primordial life, the basic selection pressure for the origin and stabilization of the primitive translational apparatus was the enhancement of the catalytic activities of these RNA-based cells in order to increase their dynamic stability and reproductive fitness. More is said on the RNA world in my other chapter (Lazcano, this volume, pp. ■■■–■■■).

The RNA world is not a radical, totally unheard of hypothesis without historical precedent. The existence of a primitive replicating and catalytic apparatus devoid of both DNA and proteins, and based solely on RNA molecules was suggested in the late 1960's by Carl Woese (1967), Francis Crick (1968) and Leslie Orgel (1968). As pointed out by the British biologist Norman W. Pirie in 1953, 'if we found a system doing things that satisfied our requirements for life but lacking proteins, would we deny it the title?'. However, in part due to deeply rooted biochemical prejudices, the idea of a RNA world has met with some healthy resistance, especially from those who argue that protein synthesis is such an essential characteristic of cells that its origin should be considered synonymous with the emergence of life itself. This pervasive view has been strongly challenged by the increasing evidence that RNA molecules are efficient and versatile catalysts. Ribozyme-mediated peptide bond formation has not been demostrated, but there are strong indications that the RNA substrate repertoire may include amino acids, as hinted by several observations, including (a) the discovery of an arginine binding site in a ribozyme from the ciliate *Tetrahymena* (Yarus 1988a); and (b) the ability of a slightly modified form of this same ribozyme to catalyze the hydrolysis of the aminoacyl bond between formylmethionine and an oligonucleotide (Piccirilli *et al.* 1992); and (c) the nearly conclusive evidence that the formation of peptide bonds can be catalyzed solely by ribosomal RNA in the absence of proteins (Noller *et al.* 1992).

God may not play dice – but Nature can be tricky. Are the proponents and adherents of the RNA world an evolutionary sect seduced by false coincidences? The answer is probably negative. In addition to its dual ability as a catalyst and as an informational macromolecule, RNA is a resilient, chemically reactive, abundant structural component of all cells, as of the ribosome where it appears to be more than a simple molecular scaffold (Noller 1991; Noller *et al.* 1992). Ribonucleotides are essential metabolic precursors in the biosynthesis of deoxyribonucleotides, and they also play a central role in contemporary cellular metabolism (as ATP, for instance, or as ribonucleotide derivatives like NADH, acetyl CoA, and FADH, among others), which may be a reflection of their ancient origin. Moreover, approximately 50% of all catalogued enzymes cannot function without coenzymes, many of which are small ribonucleotide-like molecules whose biosynthesis is intimately linked to the metabolism of RNA and its monomers (White 1982). As argued long time ago by Orgel & Sulston (1971), these coenzymes may be vestiges of pre-genetic code catalysts present in early RNA-cells before the appearance of true proteins and the development of enzymes.

Although as of December 1992 a template-dependent ribozymic RNA polymerase has not been discovered, the application of powerful molecular-biology techniques has led to the synthesis of an artificial ribozyme able to catalyze repeatedly the ligation of short oligonucleotides complementary to a RNA template (Doudna & Szostak 1989). The possibility of simple prebiotic ribozymes is supported at least in part by the discovery of the rapid, highly specific cleavage reaction catalyzed by a small, synthetic 19-nucleotide RNA molecule under physiological conditions described by Uhlenbeck (1987). Nonetheless, the RNA-world model confronts several serious challenges, including the lack of plausible primitive abiotic mechanisms to account for the formation and accumulation of ribose, which is a minor, relatively unstable component of a complex array of products formed simultaneously from the putative prebiotic self-condensation of formaldehyde (Shapiro 1986). Indeed, the problems involved with prebiotic synthesis of ribose and other nucleic acid components (Ferris 1987) have led to the suggestion that RNA itself may have been preceded by genetic polymers with a simpler backbone in which ribose presumably was replaced by acyclic, flexible compounds like glycerol, a stable, three-carbon chain compound (Joyce *et al.* 1987).

Attempts to bury the RNA world together with the spoils of other ephemeral scientific speculations may be premature and should be met with caution. In spite of the appeal of a pre-RNA era based on nucleic-acid-like molecules, (a) ribose analogs may not be suitable at all, as suggested by recent experiments showing that nucleic-acid double helixes are easily destabilized by the presence of even a few flexible glycerol-nucleosides (Schneider & Benner 1990); (b) ribozymic activity is strongly dependent on the ribose 2'-OH group, which plays a direct role in different hydrolytic, phosphorylation and condensation reactions, as well as in intron self-splicing. The absence of an equivalent hydroxyl group in glycerol and other acyclic ribose analogues seriously hinders their possible catalytic activity (Lazcano *et al.* 1992); and (c) recent experiments by A. Eschenmoser and his associates at the ETH in Zurich have shown that high yields of a ribose derivative are easily obtained by the formaldehyde-mediated phosphorylation of glycoaldehyde (Müller *et al.* 1990), suggesting that a ready prebiotic synthesis of RNA components may exist. This possibility is supported by the work of Gustaf Arrhenius and his associates (Gedulin & Arrhenius, this volume), who has shown that hydrotalcite, an abundant hydroxide mineral, concentrates the glycoaldehyde phosphate and catalyzes the formation of sugar phosphates at mild conditions. What may be actually required is an experimental redefinition of the chemical conditions for a truly prebiotic synthesis of polyribonucleotides.

Natura non facit saltus?

In spite of its inherent limitations and reductionist overtones, the RNA-world hypothesis has an intrinsic heuristic value that cannot be overstressed. At the very least, it is a causal narrative, i.e., a logical, evolutionary-oriented plausible sequence of events that attempts to explain the transition from a simple replicating system based on RNA to one involving DNA and proteins. The idea that life begun with the appearance of RNA-based cells may help to overcome the historical dispute between those that argue that the first living system was a self-replicating nucleic acid molecule, and those who

identify the origin of life with the emergence of a membrane-bounded, polymolecular, heterogeneous system endowed with basic metabolic properties. But how did the hypothetical RNA-cells suggested by Alberts, Gilbert and myself come into being?

Although we are still far from a complete understanding of the processes that may have led to the formation of the so-called prebiotic broth, the presence of a large array of organic compounds in carbonaceous meteorites, and the astonishing easiness by which amino acids, adenine, lipids and many other molecules can be synthesized under primitive conditions imply that a large set of biochemicals was readily formed on the early Earth (Oró *et al.* 1990). In particular, the chemical synthesis of amphiphilic molecules (Deamer 1986a), the non-enzymatic template-directed polymerization of nucleotides and nucleoside analogs (Orgel 1987), and the abiotic synthesis of histidyl–histidine and other small catalytic peptides (Shen *et al.* 1990c), suggest that replication-like processes, chemically active peptides, and the self-assembly of membranes from lipidic components were possible before the emergence of life.

Primitive liposomes were probably relatively simple structures, formed by small, single-chain, ionic linear fatty acids, which could easily sequester catalytic and replicative molecules (Deamer, this volume). Under laboratory conditions this process can take place in the presence of histidine, cyanamide, and several other prebiotic condensing agents and is enhanced by basic polypeptides and metallic cations (Oró & Lazcano 1990). Mixtures of different lipids are known to produce liposomes with non-selective pores, and complexes between nucleotides and metal ions could have facilitated diffusion, leading to rudimentary transport mechanisms across primitive membranes.

Precellular evolution was not a continuous, unbroken chain of progressive transformations steadily proceeding to the first living beings. Many prebiotic culs-de-sac and false starts probably took place. Emergence of the first living beings must have required the simultaneous coordination of many different components in a confluence of processes. Thus, the mere encapsulation of ribozymes, amino acids, oligopeptides and many other potential cofactors and substrates involved in RNA catalysis within liposomes was a necessary but not sufficient condition for the origin of RNA-cells. Perhaps the first forms of life did not require membranes (Lamont & Gibson 1990), but if our interpretation of the evolutionary significance of the properties of RNA molecules is correct, then a major decisive step towards the appearance of life was the emergence of a system in which energy coupling associated with membrane and proton gradients was used in template-directed ribonucleotide polymerization to produce both catalytic and replicative RNA molecules within the lipidic boundaries of primitive liposomes.

The work of Peter Gogarten and his colleagues (1989) has shown that proton pumps producing H^+ gradients appeared very early in cellular evolution, prior to the evolutionary divergence of eubacteria and archaebacteria. These complex oligomeric enzymes may have been preceded by a simpler proton-pumping pyrophosphatase with H^+-PPase and PP_i synthetase activities (Baltscheffsky & Baltscheffsky, this volume), but how proton gradients actually originated and became coupled with ions and directionality is still an unsolved problem. Very little is known of the origin of biological energy conversion mechanisms, although it is likely that ion channels and proton-selective pores appeared prior to the origin of life. Small, simple synthetic amphiphilic oligopeptides long enough to span the hydrocarbon phase of lipid bilayers have permeabilities and lifetimes resembling those of proton-selective channels and the acetylcho-

line receptor (Lear *et al.* 1988). Although these experiments have not been performed within an evolutionary context or under primitive conditions, they illustrate how small oligopeptides of prebiotic origin, or those synthesized by primitive cells with limited coding capabilities, could have been involved in ion-transport across membranes (Lazcano *et al.* 1992).

According to the scheme discussed in this chapter, survival and reproduction of primordial RNA-cells depended on ribonucleotides, lipids, and other compounds of prebiotic origin, and must have been hindered by the exhaustion of this supply. It is this direct uptake of organic molecules from the primitive environment, and not the universal distribution of glycolysis, that should be interpreted as the defining feature of the heterotrophic nature of the first cells. In spite of its simplicity, central metabolic position, and ability to function under anaerobic conditions, glycolysis as such requires a set of enzymes too complex to be expected in the first organisms. Extant bacteria have various relatively inefficient mechanisms that allow them to use organic molecules from external sources, including nitrogen bases and nucleosides (Kornberg & Baker 1992). Some of these salvage pathways may be analogous (or perhaps, even homologous) to the uptake by primitive heterotrophes of nucleic acid components from the external milieu.

Early biological evolution: the molecular chronicles

As shown by the Warrawoona fossil assemblage, an abundant, complex and highly diversified microbiota which may have included cyanobacteria, existed only 10^9 years after the Earth had formed (chapters by Schopf and Walter, this volume). Life is probably much older than these early fossils, but how long did it take for it to appear and become established? Almost nothing is known about the timescales required for the origin and evolution of bacterial metabolic pathways. The primitive environment was no microbial Eden, but the Archean paleontological record shows that once life emerged it was rapidly able to endure, diversify, and adapt itself to the stinky, harsh environmental conditions of the early Earth.

It is unlikely that the paleontological record will ever provide direct evidence of the transition from prebiotic organic molecules to the earliest cells, nor will it tell us much about the nature of the first biological systems. However, as shown more than twenty-five years ago by Emile Zuckerkandl and Linus Pauling (1965), nucleic acids and protein sequences are an extraordinarily rich source of evolutionary information. Although a cladistic approach to the origin of life is not feasible, the comparison of ribosomal RNA sequences has become an important tool in understanding the early stages of cellular evolution and has had a significant impact in our interpretation of bacterial relationships. A major achievement of this approach was the construction of a trifurcated, unrooted universal evolutionary tree in which all known organisms can be grouped in one of three major lineages: the eubacteria, the archaebacteria, and the eukaryotic nucleocytoplasm (Woese 1987b; Sogin, this volume).

The immediate predecessor of these three cellular lines was already a rather complex organism, much alike to extant bacteria in many ways. Few genes found in the three major cellular lineages have been sequenced and compared, but the sketchy picture that is already emerging of the last common ancestor of eubacteria, archaebacteria and eukaryotes, shows that it was a rather sophisticated cell with complex ribosome-mediated translation, membrane-associated H^+-ATPases engaged in active transport, histidine and purine synthetic abilities, and a set of enzymes involved in glycolysis, pyruvate oxidation, and other mainstream heterotrophic anaerobic metabolic pathways.

It is likely that genetic recombination appeared early in evolution, suggesting that Archean microbes led a life that was not totally chaste. However, gene duplications followed by further sequence divergence were probably the most important mechanism by which early cells increased their hereditary endowment. These evolutionary innovations arose in individual organisms, and then spread rapidly through ancestral bacterial populations, becoming fixed prior to their divergence into the three cellular lineages. In contrast with orthologous genes, which are duplicate sets that diverge through speciation, paralogous genes are those that diverge after a duplication event. Paralogous genes are extremely useful in rooting evolutionary trees, since one set of sequences can be used as an outgroup for the other one. As discussed by Gogarten *et al.* (1989) and by Iwabe *et al.* (1989), the sets of paralogous genes unequivocally identified in all three cellular lineages are those coding for (a) the two elongation factors that assist in protein biosynthesis; and (b) the two components of the hydrophilic ATP-synthesizing unit of ATP synthetase, an ubiquitous membrane-associated protein complex that harvests the energy associated with proton gradients, forming ATP from ADP and phosphate.

In spite of the intense dispute on the taxonomic significance of rooted universal trees derived from paralogous genes, the upper part of Fig. 1 clearly shows that the earliest detectable branching event led to the eubacterial line on the one hand, and to the archaebacterial–eukaryotic lineage on the other (Woese *et al.* 1990). Millions of years later, during the Proterozoic, some of their descendants would meet once more, becoming forever associated in intimate symbioses – but that, of course, is another chapter in the saga of cellular evolution (Margulis & Cohen, this volume).

What is important for the present discussion is to recognize that if the last common ancestor of eubacteria and archaebacteria had two sets of duplicate homologous genes coding for elongation factors and for the ATP-synthetase units, then it must have been preceded by a simpler cell with a smaller genome in which only one copy of each of these genes existed (Fig. 1). In other words, the ancestor of the eubacterial and archaebacterial lines was a prokaryote in which ATP-synthesis and protein biosynthesis were both less complex than those of the even simpler extant life forms (Lazcano *et al.* 1992). Evolutionary biologists have long argued that organisms simpler than extant bacteria existed, but such claims were based on highly evolved entities such as viruses, mitochondria, mycoplasma and others, none of which are free-living. These analogies are useful, but now an even more important task is the identification and characterization of additional (if any) sets of such pre-common ancestor genes that can lead to a more complete understanding of the biological attributes of these bygone Archean prokaryotes.

Conclusions

'What we do not know today we shall know tomorrow' concluded A.I. Oparin in his 1924 book *The Origin of Life*. 'A whole army of biologists is studying the structure and organization of living matter, while a no less number of physicists and chemists are daily revealing to us new properties of inanimate things. Like two parties of workers boring from the two opposite ends of a tunnel, they are working towards the same goal. The work has already gone a long way and very, very soon the last barriers separating the living from the non-living will crumble under the attack of patient work and powerful scientific thought.'

In spite of the spectacular results achieved by this two-way approach, there is still a huge insurmountable gulf between the results achieved by laboratory simulations, and our present-day understanding of the essential features of a truly minimal living being. Elegant experiments that combine selection and mutation of catalytic RNA molecules have been performed (Beaudry & Joyce 1992), but given the present state of both prebiotic chemistry and molecular biology, it is probably preposterous to attempt the laboratory synthesis of RNA-based life. However, the RNA-world hypothesis is amenable to experimental analysis, including the *in vitro* development of ribozymes with new substrates, the study of simple membrane-associated energy-harvesting molecules, and the characterization of RNA replicating systems within liposomes. Using the techniques of molecular phylogenetic analysis we are peeking into the molecular intimacies of the primitive organisms that preceded the bifurcation of archaebacteria and eubacteria. The preliminary results discussed here already suggest that a long series of evolutionary changes took place, after the origin of life itself but prior to the first speciation event separating the ancestors of eubacteria from those of archaebacteria.

We may be able to see even further back in time. Our molecular remembrance of things past may soon allow us to gaze into the evolution of proteins older than DNA itself. As pointed out twenty years ago by the late Margaret Dayhoff (1972), most amino acid sequences can be classified into relatively few families. Recent development of sequence databases of genes and gene products has confirmed her early insight (Gilbert 1986; Doolittle 1990). It is possible that these few families resulted from amplification processes of ancestral genes coding for proteins whose basic functional properties and structural constraints were established prior to the emergence of DNA genomes, a hypothesis that could be tested by a detailed statistical analysis of the available databases (Lazcano *et al.* 1992). This approach can be complemented by the cloning and sequencing of ancient genes, such as those coding for thymidilate synthetase and ribonucleotide reductases, the enzymes directly involved in deoxyribonucleotide biosynthesis. The world of cells with RNA genomes in which protein biosynthesis had already appeared is not totally lost.

We have gained some insights into the biological processes that took place in the early Archean world. Nevertheless, we are still very far from understanding the origin and nature of the first living beings. These are still unsolved problems – but they are not completely shrouded in mystery, and this is no minor scientific achievement. Why should we feel disappointed by our inability to even foresee the possible answers to

these luring questions? As the Greek poet Konstantinos Kavafis once wrote, Odysseus should be grateful not because he was able to return home, but because of what he learned on his way back to Ithaca. It is the journey that matters.

Acknowledgements. – I am deeply indebted to the organizers of the 1992 Nobel Symposium *Early Life on Earth* for their kind invitation to participate in this meeting. This chapter was written during a short leave of absence in which I first enjoyed the hospitality of the Ovando Foundation (México), and then of Dr. Juan Oró and his associates at the University of Houston (NASA Grant 44-005-002). I thank Drs. Gail R. Fleischaker and Juan Oró, as well as Mary Alpaugh, for their critical reading of this manuscript and many suggestions, and Drs. Harmke Kamminga, Thomas R. Cech, Gustaf Arrhenius and Joseph Piccirilli for kindly providing me with results of their work prior to publication. Work reported here has been supported in part by UNAM.IN 105289.

References: Lazcano (Transition)

Alberts, B.M. 1986: The function of the hereditary materials: biological catalyses reflect the cell's evolutionary history. *American Zoologist 26*, 781–796. [*63, 70, 71, 73, 78*]

Beaudry, A.A. & Joyce, G.F. 1992: Directed evolution of an RNA enzyme. *Science 257*, 635–641. [*68*]

Cech, T.R. & Bass, B.L. 1986: Biological catalysis by RNA. *Annual Review of Biochemistry 55*, 599–629. [*57, 62, 71, 76, 132*]

Crick, F.H.C. 1968: The origin of the genetic code. *Journal of Molecular Biology 38*, 367–379. [*63, 76, 138*]

Crick, F. 1981: *Life Itself: its origin and nature*. 200 pp. Simon & Schuster, New York. [*62*]

Dayhoff, M.O. (ed.) 1972: *Atlas of Protein Sequence and Structure 5*. 418 pp. National Biomedical Research Foundation, Washingon, D.C. [*68*]

Deamer, D.W. 1986a: Role of amphiphilic compounds on the evolution of membrane structure on the early Earth. *Origins of Life and Evolution of the Biosphere 17*, 3–25. [*65*]

Doolittle, R.F. 1990: What we have learned and will learn from sequence databases. *In* Bell, G.I. & Marr, T.G. (eds.): *Computers and DNA*, 21–31. Addison Wesley, Menlo Park, Cal. [*68*]

Doudna, J.A. & Szostak, J.W. 1989: RNA-catalyzed synthesis of complementary-strand RNA. *Nature 339*, 519–522. [*64, 71*]

Ferris, J.P. 1987: Prebiotic synthesis: problems and challenges. *Cold Spring Harbor Symposia on Quantitative Biology 52*, 29–35. [*64, 73*]

Fleischaker, G.R. 1990: Origins of life: an operational definition. *Origins of Life and Evolution of the Biosphere 20*, 127–137. [*57, 61*]

Gilbert, W. 1986: The RNA world. *Nature 319*, 618. [*62, 68, 70, 73, 88, 125, 132*]

Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E.J., Bowman, B.J., Manolson, M.F., Poole, R.J., Date, T., Oshima, T., Konishi, J., Denda, K. & Yoshida, M. 1989: Evolution of the vacuolar H^+-ATPase: implications for the origin of eukaryotes. *Proceedings of the National Academy of Sciences, USA 86*, 6661–6665. [*65, 67, 144, 153, 189, 288*]

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. & Miyata, T. 1989: Evolutionary relationship of Archaebacteria, Eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proceedings of the National Academy of Sciences, USA 86*, 9355–9359. [*67, 77, 144, 145, 153, 189, 288*]

Joyce, G.F., Schwartz, A.W., Orgel, L.E. & Miller, S.L. 1987: The case for an ancestral genetic system involving simple analogues of the nucleotides. *Proceedings of the National Academy of Sciences, USA 84*, 4398–4402. [*64, 74*]

Kamminga, H. 1992: The structure of explanation in biology and the construction of theories of the origin of life on Earth. *UROBOROS 2*, 47–65. [*61*]

Kornberg, A. & Baker, T. 1992: *DNA Replication. 2nd Edition*. 932 pp. Freeman, San Francisco, Calif. [*66, 78*]

Lamond, A.I. & Gibson, T.J. 1990: Catalytic RNA and the origin of genetic systems. *Trends in Genetics 6*, 145–149. [*65*]

Lazcano, A. 1986: Prebiotic evolution and the origin of cells. *Treballs de la Societat Catalana de Biologia 39*, 73–103. [*63, 70, 73, 77*]

Lazcano, A., Fox, G.E. & Oró, J. 1992: Life before DNA: the origin and evolution of early Archean cells. *In* Mortlock, R.P. (ed.): *The Evolution of Metabolic Function*, 237–295. CRC–Telford, Caldwell, N.J. [*52, 57, 61, 62, 64, 66, 67, 68, 78, 80*]

Lear, J.D., Wasserman, Z.R. & DeGrado, W.F. 1988: Synthetic amphiphilic peptide models for protein ion channels. *Science 240*, 1177–1181. [*66*]

Müller, D., Pitsch, S., Kittaka, A., Wagner, E., Wintner, C.E. & Eschenmoser, A. 1990: Chemie von Alpha-Aminonitrilen. Aldomerisierung von Glycolaldehyd-phosphat zu racemischen Hexose-2,4,6-triphosphaten und (in Gegenwart von Formaldehyd) racemischen Pentose-2,4-diphosphaten: rac-Allose-2,4,6-triphosphat und rac-Ribose-2,4-diphospha*Helvetica Chimica Acta 73*, 1410–1468. (2ND PART OF TITLE: racemischen Pentose-2,4-diphosphaten: rac-Allose-2,4,6 triphosphat und rac-Ribose-2,4-diphosphat sind die Reaktionshauptprodukte.) [*54, 64, 73, 74, 92, 97*]

Noller, H.F. 1991: Drugs and the RNA world. *Nature 353*, 302–303. [*57, 63, 76*]

Noller, H.F., Hoffarth, V. & Zimniak, L. 1992: Unusual resistance of peptidyl transferase to protein extraction procedures. *Science 256*, 1416–1419. [*63, 76, 77*]

Oparin, A.I. 1924: *Proizkhozhdenie zhizni*. 72 pp. Moskovskij Rabochij, Moscow. (English translation published as Appendix in Bernal, J.D. 1967: The Origin of Life. 1–345. World, Cleveland, Ohio.) [*52, 68, 124, 133, 143*]

Orgel, L.E. 1968: Evolution of the genetic apparatus. *Journal of Molecular Biology 38*, 381–393. [*63*]

Orgel, L.E. 1987: Evolution of the genetic apparatus: a review. *Cold Spring Harbor Symposia on Quantitative Biology 52*, 9–16. [*56, 65, 75*]

Orgel, L.E. & Sulston, J.E. 1971: Polynucleotide replication and the origin of life. *In* Kimball, A.P. & Oró, J. (eds.): *Prebiotic and Biochemical Evolution*, 89–94. North-Holland, Amsterdam. [*63, 72*]

Oró, J. & Lazcano, A. 1990: A holistic precellular organization model. *In* Ponnamperuma, C. & Eirich, F.R. (eds.): *Prebiological Self Organization of Matter*, 11–34. Deepak, Hampton, Va. [*56, 65*]

Oró, J., Miller, S.L. & Lazcano, A. 1990: The origin and early evolution of life on Earth. *Annual Review of Earth and Planetary Sciences 18*, 317–356. [*52, 54, 55, 63, 65, 72, 75, 89*]

Piccirilli, J.A., McConnel, T.S., Zaug, A.J., Noller, H.F. & Cech, T.R. 1992: Aminoacyl esterase acitivity of the *Tetrahymena* ribozyme. *Science 256*, 1420–1424. [*63, 77*]

Pirie, N.W. 1953: Ideas and assumptions about the origin of life. *Discovery 14*, 238–242. [*63*]

Schneider, K.C. & Benner, S.A. 1990: Oligonucleotides containing flexible nucleoside analogs. *Journal of the American Chemical Society 112*, 453–455. [*64*]

Shapiro, R. 1986: *Origins – a skeptics guide to the creation of life on Earth*. 332 pp. Summit, New York, N.Y. [*64, 73*]

Shen, C., Lazcano, A. & Oró, J. 1990c: The enhancement activities of histidyl-histidine in some prebiotic reactions. *Journal of Molecular Evolution 31*, 445–452. [*56, 65, 75*]

Uhlenbeck, O.C. 1987: A small catalytic oligoribonucleotide. *Nature 328*, 596–600. [*64, 73*]

White, H.B. 1982: Evolution of coenzymes and the origin of pyridine nucleotides. *In* Everse, J., Anderson, B. & You, K.S. (eds.): *The Pyridine Nucleotide Coenzymes*, 2–17. Academic Press, New York, N.Y. [*63, 72, 73*]

Woese, C.R. 1967: *The Genetic Code: The Molecular Basis for Genetic Expression*. 200 pp. Harper & Row, New York, N.Y. [*63, 76, 77*]

Woese, C.R. 1987b: Bacterial evolution. *Microbiological Reviews 51*, 221–271. [*39, 66, 144, 152, 156, 162, 163, 182, 301, 335*]

Woese, C.R., Kandler, O. & Wheelis, M.L. 1990: Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences, USA 87*, 4576–4579. [*62, 67, 144, 145, 152, 153, 159, 290, 303*]

Yarus, M. 1988a: Specificity of arginine binding by the *Tetrahymena* intron. *Biochemistry 28*, 980–988. [*63*]

Zuckerkandl, E. & Pauling, L. 1965: Molecules as documents of evolutionary history. *Journal of Theoretical Biology 8*, 357–366. [*66, 144, 152, 182, 468*]

The Significance of Ancient Paralogous Genes in the Study of the Early Stages of Microbial Evolution

Antonio Lazcano

Departamento de Biología, Facultad de Ciencias-UNAM, Apdo. Postal 70-407, Cd. Universitaria, México 04510, D.F., México

Keywords: RNA world, ancient paralogous genes, early microbial evolution

Summary

Although how life originated in our planet is unknown, there is considerable evidence suggesting the existence of an early stage of biological evolution during which RNA molecules and ribonucleotides played a major role in cellular processes. It is argued that the use of ancient sets of paralogous genes (such as those coding for the two elongation factors involved in protein biosynthesis, and for the α and ß subunits of F-type ATPases), may provide information about simpler, less-regulated biological processes that apparently took place in ancestral cells predating the last common ancestor of the eubacterial and the archaebacterial lineages. Some examples of putative ancestral sets of paralogous genes are provided.

1. Introduction

Considerable progress has been achieved in the past fifty years in the understanding of the origin and early evolution of life, but this field of scientific enquiry is haunted by several major problems (10). These include (a) our inability to achieve accurate reconstructions, even at a broad scale, of the environmental conditions of the primitive Earth; (b) the difficulties involved in explaining the prebiotic synthesis of RNA molecules; (c) the lack of a working model of a primitive cell based solely on the catalytic and replicative properties of RNA; and (d) the existence of a huge gap in the reconstruction of the evolutionary events that took place after the putative primordial RNA-based forms of life, and the last common ancestor of extant cellular lines. In this paper I will argue that the detection and evolutionary analysis of ancient sets of paralogous genes whose sequences are found today in eubacteria, archaebacteria and eukaryotes, are a powerful tool in the characterization of simpler biological systems that preceded the divergence of early cells into the two major prokaryotic lineages.

2. Life in the RNA World

The detailed chemical nature of the first genetic polymers and the catalytic agents involved in primordial metabolic processes are still unknown. The problems involved with the prebiotic synthesis and accumulation of D-ribose, pyrimidines, nucleotides and oligonucleotides has led to the suggestion that RNA was

Trends in Microbial Ecology
R. Guerrero & C. Pedrós-Alió (eds.)

preceded by simpler genetic polymers of prochiral open-ring nucleoside analogues based on glycerol or other equivalent compounds (6). Nonetheless, it is generally agreed that RNA molecules played a major role during the early stages of biological evolution. This possibility is supported by the dual abilities of RNA molecules to store genetic information and to act as catalysts. There is increasing experimental evidence of the catalytic flexibility of ribozymes, which have been shown to act not only over nucleic acid substrates, but also appear to play a major role in peptide bond formation (9). These results support the hypothesis that ribosome-mediated peptide bond formation emerged within the lipidic boundaries of RNA-based cells (7). The RNA world was probably a short-lived stage of biological evolution (5), but it left major marks in contemporary cells: RNA molecules play a central role in protein synthesis and other biological processes, and ribonucleotides (a) are universal precursors in the biosynthesis of deoxyribonucleotides; (b) are an essential moiety of a large number of coenzymes; and (c) form part of a cellular system of distress signals or alarmones, which are modified ribonucleotides such as cAMP, ppGpp, pppGpp, and AppppA (7).

3. Early Cellular Evolution

Phylogenetic analysis of small subunit ribosomal RNA has firmly established the existence of three cellular lineages represented today by eubacteria, archaebacteria, and the eukaryotic nucleocytoplasm (13). The comparison of the traits found among these different lineages suggest that their last common ancestor, or "progenote" (13), was already a complex cell comparable in many ways to modern prokaryotes (7). Accordingly, this "progenote" could not be an immediate descendant of the RNA world. It is not possible to root phylogenetic trees based solely on the comparison of small subunit rRNA sequences, since no known organism can be used as an outgroup. However, rooting can be achieved using paralogous genes, i.e., homologous sequences that diverged after gene duplication, by using one set of them as an outgroup for the other one. Using two such sets, i.e., those coding for the α and the β subunits of the hydrophilic portions of the F_0F_1-type ATPases (2), and for the two elongation factors involved in protein biosynthesis (3), phylogenetic trees have been constructed that shown that all the major archaebacterial lines are more closely related to eukaryotes than to eubacteria (2, 3, 14). This conclusion is in agreement with phylogenetic reconstructions based on the comparison of DNA dependent RNA polymerases and ribosomal proteins, as well as on the distribution of other traits common to archaebacteria and eukaryotes (7, 11).

The rooting technique described above (2, 3) is possible because the last common ancestor of eubacteria and archaebacteria was already endowed with two homologous genes coding for two elongation factors, as well as with F-type ATPases having homologous α and β subunits. Accordingly, it can be concluded that the population of ancestral cells were in turn preceded by simpler ones, in which protein synthesis took place with only one type of an elongation factor, and with ATPases lacking the α regulatory subunit. This implies that cells predating the last common ancestor of the two prokaryotic lineages lacked the sophisticated

regulatory abilities found in contemporary bacteria. Such conclusion supports previous suggestions on the existence of primordial metabolic pathways based on enzymes of broad substrate specificity and limited regulatory mechanisms (4, 15).

Based on the available databanks, it has been argued (3) that other sets of paralagous genes may include those coding for (a) the lactate- and malate dehydrogenases; (b) the valyl-tRNA and isoleucyl-tRNA synthetases; and (c) the LepA protein and the initiation factor 2. Additional evidence of ancestral paralogous duplications may be found in genes coding for enzymes involved in nucleic acid metabolism. The monophyletic origin of two different carbomoyltranferases involved in the biosynthesis of arginine and of pyrimidines (12) suggest that the genes that code for them are part of a wider paralogous assemblage whose duplication and divergence predated the evolutionary separation of eubacteria and archaebacteria. In addition, recent evidence of the presence of members of the B family of DNA polymerases in the three cellular lineages, as well as of their homology with the A family of DNA polymerases (1), implies that these enzymes are also coded by ancient paralogous genes that can be used to infer the existence of simpler ancestral cells with only one DNA polymerase, that predated the last common ancestor of the three cellular lineages.

4. Conclusions

Since paralogous sequences do not provide evidence of speciation but on the order of gene duplications, they have been avoided in phylogenetic comparisons. As argued in this paper, however, they are a powerful tool that may allow us to look further back in time, and glimpse into the characteristics of ancestral cells predating the last common ancestor of the two prokaryotic lineages. In addition to the examples discussed in the previous section, the sequence alignment of the valyl-tRNA and isoleucyl-tRNA synthetases (3), as well as the evolutionary grouping of the aminoacyl-tRNA synthetases for glutamic acid/glutamine, and for aspartic acid/asparagine (8), strongly suggest that these enzymes are coded by paralogous genes that in principle could be used to trace the evolution of the extant genetic code from a simpler version that must have existed long before the divergence of eubacteria and archaebacteria. Since the study of very early stages cellular evolution could benefit from this type of evolutionary inferences, more effort should be devoted to the development of criteria for identification, sequencing and evolutionary analysis of additional ancestral paralogous genes.

Acknowledgement

Work reported here has been supported in part by UNAM IN 105289.

References

1. Forterre, P. (1992) Nucleic Acid Res. 20: 1811

2. Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B.J., Manolson, M.F., Poole, R.J., Date, T., Oshima, T., Koshini, J., Denda, K., & Yoshida, M. (1989) Proc. Natl. Acad. Sci. USA 86: 6661

3. Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., & Miyata, T. (1989) Proc. Natl. Acad. Sci. USA 86: 9355

4. Jensen, R.A. (1976) Annu. Rev. Microbiol. 30: 409

5. Joyce, G.F. (1991) The New Biologist 3: 399

6. Joyce, G.F, Schwartz, A.W., Orgel, L.E., & Miller, S.L. (1987) Pro. Natl. Acad. Sci. USA 84: 4398

7. Lazcano, A., Fox, G.E. & Oró, J. (1992) in R.P. Mortlock (ed), The Evolution of Metabolic Function (CRC Press, Boca Raton), 297

8. Nagel, G.M. & Doolittle, R.F. (1991) Proc. Natl. Acad. Sci. USA 88: 8121

9. Noller, H.F., Hoffarth, V., & Zimniak, L. (1992) Science 256: 1416

10. Oró, J., Miller, S.L., & Lazcano, A. (1990) Annu. Rev. Earth Planet.Sci. 18: 317

11. Sogin, M. (1991) Current Opinion Genet. Develop. 1: 457

12. Van Vliet, F., Cunin, R., Jacobs, A., Piette, J., Gigot, D., M. Lawereys, Piérard, A., & Glansdorff, N. (1984) Nucleic Acid Res. 12: 6277

13. Woese, C.R. (1987) Microbiol. Rev. 51: 221

14. Woese, C.R., Kandler, O., & Wheelis, M.L. (1991) Proc. Natl. Acad. Sci. USA 87: 4576

15. Yeas, M. (1974) J. Theoret. Biol. 44: 145

Adv. Space Res. Vol. 15, No. 3, pp. (3)345–(3)356, 1995

0273–1177/95 $7.00 + 0.00

4

ON THE LEVELS OF ENZYMATIC SUBSTRATE SPECIFICITY: IMPLICATIONS FOR THE EARLY EVOLUTION OF METABOLIC PATHWAYS

A. Lazcano,* E. Díaz-Villagómez,* T. Mills** and J. Oró**

* *Departamento de Biología, Facultad de Ciencias-UNAM, Apdo. Postal 70-407, Cd. Universitaria, México 04510, D.F. Mexico*
** *Department of Biochemical and Biophysical Sciences, University of Houston, Houston, TX 77204-5934, U.S.A.*

ABSTRACT

The most frequently invoked explanation for the origin of metabolic pathways is the retrograde evolution hypothesis. In contrast, according to the so-called "patchwork" theory, metabolism evolved by the recruitment of relatively inefficient small enzymes of broad specificity that could react with a wide range of chemically related substrates. In this paper it is argued that both sequence comparisons and experimental results on enzyme substrate specificity support the patchwork assembly theory. The available evidence supports previous suggestions that gene duplication events followed by a gradual neoDarwinian accumulation of mutations and other minute genetic changes lead to the narrowing and modification of enzyme function in at least some primordial metabolic pathways.

INTRODUCTION

Prompted by the discovery of the catalytic abilities of some RNA molecules, the idea that primordial metabolism was based on ribozymes has gained considerable acceptance /1/. There is considerable evidence suggesting that RNA played a major role during the early stages of biological evolution, including the fact that ribonucleotides (a) are universal biosynthetic precursors of deoxyribonucleotides; (b) form part of a cellular system of distress signals or alarmones /2/; and (c) are an essential moiety of a significant percentage of coenzymes /3/. There is considerable intellectual appeal in the RNA world hypothesis, but a number of as yet unsolved issues remain, including that of the prebiotic availability of ribozymes /4, 5/. Furthermore, although there is increasing experimental evidence of the catalytic versatility of ribozymes /6-9/, we still lack a working model of primitive life forms entirely based on ribozymes.

How did the first catalytic proteins emerge? As noted more than 25 years ago by M. Eden /10/, it is extremely unlikely that catalytically useful proteins could have originated from the

spontaneous random assembly of their amino acid monomers. The synthesis of chemically active peptides /11, 12/ and other highly reactive organic molecules under possible primitive conditions has been reported /13/, but there is a growing number of findings supporting the possibility that protein synthesis evolved in RNA-based cells /14-16/. This evidence includes (a) the discovery of an highly specific arginine-binding site in a ribozyme from the ciliate *Tetrahymena* /17/; (b) the ability of a slightly modified form of this same ribozyme to hydrolyze the aminoacyl bond between formylmethionine and a small oligonucleotide /8/; and (c) the observations by Noller *et al.* /9/ suggesting that protein-depleted ribosomes can catalyze peptide bond formation.

It is unlikely that the first proteins were already endowed with catalytic properties. Tertiary structure plays a major role in extant enzymatic activity, and it has been argued that the earliest RNA-coded proteins were small nucleic acid-binding oligopeptides involved in the stabilization of ribozyme catalytic conformations /1, 16, 18/. It has also been suggested that primordial RNA genetic material had an intron/exon structure involving autocatalytic introns, whose evolution eventually led to proteins being assembled from exons from the very beginning /18/. This model provides an explanation for the acquisition of new enzymatic abilities and metabolic traits in early Archean microbial populations. As summarized elsewhere by D. L. Hartl /19/, enzymes with truly novel catalytic activities evolve due to combinatorial processes involving exon reassortment /18, 20-23/ and the rearrangement of smaller functional units /24/.

Support for the mechanism discussed by Hartl /19/ has been provided by many examples of exon shuffling, including the startling discovery that the gene coding for the mammalian low-density lipoprotein receptor is a mosaic that shares a number of homologous exons with the genes for epidermal growth factor and the complement blood-clotting factors IX and X /25/. However, in this paper we will argue that both sequence comparisons and experiments on enzyme substrate specificity, suggest that the orthodox neoDarwinian gradualist step-wise mode of evolution may be succesfully applied to explain the appearance of at least some metabolic pathways. These may have been assembled in early Archean times from enzymes originally endowed with broad substrate specificity and by subunit shuffling.

DID METABOLIC PATHWAYS EVOLVE BACKWARDS?

Although the attempt to understand the evolution of microbial metabolism was one of the main intellectual drives underlying the 1920's development of A. I. Oparin's theory on the origin of life /26/, it was not until 1945 when the first explanation of the emergence of metabolic pathways was developed by Norman H. Horowitz /27/. According to this hypothesis, primordial heterotrophic bacteria had acquired different biosynthetic abilities in a step-wise, sequential development of enzymes in a reverse order as found in extant pathways. Thus, if the contemporary biosynthesis of compound A involved the sequential transformation of precursors D, C, and B mediated by enzymes ***c***, ***b***, and ***a***, as shown below,

$$\mathrm{D} \xrightarrow{c} \mathrm{C} \xrightarrow{b} \mathrm{B} \xrightarrow{a} \mathrm{A}$$

then it is unlikely that simultaneous mutations will give rise to enzymes ***a***, ***b***, and ***c***. Based both on the ideas of A. I. Oparin /28/ and on the Beadle-Tatum one gene-one enzyme theory, Horowitz /27/ assumed that once substance A became depleted in the prebiotic environment, its synthesis from a closely related compound B was possible due to a single mutation from enzyme ***a***. Similarly, when B was exhausted, enzyme ***b*** would arise from enzyme ***a*** by a single mutation. Thus, the last enzyme ***a*** in the biosynthetic pathway was the first to appear,

and the first enzyme *c*, the last one /27/. This idea, also known as the retrogade hypothesis, was further refined by Horowitz in a later publication /29/, where he assumed that clusters of genes coding for enzymes involved in biosynthetic pathways are the result of early tandem gene duplication events followed by divergence.

As discussed below, changes in substrate specificity due to relatively few mutations can be invoked to explain the use of new amides by contemporary eubacteria /30/. Horowitz's retrogade hypothesis is also supported by results from abiotic chemical synthesis, and from degradation processes, during which some of the intermediates are found to be identical or similar to those produced by metabolic pathways /31/. This appears to be case of the alkaline degradation of glucose-6-phosphate /32/, as well as of the prebiotic synthesis of purines /33, 34/, orotic acid /35/ and uracil /36-37/.

The retrograde evolution hypothesis can used for a partial description of the origin of the eubacterial biosynthesis of methionine, since it has been shown that two sequential steps are mediated by the homologous enzymes β-cystathionase and cystathionine γ-synthase /38/. However, the Horowitz hypothesis /27, 29/ does not explain the origin of regulatory mechanisms involved in catabolism, or the development of metabolic pathways involving sequential series of dissimilar reactions. Additional objections to retrograde evolution have been summarized elsewhere /39/, and include the following:

(a) most metabolic intermediates are chemically unstable molecules /40-43/, and it is difficult to explain their synthesis and accumulation in both the prebiotic and extant environments;

(b) many of these metabolic intermediates are phosphorylated compounds that could not permeate primordial membranes in the absence of specialized transport systems that were probably absent in primitive cells /41, 43/;

(c) according to the retrogade evolution hypothesis, succesive steps in metabolic pathways should involve similar chemical transformations. With the exception of the *Escherichia coli* methionine biosynthesis discussed above /38/, no additional examples are known that satisfy this condition /44, 45/; and

(d) if gene duplications are invoked to account for the different enzymes involved in the same metabolic pathway /27, 29/, then these enzymes must share structural similarity. However, no evidence of relatedness has been found among the different genes forming the *E. coli* tryptophan operon /46/. Analysis of additional partial operon sequences that are currently available also fails to support the retrograde hypothesis.

THE PATCHWORK ASSEMBLY OF METABOLIC PATHWAYS

Although the retrograde hypothesis has frequently been invoked to explain the development of successively longer biosynthetic pathways /34, 47-49/, the available evidence suggests that its application to the understanding of evolution of metabolism is rather limited. In contrast to the Horowitz hypothesis, S.G. Waley /50/ suggested the step-wise model of evolutionary development of metabolic pathways, by assuming that the evolution of a duplicate copy of a gene of an enzyme resulting in minimal changes in conformation could lead to a new protein catalyzing a similar type of reaction.

As an alternative to the retrograde evolution hypothesis /27, 29/, M. Ycas /51/, R. A. Jensen /43/, and R. A. Jensen and G.S. Byng /52/, also argued that primordial biosynthetic pathways

were assembled by the recruitment of slow, unefficient enzymes of broad substrate specificity. Such relatively unspecific early enzymes may have represented a mechanism by which primitive cells with small genomes could overcome their limited coding abilities /51, 52/. Such "patchwork assembly" of metabolic pathways by the assemblage of inefficient catalysts /45/ is consistent with the notion of a less accurate, error-prone primordial primitive translation apparatus synthesizing small "statistical" enzymes that could react with a relatively wide range of chemically related substrates /53/.

The Waley-Ycas-Jensen ideas are not mutually exclusive, since they are all based on the same basic idea: that the origin of extant metabolic pathways is to be found in a relatively limited variety of primordial genes that underwent amplification events followed by divergence. As reviewed by W.H. Li /54/, several hypothesis haven been proposed to account for the emergence of novel genes from redundant duplicates. According to the patchwork theory, new enzymes arise by gene duplication followed by divergence, which allows for the narrowing and modification of their substrate specificity. The traditional gradualist neoDarwinian evolutionary explanation is advocated by this hypothesis to describe the diversification of ancestral enzymes of broad specificity into families of related catalysts on the basis of point mutations, small deletions, and other minute genetic changes. It is not known how many changes are required to lead to the appearance of a new enzymes from a previous one, but directed mutagenesis studies suggest that in several cases very few genetic changes are actually required /58, 59/.

The recruitment of enzymes belonging to different metabolic pathways to serve novel biosynthetic routes is well documented under laboratory conditions. These are the so-called "directed evolution experiments", in which microbial populations are subjected to a strong selective pressure leading to heteretrophic phenotypes capable of using new substrates /55/. Extensive studies by R.P. Mortlock and his associates /56, 91/ have shown that under laboratory conditions prokaryotic catabolic evolution can lead to the use of new pentoses. The recruitment of previous enzymes to function in the new pathways typically followed the loss of existing repressive control by single mutations in regulator genes /44, 45, 55-57/.

The Waley-Ycas-Jensen model of metabolic evolution also implies that enzymes catalyzing similar biochemical reactions in different pathways must be homologous. It has been argued that this explanation underlies the observed homology between the amino acid sequences of the *Bacillus subtilis* threonine synthase and the tryptophan synthase ß chain from various organisms /60/. The phylogenetic analysis of different homologous tryptophan synthase ß chains from eubacteria, archaebacteria and eukaryotes /61, 62/ has shown that this protein is coded by a highly conserved gene that was probably already present in the last common ancestor of the three extant cellular lineages /39/. Several events of horizontal transfer of the tryptophan synthase ß gene may have taken place, resulting in a complex gene phylogeny /62/. However, distance matrix comparisons (not shown) between an extended databank of tryptophan ß chain and threonine synthase gene sequences have confirmed the suggestion /60/ that at least several small domains are shared between these two pyridoxal phosphate-dependent enzymes.

As noted by R. A. Jensen /63/, the patchwork theory is also supported by sequence comparisons that have demostrated that two different carbamoyl transferases involved in the biosynthesis of arginine and of pyrimidines have evolved from a common ancestor /64/. Biosynthesis of pyrimidines must be a very ancient pathway /39/, and although no sequences of archaebacterial carbamoyl transferases are currently available, it has been suggested that they may have resulted from paralogous duplications that took place prior to the evolutionary divergence of the three cellular lines /2/. Both the *E. coli argF-argI* gene pair coding for two homologous ornithine carbamoyl transferases (OTCases) involved in arginine biosynthesis, as

well as the *pyrB* gene encoding the catalytic monomer of aspartate carbamoyl transferase (ATCase), may be the result of a near-tandem duplication of the gene coding for an ancestral carbamoyl transferase endowed with substrate ambiguity /52, 64/.

Additional evidence supporting the Waley-Ycas-Jensen theory of enzyme evolution and assembly of metabolic pathways includes the following:

(a) sequence comparison analysis and functional homology studies that have shown that the *E. coli* acetohydroxy acid synthase (that participates in the first steps of the biosynthesis of branched-chain amino acids), and pyruvate oxidase (a flavoprotein dehydrogenase involved in the decarboxylation of pyruvate to acetate), are both derived from an ancestral flavin-dependent enzyme /65/;

(b) site-specific amino acid substitution experiments that have shown that the substrate specificity of an eubacterial NAD-dependent lactate dehydrogenase can be modified leading to a specific, highly active malate dehydrogenase simply by changing only one amino acid /58/;

(c) seven point mutations can alter the preference for NADPH as a cofactor of glutathione reductase, a dimeric flavin-dependent disulphide oxidoreductase /66/. Recent structure determinations of the human gluthatione reductase and *E. coli* thioredoxin reductase 3, as well as of other different enzymes belonging to this same family /67, 68/, have shown that these oligomeric enzymes are very similar, although they are involved in a wide spectrum of catalytic functions and biological roles /59/. Tertiary structure comparisons of these different enzymes strongly suggest that during their divergence from a common ancestor, small, simple alterations of the relative spatial orientation of their different domains resulted in major specificity changes and reconfigurations of their catalytic sites /59/; and

(d) in *B. subtilis* the sequences coding for the tryptophan biosynthetic pathway are organized in the gene cluster *trpEDCFBA*, a highly conserved operon organization that has been conserved in other prokaryotes /61/. The first enzyme in this pathway is anthranilate synthase, which is involved in the catalytic conversion of chorismatic acid and glutamine to anthranilic acid and pyruvate /93/. Anthranilate synthase in an oligomeric enzyme formed by two dissimilar subunits, α and β. The α subunit catalyzes the NH_3-dependent synthesis of anthranilate, and is coded by the *trpE* gene. The amide transfer capability is provided to the enzyme complex by its β subunit, and it is coded by the *trpG* gene, which is absent in the *B. subtilis trp* operon /62, 93/. Quite surprisingly, the amide transfer domain involved in tryptophan biosynthesis is coded by a gene from a separate metabolic pathway, that of folate biosynthesis. The first step in the conversion of chorismic acid to folic acid is the conversion of chorismate to *p*-aminobenzoic acid at the expense of glutamine. This reaction is quite similar to the first reaction of the tryptophan biosynthesis, the only difference being the orientation of the functional groups on the product, *para* in the folate pathway (*p*-aminobenzoic acid), and *ortho* in the tryptophan case (anthranilate aka *o*-aminobenzoic acid). Thus, the tryptophan and folate biosynthetic pathways provide an example of subunit shuffling, i.e., of a gene coding for at least one subunit of two different oligomeric enzymes participating in distinct but similar reactions in different pathways. .

ON POLYMERASE SUBSTRATE SPECIFICITY

Altough specificity is one of the long-cherished concepts of biochemistry, within certain limits different enzymes may use a broader range of related substrates than it is generally acknowledged. This is particularly true of both template and non-template dependent

polymerases. Polynucleotide phosphorylase (PNPase) is one such case. It is an enzyme apparently involved in RNA turnover, that can also catalyze the non template-dependent polymerization of ribonucleoside diphosphates /69/. However, in the presence of Mn^{++} PNPase substrate specificity is altered, leading to the formation of chains of more than ten oligodeoxyribonucleotides /70, 71/. Considerable evidence of the lack of strict substrate and template specificity of other different polymerases is also available, and has been summarized elsewhere /72, 73/. Metallic cations such as Mn^{++} are known to alter the substrate specificity of all major types of nucleic acid polymerase. However, the production of mixed-oligomers or of the wrong polymer has also been observed in the absence of cations that are known to alter the behaviour of viral and cellular RNA polymerases, DNA polymerases and DNA primases /72, 73/. Conclusive evidence of this phenomenon has been obtained by Konarska and Sharp /74/, who have described the replicase behaviour (i.e., RNA-directed RNA polymerization) of a viral DNA-dependent RNA polymerase.

Reverse transcriptases (RTs), i.e., the RNA-dependent DNA polymerases involved in the replication of different retroelements /75/ have also been shown to exhibit considerable substrate ambiguity, both *in vitro* and and *in vivo* /72/. It is generally assumed that RTs can only use deoxyribonucleotide substrates /92/. However, ribonucleotides have been found in the unintegrated linear spleen necrosis (SN) retroviral DNA, indicating that the SN retroviral-RT can incorporate them into a growing DNA chain /76/. This result is consistent with reports of minus-strand RNA molecules of different length in cells infected by tumor-producing retroviruses /77, 78/.

The lack of strict substrate discrimination is also true of different DNA-dependent DNA polymerases. Although the accuracy of cellular and viral DNA replication is a result of the high levels of DNA polymerases fidelity /79/, these enzymes are endowed with less specificity than it is usually acknowledged. This is true both *in vitro* and *in vivo* /72, 73/. As summarized by Mathews *et al.* /80/, in permeabilized phage T4-infected cells both deoxyribonucleoside monophosphates and ribonucleoside diphosphates are much more efficiently used by DNA polymerase than its normal, proper substrates, i.e., deoxyribonucleoside triphosphates (dNTPs). This observation, as well as additional evidence on (a) the temporal coincidence of dCTP and dTTP synthesis with viral replication /81/; (b) the intracellular interactions between DNA replication proteins and enzymes involved in dNTP biosynthesis /82/; and (c) the high concentration gradients of dNTPs in the vicinity of DNA replication forks /80/, support the idea that DNA precursor biosynthesis is closely coordinated with DNA replication itself /80, 83/.

The results summarized above suggests that although DNA polymerases exhibit considerable replication fidelity, they also have high levels of substrate promiscuity. Thus, the deoxyribonucleotide synthesizing apparatus (i.e., thymidylate synthase, ribonucleotide reductase, etc.) should be considered part of the bacterial enzymatic complex involved in DNA replication, i.e., the prokaryotic replisome /80/. Since the eukaryotic ribonucleotide reductase is located outside the nuclear membrane, dNTP channeling has been invoked to account for the coupling of deoxyribonucleotide biosynthesis with DNA replication, which takes place inside the nucleus /84/. However, it is likely that the replication of some eukaryotic viruses requires a direct physical association between the cellular ribonucleotide reductase and the viral replisome. This may be the case of retroviruses (including, of course, the infamous AIDS virus), whose biological cycle involves a RNA -> DNA transition in the cytoplasm of infected eukaryotic cells.

It has been argued that all template-dependent polymerases may share a common ancestor /85, 86/. However, the low specificity of polymerases suggests that the evolutionary transition from a primitive replicase into the extant DNA-dependent DNA polymerases did not require the early

appearance of a whole series of different enzymes. The evidence reviewed in this paper suggests that this conversion to different but chemically related genetic macromolecules, could have been mediated by a low-specificity nucleic acid polymerase, in which novel catalytic properties were acquired by few conformational changes. Since cellular DNA polymerases apparently have not developed strict substrate specificity, the observations summarized by C.K. Mathews and his colleagues /80/ also suggest that the coupling and close functional coordination of the ancestral ribonucleotide reductase with the enzymatic apparatus involved in nucleic acid replication was a major, decisive step in the evolutionary transition from RNA to DNA cellular genomes /87/.

CONCLUSIONS

As noted by S. Granick /88/, the Horowitz hypothesis establishes an evolutionary link between prebiotic chemistry and the development of biochemical phenomena. However, in this paper we have critically reviewed and summarized a number of observations and experiments that suggest that it fails to explain the observed characteristics of widely distributed metabolic pathways. In contrast, the available evidence suggest that the simple traditional neoDarwinian accumulation of random, minute genetic changes following early gene duplications advocated by the Waley-Ycas-Jensen hypothesis can be succesfully applied to describe the evolution of some ancient metabolic pathways like pyrimidine biosynthesis ina world in which catalytic proteins had already emerged. Since prokaryotes are haploid organisms with only a single set of genes, it is likely that most of these small genetic changes were expressed as soon as they arose, resulting in a rapid mode of metabolic evolution.

The existence of amphibolic enzymes in the tryptophan and folate anabolic pathways clearly exemplifies the existence of proteins catalyzing simlar reactions in completely separated pathways /93/. Detailed scrutiny has revealed that the amide transfer in the *B. subtilis* tryptophan biosynthesis is not the result of subunit specificity modifications due to the accumulation of small changes, but may have resulted from the loss of one member of a set of paralogous genes /62/. Nonetheless, it demostrates that enzymatic and metabolic ambiguity is present in extant bacteria, and may have occured in primitive prokaryotes.

Because of the lack of highly defined substrate and template specificity, early enzymes may have been multifunctional catalysts acting upon similar substrates. Retrogade evolution may have taken place in some cases (including parts of the methionine biosynthetic pathway), but it is likely that major metabolic innovations resulted from other evolutionary mechanisms. These include (a) the horizontal transfer of genes; (b) exon shuffling; (c) the reconfiguration of catalytic sites and substrate-binding domains by small spatial orientation changes in oligomeric enzymes; and (d) the patchwork assembly of enzymes, including subunit shuffling. As argued elsewhere /2/, the gene amplification events underlying the emergence of enzyme specificity /43, 50-52/ may have been paralogous duplications that took place prior to the evolutionary separation of the eubacterial, archaebacterial and eukaryotic lineages. Evidence of horizontal transfer of genes coding for mainstream metabolic enzymes such as glyceraldehyde-3-phosphate dehydrogenase /89/, glutamine synthase /90/ and the tryptophan synthase ß chain /62/, implies that strict criteria for the proper identification of Archean paralogous genes should be developed in order to avoid mistakes in the reconstruction of gene phylogenies. Since sets of paralogous genes provide important insights into the nature of metabolic pathways of cellular systems predating the last common ancestor of the two prokaryotic lineages /2/, more attention should be given to their study.

ACKNOWLEDGEMENTS

We are indebted to Drs. Gerda Horneck and Herrick Baltcheffsky for their patience and encouragement. We thank Monsieurs Alejandro Sosa-Peinado and Ervin Silva for providing us with several useful references. Work reported here has been supported in part by UNAM.IN 105289 (A.L.) and NAGW-2788 (J.O.).

REFERENCES

1. A., Lazcano. In *Early Life on Earth: Nobel Symposium No. 84*, ed. S. Bengtson, Columbia University Press, New York (in press)
2. A. Lazcano. In *Proceedings of the Sixth International Symposium in Microbial Ecology*, eds. R. Guerrero and C. Pedrós-Alios, Soc. Catalana de Biología, Barcelona (in press)
3. H. B. White III. *Jour. Mol. Evol.* **7**, 1 (1976)
4. R. Shapiro. *Origins -A Skeptics Guide to the Creation of Life*, Summit Books, New York (1986)
5. G. F. Joyce, Schwartz, A.W., Orgel, L.E. and Miller, S.L., *Proc. Natl. Acad. Sci. USA* **84**, 4398 (1987)
6. T. R. Cech and Bass, B.L. *Annu. Rev. Biochem.* **55**, 599 (1986)
7. A. A. Beaudry and Joyce, G.F. *Science* **257**, 635 (1992)
8. J. A. Piccirilli, McConnell, T.S., Zaug, A.J., Noller, H.F. and Cech, T.R. *Science* **256**, 1420 (1992)
9. H. F. Noller, Hoffarth, V. and Zimniak, L. *Science* **256**, 1416 (1992)
10. M. Eden. In *Mathematical Challenges to the NeoDarwinian Interpretation of Evolution*, eds. P.S. Moorhead and M.M. Kaplan,Wistar Institute Press, Philadelphia, p. 5 (1967)
11. A. Brack, and Barbier, B. *Adv. Space Res.* **9**, 83 (1989)
12. C. Shen, Lazcano, A., and Oró, J. *Jour. Mol. Evol.* **31**, 445 (1990)
13. J. Oró, Miller, S.L. and Lazcano, A. *Annu. Rev. Earth and Planet. Sci.* **18**: 317 (1990)
14. B. Alberts. *Am. Zool.* **26**, 781 (1986)
15. W. Gilbert. *Nature* **319**, 618 (1986)
16. A. Lazcano. *Treballs Soc. Catalana Biol.* **39**, 73 (1986)
17. M. Yarus. *Science* **240**, 1751 (1988)
18. W. Gilbert. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 901 (1987)

19. D. L. Hartl. *Genetics* **122**, 1 (1989)

20. C. C. C. Blake. *Nature* **273**, 267 (1978)

21. C. C. C. Blake. *Nature* **277**, 598 (1979)

22. W. Gilbert. *Nature* **271**, 501 (1978)

23. W. Gilbert. *Science* **228**, 823 (1985)

24. D. C. Phillips, Sternberg, M.J.E. and Sutton, B.J. in *Evolution from Molecules to Man*, ed. D.S. Bendall, Cambridge University Press, p. 145 (1983)

25. T. C. Südhof, Goldstein, J.L., Brown, M.S. and Russell, D. W. *Science* **228**, 815 (1985)

26. A. Lazcano. In *Environmental Evolution: Effect of the Origin and Evolution of Life on Planet Earth*, eds. L. Margulis and L. Olendzenski, MIT Press, Cambridge, Mass., p.57 (1992)

27. N. H. Horowitz. *Proc. Natl. Acad. Sci. USA* **31**, 153 (1945)

28. A. I. Oparin. *The Origin of Life*, Macmillan, New York (1938)

29. N. H. Horowitz. In *Evolving Genes and Proteins*, eds. V. Bryson and H.J. Vogel, Academic Press, New York, p. 15 (1965)

30. J. L. Betz, Brown, P.R., Smyth, M.J. and Clarke, P.H. *Nature* **247**, 261 (1974)

31. G. D. Hegeman and Rosenberg, S.L. *Annu. Rev. Microbiol.* **24**, 429 (1970)

32. C. Degami and Halmann, M. *Nature* **216**, 1207 (1967)

33. J. Oró. *Biochem. Biophys. Res. Comm.* **2**, 407 (1960)

34. S. L. Miller and Orgel, L.E. *The Origins of Life on Earth*, Prentice-Hall, Englewood Cliffs (1974)

35. Y. Yamagata, Sasaki, K., Takaoka, O., Sano, S., Inomata, K., Kanemitsu, K., Inoue, Y. and Matsumoto, I. *Origins of Life* **20**, 389 (1990)

36. J. P. Ferris and Joshi, P.C. *Science* **201**, 361 (1978)

37. J. P. Ferris, Joshi, P.C., Edelson, E. and Lawless, J.G. *J. Mol. Evol.* **11**, 293 (1978)

38. J. Belfaiza, Parsot, C., Martel, A., Bouthier de la Tour, C., Margarita, D., Cohen, G.N. and Saint-Girons, I. *Proc. Natl. Acad. Sci. USA* **83**: 867 (1986)

39. A. Lazcano, Fox, G.E. and Oró, J. In *The Evolution of Metabolic Function*, ed. R.P. Mortlock, CRC Press, Boca Raton, p. 237 (1992)

40. J. L. Cánovas, Ornston, L.N. and Stanier, R.Y. *Science* **156**, 1695 (1967)

41. L. N. Ornston. *Bacteriol. Rev.* **35**, 87 (1971)

42. H. Hartman. *J. Mol. Evol.* **4**, 359 (1975)

43. R. A. Jensen. *Annu. Rev. Microbiol.* **30**, 409 (1976)

44. T. T. Wu, Lin, E. C. C. and Tanaka, S. *J. Bacteriol.* **96**, 447 (1968)

45. P. H. Clarke. In *Evolution in the Microbial World*, eds. M.J. Carlile and J.J. Skehel, Cambridge University Press, Cambrigde, p. 183 (1974)

46. C. Yanofsky. *Mol. Biol. Evol.* **1**, 143 (1989)

47. E. Broda. *The Evolution of the Bioenergetic Processes*, Pergamon Press, Oxford (1975)

48. A. Lazcano. *El Origen de la Vida: Evolución Química y Evolución Biológica*, Trillas/ANUIES, México (1976)

49. L. Margulis. *Symbiosis in Cell Evolution*, Freeman Co., San Francisco (1981)

50. S. G. Waley. *Comp. Biochem. Physiol.* **30**, 1 (1969)

51. M. Ycas. *J. Theoret. Biol.* **44**, 145 (1974)

52. R. A. Jensen and Byng, G.S. in *Isosymes: Current Topics in Biological and Medical Research*, eds. M.C. Rattazi, J.G. Scandios and G.S. Whitt, Alan R. Liss Co., New York, p. 143 (1982).

53. C. R. Woese. Proc. *Natl. Acad. Sci. USA* **54**, 1546 (1965)

54. W. H. Li. In *Evolution of Genes and Proteins*, eds. M. Nei and R.K. Koehn, Sinauer Press, Sunderland, p. 14 (1983)

55. P. H. Clarke. In *Evolution from Molecules to Man*, ed. D.S. Bendall, Cambridge University Press, Cambridge, p. 283 (1983)

56. R. P. Mortlock and Gallo, M. A. in *The Evolution of Metabolic Pathways*, ed. R.P. Mortlock, CRC Press, Boca Raton, p. 1 (1992)

57. R. P. Mortlock. *Adv. Microb. Physiol.* **13**, 1 (1976)

58. H. M. Wilks, Hart, K.W., Feeney, R., Dunn, C.R., Muirhead, H., Chia, W.N., Barstow, D.A., Atkinson, T., Clarke, A.R. and Holbrook, J.J. *Science* **242**, 1541(1988)

59. G. A. Petsko. *Nature* **352**, 104 (1991)

60. C. Parsot. *Proc. Natl. Acad. Sci. USA* **84**, 5207 (1987)

61. W. L. Lam, Cohen, A., Tsouluhas, D. and Doolittle, W.F. *Proc. Natl. Acad. Sci. USA* **87**, 6614 (1990)

62. I. P. Crawford and Milkman, R. In *Evolution at the Molecular Level*, eds. R.K. Selander, A.G. Clarke and T.S. Whittam, Sinauer, Sunderland, p. 77 (1991)

63. R. A. Jensen. in *The Evolution of Metabolic Function*, ed. R.P. Mortlock, CRC Press, Boca Raton, p. 205 (1992)

64. F. Van Vliet, Cunin, R., Jacobs, A., Piette, J., Gigot, D, Lauwereys, M., Pierand, A. and Glansdorff, N. *Nucleic Acid Res.* 12, 6277 (1984)

65. Y. Y. Chang and Cronan, J.E., *Jour. Bacteriol.* 170, 3937 (1988)

66. N. S. Scrutton, Berry, A. and Perham, R.N. *Nature* 343, 38 (1990)

67. N. W. Schiering, Kabasch, M.J. Moore, M.D. Distefano, C.T. Walsh and E.F. Pai. *Nature* 352, 168 (1991)

68. J. Kuriyan, Krishna, T.S.R., Wong, L., Guenther, B., Pahler, A., Williams, C.H. and Model, P. *Nature* 352, 172 (1991)

69. M. Grunberg-Manago and Ochoa, S. *Jour. Am. Chem. Soc.* 77, 3165 (1955)

70. W. J. Hsieh. *Jour. Biol. Chem.* 246, 1780 (1971)

71. S. Gillam, Jahnke, P. and Smith, M. *Jour. Biol. Chem.* 253, 2532 (1978)

72. A. Lazcano, Valverde, V., Hernández, G., Gariglio, P., Fox, G.E. and Oró, J. *Jour. Mol. Evol.* 35, 524 (1992)

73. A. Lazcano, Llaca, V., Cappello, R., Valverde, V. and Oró, J. *Adv. Space Res.* 12, 207 (1992)

74. M. M. Konarska and Sharp, P.A. *Cell* 57, 423 (1989)

75. H. M. Temin. *Mol. Biol. Evol.* 2, 455 (1985)

76. I. S. Chen and Temin, H.E. *J. Virol.* 33, 1058 (1980)

77. N. Biswal and Benyesh-Melnick, M. *Proc. Natl. Acad. Sci. USA* 64, 1372 (1969)

78. E. Stavnezer, Ringold, G., Varmus, H.E. and Bishop, M. *Jour. Virol.* 20, 342 (1976)

79. T. A. Kunkel. *Jour. Biol. Chem.* 267, 18251 (1992)

80. C. K. Mathews, Moer, L.K. Wang, Y. and Sargent, G.R. *Trends Biochem. Sci.* 13, 394 (1988)

81. P. K. Tomich, Chiu, C.S., Wovcha, M.G., Greenberg, G.R. *Jour. Biol. Chem.* 249, 7613 (1974)

82. K. S. Cook, Wirak, D.O., Seashoti, A.F. and Greenberg, G.R. *J. Biol. Chem.* 263, 6202 (1988)

83. C. K. Mathews and Slabaugh, M.B. *Exp. Cell Res.* **162**, 285 (1986)

84. C. K. Mathews, North, T.W. and Reddy, G.P.V. *Adv. Enzyme Regul.* **18**, 133 (1979)

85. A. Lazcano, Fastag, J., Gariglio, P., Ramírez, C. and Oró, J. *Jour. Mol. Evol.* **27**, 365 (1988)

86. M. Delarue, Poch, M., Tordo, N., Moras, D. and Argos, P. *Protein Engineering* **3**, 461 (1990)

87. A. Lazcano, Guerrero, R., Margulis, L. and Oró, J. *Jour. Mol. Evol.* **27**, 283 (1988a)

88. S. Granick. *Annals New York Acad. Sci.* **69**, 292 (1957)

89. R. F. Doolittle, Feng, D.F., Anderson, K.L. and Alberro, M.R. *Jour. Mol. Evol.* **31**, 383 (1990)

90. M. W. Smith, Feng, D.F. and Doolittle, R.F. *Trends Biochem. Sci.* **17**, 489 (1992)

91. R. P. Mortlock, Fossitt, D.D. and Wood, W.A. *Proc. Natl. Acad. Sci. USA* **54**, 572 (1965)

92. I. M. Verma. *Biochem. Biophys. Acta* **473**, 1 (1977)

93. I. P. Crawford. *Annu. Rev. Microbiol.* **43**, 567 (1989)

MICROBIOLOGÍA SEM **11** (1995), 185–198

Cellular evolution during the early Archean: what happened between the progenote and the cenancestor?

Antonio Lazcano

Laboratorio de Microbiología, Departamento de Biología, Facultad de Ciencias, Universidad Nacional Autónoma de México, México, D.F., México

Summary

Although cladistic techniques cannot be applied to the understanding of the origin of life itself, at the time being the comparison of macromolecules is not only the most powerful tool for inferring the branching order of the three cell lineages, but also for providing some insights into the nature of the biological systems that preceded their last common ancestor. It is argued that this information cannot be extrapolated to support the hypothesis that the first living systems were hyperthermophiles that emerged in deep sea vents or in other extreme environments. The significance of detecting and characterizing paralogous genes that duplicated before the divergence of the last common ancestor of all extant life to understand some of the mechanisms that led to the establishment of biochemical pathways is also discussed.

Key words: cellular evolution, Archean times, molecular cladistics, progenote, cenancestor

Resumen

A pesar de que las técnicas cladísticas no son por sí mismas aplicables al conocimiento del origen de la vida, en la actualidad la comparación de macromoléculas no es sólo la herramienta más potente para determinar el orden de divergencia de los tres linajes celulares, sino también para proporcionar algunas indicaciones sobre la naturaleza de los sistemas biológicos que precedieron a su antepasado común. Esta

Correspondence to: Antonio Lazcano. Departamento de Biología. Facultad de Ciencias. Universidad Nacional Autónoma de México. Ciudad Universitaria. Apartado 70-407. México, D.F. 04510. México. Tel.: +52-5-6224823. Fax: +52-5-6160451.

información no puede extrapolarse para fomentar la hipótesis de que los primeros sistemas vivos fueron hipertermófilos que surgieron en los "deep sea vents" o en otros ambientes extremos. En este artículo se discute también la importancia que tiene la detección y caracterización de genes parálogos, que se duplicaron antes de la divergencia del último antepasado común, para el conocimiento de algunos mecanismos que condujeron al establecimiento de rutas bioquímicas básicas.

Introduction

Although molecular techniques have been used in phylogenetic studies since the turn of the century (44), it was not until much latter that it was fully realized that protein and nucleic acid sequences are historical documents (73) that contain an extraordinarily rich amount of evolutionary information of unsurpassed value whose retrieval has led to several major conceptual revolutions in contemporary biology. However, it is also true that this approach has remarkable limitations. Several attempts have been made to extrapolate the results of macromolecular comparisons back into the stages in which the basic characteristics of living systems were first established (12, 60). Nevertheless, these endeavours have been hindered by our almost complete ignorance of the nature of the first living systems and thus they lack, as argued below, evolutionary significance. Indeed, it is becoming increasingly clear that, although the development of molecular phylogeny has greatly deepened our understanding of the early stages of cellular evolution, the complex—and perhaps unfathomable—issue of the origin of life is not amenable in itself to cladistic analysis.

Of course, the significance of molecular cladistics in the discovery of the three major cell lineages, i.e., the eubacteria, the archaeobacteria, and the eukaryotic nucleocytoplasm (69) cannot be underscored. As argued throughout this paper, the evolutionary analysis of the available databases can provide important insights not only on the nature of their last common ancestor, but also on the biological events that preceded it, during which some of the essential traits of basic metabolic pathways were shaped (71). Therefore, the purpose of this paper is to summarize our current knowledge of an early stage of cellular evolution that took place before the diversification events of the last common ancestor of all extant life forms, but after the appearance of biological systems, very likely of cellular nature, already endowed with a genetic code and a ribosome-mediated protein synthesis.

Progenotes, cenancestors and molecular cladistics

Although attempts to find the proper place of microbes in phylogenetic trees date back to the work of Ernst Haeckel and other 19th century scientists, it was not until much latter that the first detailed theories on prokaryotic evolution and classification where first suggested. On the basis of morphological features, Kluyver and van Niel (32) suggested that three different lines of descent formed by the spirilla, the sporulating Gram-positive bacilli, and the high actinobacteria, respectively, were descendants of primitive coccoidal bacteria. That same year A. I. Oparin, on the basis of a detailed comparison of the basic biochemical processes and energy-generating metabolic pathways, published the Russian edition

of his book on the origin of life, in which he suggested an evolutionary scheme that begun with anaerobic heterotrophy and proceeded, in a gradual, stepwise evolutionary process that eventually led to oxygen-producing photosynthetic cyanobacteria (45).

Although the points of view suggested by Oparin have an enormous heuristic value, which eventually led to the establishment of an entire field of scientific research devoted to the scientific study of the origin of life, it is no longer possible to assume that the first living system was a *Clostridium*-like, anaerobic fermenter (55) or a *Mycoplasma*-type of prokaryote (47, 67). The nature of the first living forms is still an open question, but as summarized by Woese (69), molecular cladistic studies have shown that simplicity is not equal to primitiveness: the wall-less aphragmobacteria phenotype is polyphyletic and has evolved independently both among the eubacteria and the archaeobacteria, while the anaerobic relatively simply clostridial lifestyle is part of the Gram-positive, low GC branch that is very far away from the oldest eubacterial phenotypes.

It is now generally accepted that the development of molecular cladistics has led to major changes in our current understanding of microbial evolution. A major achievement of this approach has been the use of small subunit ribosomal RNA (rRNA) sequences as phylogenetic markers. The advantages of using these phylogenetic markers are: (i) the universal distribution of their genes in all known organisms and organelles of cellular origin; (ii) the fact that they always serve the same function; (iii) their improbable lateral gene transfer; and (iv) the relatively slowly change in primary sequence compared to other molecular clocks and may be thus used for the construction of deep phylogenies (69).

It is frequently forgotten that the evolutionary comparison of 16S rRNA-like genes allows the construction of cladograms depicting the phylogeny of genes, but not of organisms. However, the properties of rRNA genes as phylogenetic markers make them valuable instruments, which have allowed the construction of a trifurcated, unrooted tree in which all known organisms can be grouped in one of the three major cell lineages: the eubacteria, the archaeobacteria, and the nucleocytoplasmic component of eukaryotes (69). Since in unrooted rRNA-based cladograms no single major branch predates the other two, and all three derive from a common ancestor, it was concluded that the latter corresponded to an ancestral form of life much simpler than contemporary prokaryotes, i.e., the progenote (70). This hypothetical entity was defined as a primitive system with a rudimentary translation machinery, in which phenotype and genotype still had an imprecise, rudimentary linkage relationship (70). A model for the progenote described it as endowed with a fragmented, disaggregated genome formed of double-stranded RNA genes, many of which could have existed in multiple copies (68). Independent evolutionary refinements along the three lines of descent not only led into its aggregation into DNA genomes, but led also to the differences found among the transcription and translation machineries of eubacteria, archaeobacteria and eukaryotes after their divergence from their last common ancestor (69).

No outgroups are known for rRNA based phylogenies, which specify branching relationships but not the position of the universal ancestor. Therefore, it is not possible to identify in them the oldest cellular phenotype. Nevertheless, speculations on the antiquity of a trait may be justified on empirical generalizations based both on the trait's essential role and on its wide distribution. Thus, a partial description of the last common ancestor of the three main branches may be infered from the distribution of homologous characters among its descendants, i.e., by comparing eubacteria, archaeobacteria and eukaryotes and see which monophyletic traits are common to the three of them. It can be argued that any feature found in all three lines was probably present in the ancestral organism from which they are

derived, i.e., that genes present in the main branches of the universal rRNA tree which are not the result of horizontal transfer must have been also present in their evolutionary progenitor.

As shown in Table 1, the set of such traits that have been identified as of 1994 is still small, but the picture of the last common ancestor that can be constructed is that of a rather sophisticated cell. As suggested by the presence of genes involved in major anabolic processes that include the biosynthesis of purines, pyrimidines, coenzymes, arginine, tryptophan, histidine and the branched-chained amino acids, i.e., valine, isoleucine, and leucine (Table 1), its biosynthetic abilities were comparable to those of modern cells. The occurrence of insulin-like peptides and cAMP in the three cell lines suggests that signalling molecules involved in intracellular and cell-to-cell communication had already appeared in their last common ancestor, as well as heat-shock proteins and other molecules that may have been involved in responses to environmental insults and stress conditions.

TABLE 1. Homologous traits common to the three cellular domains[a]

(i) Traits involved in replication and protein biosynthesis	
DNA polymerase B*	elongation factor 1α/Tu*
Gyrase B	elongation factor G/2*
DNA topoisomerase II	isoleucyl-tRNA synthetase*
RNA polymerases*	ribonuclease P
polynucleotide phosphorylase	ribosomal proteins S9, S10, S17, S15, L2, L3, L6, L10, L11, L22, and L23
(ii) Traits involved in energy generation processes and in biosynthetic pathways	
F-type ATPase α subunit*	arginosuccinate synthetase
F-type ATPase β subunit*	aspartate aminotransferase*
carbamoyl-phosphate synthetase*	cytrate synthetase
glucose 6-phosphate dehydrogenase	enolase
glutamate dehydrogenase II*	glutamine synthetase
malate dehydrogenase*	phosphoglycerate kinase
pyruvate: ferrodoxin oxidoreductase	porphobilinogen synthase
histidinol phosphate aminotransferase	purine biosynthetic genes
tryptophan biosynthetic genes	branched-chain amino acid biosynthetic genes
(iii) Traits involved in environmental response and chemical signalling	
cAMP	
insulin-like polypeptides	
heat shock protein 70	
Mn/Fe superoxide dismutases*	
photolyases	

[a] Based on Lazcano et al. (38), Benner et al. (7), and Doolittle and Brown (13).

* Paralogous duplicate.

The structural similarities shared by the proteins found in all three lines of descent suggest that considerable fidelity already existed on the then operative genetic system of their last common ancestor, which might have been already based on double-stranded DNA molecules (38). This conclusion is supported by the analysis of the available sequence databases, as well as by information derived from antibiotic sensitivity and antibody response, which suggests that the ancestor of the three lines already encoded oligomeric RNA polymerases, DNA topoisomerases, DNA polymerases with proof-reading activity (17), and photolyases involved in the monomerization of UV-pyrimidine dimers (62). Thus, it was already endowed with mechanisms assuring both high-fidelity replication and repair of UV-induced DNA damage, which would be extremely valuable in the anoxic primitive environment.

Although the presence of superoxide dismutase in *Methanobacterium thermoautotrophicum* (63), and of cytochrome oxidase in *Sulfolobus acidocaldarius* (8) raises the troublesome possibility that ancient organisms found at the base of the archaeobacterial branch had aerobic traits, these two enzymes are in fact part of defense mechanisms that may have evolved once oxygen had accumulated in the primitive atmosphere. In fact, evidence that anaerobiosis is an ancient trait is supported not only by the universal distribution of at least some of the glycolytic enzymes found in mainstream heterotrophic anaerobic metabolic pathways (18), but also by the identification of anaerobic eubacterial ribonucleotide reductase III as the oldest of these enzymes involved in the synthesis of deoxyribonucleotides (49). This finding suggests that the RNA to DNA evolutionary transition (39) took place in an oxygen-poor primitive environment.

The traits shared by the three main cell lines (Table 1) are far to numerous and complex to assume that they have evolved independently, i.e., they are of polyphyletic origin, or that they are the result of massive horizontal transfer. Thus, although inferences on early life may be hindered by cell fusion events or lateral flow of genetic sequences (20, 24, 59), the data summarized in Table 1 not only suggests that both in basic organization of the genetic apparatus and in its metabolic abilities eubacteria, archaeobacteria and the eukaryotic nucleocytoplasm are ultimately related and the three lines descend from a common ancestor, but also that the latter was not a protocell or any other pre-life progenitor system. It is likely that the last common ancestor of all known forms of life was in fact comparable to modern bacteria in its biological complexity, ecological adaptability, and evolutionary potential. Accordingly, the original definition of progenote cannot be used for the genetic entity from which the eubacterial, archaeobacterial, and eukaryotic branches diverged. Progenotes, if they ever existed, must have become extinct long before the separation of the three lineages, whose last common ancestor may be more appropiately described by using the term *cenancestor*, a neologism coined by Fitch and Upper (16) using a Greek prefix that can be translated both as *last* and as *common*. As discussed below, partial understanding of the processes that took place before the appearance of cenancestor may be achieved by analysing the sequences of genes that duplicated before the separation of the three major branches.

A hot origin of life?

The results summarized in the previous section indicate that the most basic questions pertaining to the origin of life relate to much simpler entities predating by a long series of evolutionary changes the cenancestor, i.e., the earliest ancestor that we can detect using rRNA-based phylogenetic trees. As noted

above, the identification of ancestral conditions is not possible for unrooted rRNA cladograms because there is no known organism that can be used as an outgroup. However, such problem can be overcomed by making outgroup comparisons using paralogous genes, i.e., genes that diverged after a duplication event and not through speciation (15), by using one set of paralogous genes as an outgroup for the other set (55).

This technique was employed a few years ago by Iwabe et al. (27) and by Gogarten et al. (21), who identified the sets that code for (i) the elongation factors (EF–G, EF–Tu) that assist in protein biosynthesis; and for (ii) the α and β hydrophilic components of F-type ATP synthethases as paralogous duplicates. Using different computing techniques, both groups independently placed the root of the universal tree between the eubacteria, on the one side, and the archaeobacteria and eukaryotes on the other. This result not only implies that the eubacteria are the oldest recognizable cellular phenotype, but also that a significant portion of the eukaryotic nucleocytoplasm is derived from archaeobacteria.

Although the eubacterial rooting of universal phylogenetic trees has been disputed and the problem remains open as one of the most challenging issues in evolutionary biology (6, 13, 17), the hypothesis that eukaryotes and archaeobacteria are sister groups is in fact supported by a number of molecular traits that are shared by their transcription machineries (30). Placing the root of universal trees at the eubacterial lineage has also groups all the known hyperthermophiles at the base of the two prokaryotic branches of rRNA trees (1, 60). This confirms previous suggestions that mesophilic eubacteria and archaeobacteria are the descendants of ancient hyperthermophiles, i.e., of organisms that grow optimally at temperatures in the range 75–100°C.

Perhaps not surprisingly, the phylogenetic distribution of heat-loving bacteria has been interpreted to support those advocating a hot origin of life. Such ideas are not new: they were discussed with considerable detail in the late 19th century by the German chemist E. Pflüger as part of his HCN-based theory on the origin of life (46). More recently, the recognition that some cyanobacteria have heat-tolerant modes of life led to Harvey (23), Copeland (10) and Scher (52) to defend the possibility of a heterotrophic emergence of life in high-temperature environments. Modern equivalents of these ideas suggest that life appeared in geothermically heated, high-temperature environments such as those found today in deep sea vents (26), or in other sites in which mineral surfaces have been hypothesized to have played a major role in the appearance of primordial chemolithoautotrophic biological systems (11, 29, 66).

However, several objections can be raised against this possibility. First of all, a high temperature regime for the origin of life would rapidly lead to an irreversible destruction of organic compounds, and thus to a very short lifetime for amino acids, purines, pyrimidines, and other biochemical compounds that are generally assumed to have been essential for the first organisms (42). It is also difficult to reconcile the possibility of an RNA world with the hot-origin-of life hypothesis (7, 35). Although the significance of ribozymes in the emergence of the biosphere is still an open question, the presence of the 2'–OH group in ribose makes RNA an extremely thermolabile polymer that is much more sensitive to cleavage than DNA at the temperatures typical for hyperthermophiles (40).

The claim that the phylogenetic distribution of hyperthermophiles supports the hypothesis that life emerged in a hot, sizzling environment (29) is in fact based on the unwarranted assumption that the root of universal evolutionary trees based on macromolecules can be extended back in time down to prebiotic

epochs. However, this may be a premature conclusion; although it is true that from a cladistic viewpoint a characteristic state found only in the deepest branches can be interpreted as primitive (61), no species exists today with all traits in the ancestral state. Indeed, heat-loving bacteria share with all other known organisms the same basic features of genome replication, gene expression, ATP-based energy producing mechanisms, and basic biosynthetic pathways. Thus, hyperthermophiles are cladistically ancient organisms, not primitive ones. This conclusion implies, of course, that a heat-loving lifestyle may be a relic of early Archean times, i.e., a secondary adaptation that evolved in population of even older mesophilic bacteria (9, 34), perhaps as a result of high temperature regimes that may have resulted from major asteroidal impacts during the late bombardment period that characterized the final stages of accretion of our planet (57) or by other, still uncharacterized selection pressure.

The confirmation of the hypothesis that hyperthermophily is indeed an ancient secondary adaptation requires an understanding of the nature of the biochemical adaptations involved in the adaptation to extreme temperatures. Although this is still a largely open question, it appears to depend on a wide spectrum of different mechanisms, which may include histone-like proteins, numerous post-transcriptional RNA modifications, high intracellular salt concentrations, multienzyme complexes protecting small intermediary metabolites, DNA-binding proteins, polyamines, and reverse gyrase, a type I topoisomerase that twists DNA into a positive supercoiled double-stranded chain (2, 9, 48, 56). Molecular cladistic analysis and a detailed understanding of the role that these and other traits play in a heat-tolerant lifestyle are required before the significance of the basal position of hyperthermophiles in phylogenetic trees is fully assessed. Until such analysis are accomplished, it is premature to take the possibility of a hot origin of life for granted..

Biological evolution before the cenancestor

Despite minor differences, the universal distribution of molecular biology processes not only provide direct evidence of the monophyletic origin of all known forms of life; they also imply that the sets of genes encoding the different components of these complex traits became frozen long time ago, i.e., major changes in them are lethal and very strongly selected against. However, it is clear that these complex, multigenic traits must have evolved through a series of simpler states. Unfortunately, no evolutionary intermediate stages or ancient simplified versions of ATP production, DNA replication and ribosome-mediated protein synthesis have been discovered in extant organisms. The absence of any known cladistically primitive taxa that originated before the freezing of these biological processes is indeed a major obstacle in the reconstruction of the sequence of the evolutionary development of Archean cells.

However, there is a way of inferring some of the characteristics of the primitive entities from which the cenancestor evolved (36). As noted above, the presence of paralogous genes implies that the last common ancestor of the three cell lineages was already endowed with two homologous genes coding for two elongation factors, as well as with F-type ATPases having homologous α and β subunits (21, 27). Accordingly, if the cenancestor had two sets of duplicate homologous genes coding for elongation factors and for the ATP-synthethase units, then it must have been preceded by a simpler cell with a

smaller genome in which only one copy of each of these genes existed, i.e., by cells in which protein synthesis required the presence of only one elongation factor, and with ATPases that lacked the α regulatory subunit (36).

Although the list of such precenancestral paralogous genes is still small (Table 1), it includes, in addition to the elongation factors and the ATPase subunits, hexameric glutamate dehydrogenases (6), glutamine synthethases (31, 64), the heat-shock protein family (22), and DNA topoisomerases I and II and DNA polymerase families A and B (17), as well as the large subunit of carbamoyl-phosphate synthase (CPSase), a protein formed by two homologous halves that resulted from an internal (i.e., partial) duplication followed by a gene fusion event (53; Lazcano, Puente, and Gogarten, in prep.). Additional products of gene pairs that may have duplicated during early Archean times have been summarized elsewhere, and include those coding for superoxide dismutases, carbamoyl transferases, dehydrogenases, pyruvate oxidase and acetohydroxyacid synthase, and several aminoacyl-tRNA synthethases (19).

In other words, detection and analysis of paralogous sequences common to the three lines are a potential source of evolutionary information which may provide direct insights to the organization and encoding capacities of genetic systems predating the cenancestor. Such paralogous sequences whose duplication preceded the cenancestor imply that before these gene amplification events, simpler living systems existed that lacked at least some of the complex regulated biochemical processes found in extant cells. That is, the cenancestor was the descendant of earlier life forms in a genetic code, and ribosome-mediated protein biosynthesis already existed, but in which (i) the large subunit of CPSase had half the molecular weight of its modern equivalent; (ii) protein biosynthesis could take place with only one elongation factor; (iii) F-type ATPases lacked the α regulatory subunit, and (iv) the DNA replication and repair machineries involved one only DNA polymerase ancestral to DNA polymerase I and II.

Precenancestral cells must have been less complex than even the simplest extant life forms, lacking the large set of enzymes and some of the sophisticated regulatory abilities of contemporary prokaryotes. Contemporary equivalents are not known, perhaps because they have been completely obliterated by their more successful descendants. However, the available databases provide direct evidence of Archean cells in which biological functions were apparently dependant on primitive, less-complex enzymes. This conclusion may be interpreted as supporting the hypothesis that early biosynthetic pathways were mediated by small, inefficient enzymes of broad substrate specificity (14, 21, 28).

Gene duplication and early cellular evolution

How was the extant set of genetic sequences built from an earlier, simpler genome? Evolution of early Archean microbes must have required important increases of their genome sizes, i.e., a major expansion of their coding abilities. The different mechanisms that may modify cellular DNA content in contemporary organisms are shown in Fig. 1. Their relative significance in prokaryotic genome size evolution should be analysed. For instance, although endosymbiosis has been a major driving force in the emergence and evolution of eukaryotic cells (41), it is unlikely that it ever played a major role in shaping prokaryotic evolution.

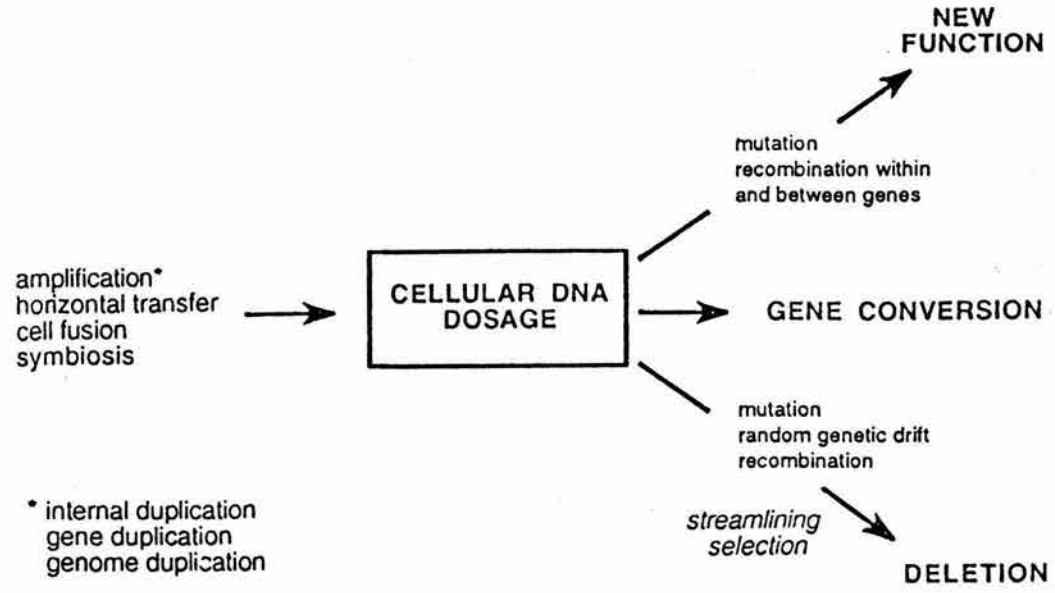


FIG. 1. Mechanisms that increase cellular DNA content and their possible evolutionary fate (43).

Moreover, cell fusion events are well-documented in *Oxytrichia* and other single-cell protists, but have not been described in bacteria. Nevertheless, it has been suggested that they may have taken place during early Archean times, thus providing with an explanation of the discrepancies between the 16S rRNA phylogeny and the protein-based trees (20, 59, 72). Besides the observation that multichromosomal mutant *Escherichia coli* cells can be obtained by blocking different stages during cell division (5), could be interpreted to support the possibility of dramatic increases in DNA content resulting from a series of bacterial genome doublings (50).

The role of horizontal transfer of genes in the expansion of the coding abilities of Archean cells should not be overlooked. The nitrogen-fixing eubacteria *Azotobacter vinelandii* is endowed with multiple copies of rather large plasmids that have increased its DNA content by a factor of 40 as compared to *Escherichia coli* (25). Evidence of lateral acquisition of genes can be recognized among different prokaryotes (58), and some of these events may have taken place shortly after the diversification of the three cell lineages (24).

However, there are several independent indications that gene duplications played a decisive role in the evolutionary development of the encoding capabilities of ancient genomes. This possibility is supported not only by the products of duplications detectable in the three cell lineages (Table 1), but also by the statistical analysis of the *Escherichia coli* sequence databases, that has shown that about 40% of the proteins whose sequence is now available are the result of duplication events (33).

If duplication was one of the major forces that shaped shaping ancient genomes, then it may be possible to understand the rapid development of levels of biochemical complexity and ecological diversity suggested by morphological and isotopic evidence showing that stromatolite-building phototactic

bacteria already existed 3.5×10^9 years ago (54). The age and complexity of these early remnants of Archean life not only suggests that the basic features of DNA-based cellular genomes had been established long before the deposition of these early fossils, but also that the major features separating the archaeobacterial and eubacterial branches were already established during the early Archean times.

Although there are many uncertainties surrounding the origin of life and the early evolution of the biosphere, the possibility that duplication events were the most important mechanism for increasing the size and complexity of prokaryotic genomes allows an estimate of the time required for the emergence of an oscillatorian-like cyanobacteria similar to the morphotypes discovered in the Warrawoona assemblage (37). Duplication events appear to occur spontaneously at relatively high, constant rates of 10^{-5} to 10^{-3} gene duplications per gene per cell generation both among eubacteria and eukaryotes (3, 51, 65). By using the lowest value of 10^{-5}, and by assuming that only 10% of the duplications are neutral, a rate of duplicon accretion of one nucleotide pair per year has been estimated, which implies that only seven million would be required to go from a 100 gene DNA/protein organism to a 7000 genes filamentous cyanobacteria (37). Of course, many of these figures are ridden by a large number of uncertainties, but they suggest that there is no compelling reason to assume that the entire process required more than 10 million years or so (37, 43). It is likely that the assumption that the emergence of life was an extremely slow process is nothing more than a deeply-rooted intellectual prejudice whose origin may be found in some of the most conservative traditions of neoDarwinism. As a matter of fact, the understanding of the mechanisms that may help to understand why the origin and evolution of early life took place in a relatively short span of geological time should be combined with the analysis of the processes underlying the lenghty periods of evolutionary stasis during which the emergence of metabolic novelties in different prokaryotic lineages has been limited or inhibited.

Concluding remarks

Because of its very nature, molecular cladistics separates clusters of adaptive characters into a nested hierarchical set which is generally expected to reflect the temporal sequence of their evolutionary acquisition. It is not surprising that such approach, which has all the demerits of a reductionist one-trait approach to biological evolution, has also led to incomplete description of cellular evolution. This limitation may be particularly clear in the failure molecular trees to include branch fusion events (i.e., anastomosis of lineages) that can describe eukaryotes as highly integrated components of evolutionary consortia, but also in the unjustified attempts to extrapolate the root of molecular phylogenies into the origin of life itself or even before (4).

However, the evidence reviewed in this paper suggests that although the analysis of macromolecular sequences cannot be extended back into prebiotic times, it is a powerful tool whose full potential may have not been fully realized. In particular, the identification of paralogous genes that duplicated before the divergence of the cenancestor can provide major insights into the nature of biological processes whose characteristics cannot be infered from the palaeontological record or from the other traditional approaches. Recognition of this possibility implies that what have been calling the root of universal trees corresponds in fact to the tip of their trunks, and in order to obtain insights into primitive cells, we must learn to read the valuable evolutionary information still contained in them.

The evidence that relevant information concerning biochemical characteristics of cells older than the three domains may be derived from ancestral paralogous genes is persuasive, and major attention should be devoted to its retrieval and interpretation. Accordingly, the design of a research strategy for the identification, sequencing and evolutionary comparison of sets of paralogous genes that originated before the separation of eubacteria, archaeobacteria and eukaryotes should be considered a major priority in our efforts to understand the evolutionary history of ancient cells, and could help to reduce in part the gap that exists in current descriptions of the evolutionary transition between the RNA world, the progenote stages, and the last common ancestor of all extant organisms.

Acknowledgments

I am indebted to Professors Lynn Margulis, Monica Riley, and Stanley L. Miller for many useful discussions. I thank Drs. Ford Doolittle, Renato Fani, Patrick Forterre, J. Peter Gogarten, and their coauthors, for providing me with copies of their results prior to publication.

References

1. Achenbach-Richter, L., Gupta, R., Kandler, K. O., Woese, C. R. (1987). Were the original eubacteria thermophiles? System. Appl. Microbiol. **9**, 34–39.
2. Adams, M. W. W. (1993). Enzymes and proteins from organisms that grow near and above 100°C. Annu. Rev. Microbiol. **47**, 627–658.
3. Anderson, R. P., Roth, J. R. (1977). Tandem genetic duplications in phage and bacteria. Annu. Rev. Microbiol. **31**, 473–505.
4. Becerra-Bracho, C., Silva, E., Velasco, A. M., Lazcano, A. (1995). Molecular biology and the reconstruction of microbial phylogenies: des liaisons dangereuses? *In* Collado, J., Smith, T., Magasanik, B. (ed.), Integrative Approaches to Molecular Biology. MIT Press, Cambridge, MA, in press.
5. Begg, K. J., Donachie, W. D. (1991). Experiments on chromosome separation and positioning in *Escherichia coli*. New Biol. **3**, 1–11.
6. Benachenhou-Lahfa, N., Forterre, P., Labedan, B. (1993). Evolution of glutamate dehydrogenase genes: evidence for two paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life. J. Mol. Evol. **36**, 335–346.
7. Benner, S. A., Cohen, M. A., Gonnet, G. H., Berkowitz, D. B., Johnsson, K. P. (1993). Reading the palimpsest: contemporary biochemical data and the RNA world. *In* Gasteland, R. F., Atkins, J. F. (ed.), The RNA World, pp. 27–70. Cold Spring Harbor Lab. Press, Cold Spring Harbor, NY.
8. Castresana, J., Lübben, M., Saraste, M., Higgins, D. G. (1994). Evolution of cytochrome oxidase, and enzyme older than atmospheric oxygen. EMBO J. **13**, 2516–2525.
9. Confalonieri, F., Elie, C., Nadal, M., Bouthier de la Tour, C., Forterre, P., Duguet, M. (1993). Reverse gyrase: a helicase-like domain and a type I topoisomerase in the same polypeptide. Proc. Natl. Acad. Sci. USA **90**, 4753–4758.
10. Copeland, J. J. (1936). Thermophilic microorganisms. Ann. New York Acad. Sci. **69**, 328–335.
11. Danchin, A. (1990). Homeotopic transformation and the origin of translation. Prog. Biophys. Molec. Biol. **54**, 81–86.

12. Dayhoff, M. O. (1969). Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Spring, MD.
13. Doolittle, W. F., Brown, J. R. (1994) Tempo, mode, the progenote and the universal root. Proc. Natl. Acad. Sci. USA **91**, 6721–6728.
14. Fani, R., Liò, P., Lazcano, A. (1995). Molecular evolution of the histidine biosynthetic pathway. J. Mol. Evol., in press.
15. Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. Syst. Zool. **9**, 117–133.
16. Fitch, W. M., Upper, K. (1987). The phylogeny of tRNA sequences provides evidence of ambiguity reduction in the origin of the genetic code. Cold Spring Harbor Symp. Quant. Biol. **52**, 759–767.
17. Forterre, P., Benachenhou-Lahfa, N., Confalonieri, F., Duguet, M., Elie, C., Labedan, B. (1993). The nature of the last universal ancestor and the root of the tree of life, still open questions. BioSystems **28**, 15–32.
18. Fothergill-Gilmore, L. A., Michels, P. A. M. (1993). Evolution of glycolysis. Prog. Biophys. Molec. Biol. **59**, 105–235.
19. García-Meza, V., González-Rodríguez, A., Lazcano, A. (1994). Ancient paralogous duplications and the search for Archean cells. *In* Fleischaker, G. R., Colonna, S., Luisi, P. L. (ed.), Self-Reproduction of Supramolecular Structures: from Synthetic Structures to Models of Minimal Living Systems, pp. 231–246. Klüwer Academic Press, Dordrecht, Netherlands.
20. Gogarten, J. P. (1994). Which is the most conserved group of proteins? Homology-orthology, paralogy, xenology, and the fusion of independent lineages. J. Mol. Evol. **39**, 541–543.
21. Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, J., Date, T., Oshima, T., Konishi, L., Denda, K., Yoshida, M. (1989). Evolution of the vacuolar H^+-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA **86**, 6661–6665.
22. Gupta, R. S., Singh, B. (1992). Cloning of the HSP70 gene from *Halobacterium marismortui*: relatedness of archaebacterial HSP70 to its eubacterial homologs and a model of the evolution of the HSP70 gene. J. Bacteriol. **174**, 4594–4605.
23. Harvey, R. B. (1924). Enzymes of thermal algae. Science **LX**, 481–482.
24. Hilario, E., Gogarten, J. P. (1993). Horizontal transfer of ATPase genes—the tree of life becomes a net of life. BioSystems **31**, 111–119.
25. Holloway, B. W. (1993). Genetics for all bacteria. Annu. Rev. Microbiol. **47**, 659–683.
26. Holm, N. G. (ed.) (1994). Marine Hydrothermal Systems and the Origin of Life. Klüwer Academic Press, Dordrecht, Netherlands.
27. Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., Miyata, T. (1989). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes infered from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA **86**, 9355–9359.
28. Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. Annu. Rev. Microbiol. **30**, 409–427.
29. Kandler, O. (1994). The early diversification of life. *In* Bengtson, S. (ed.), Early Life on Earth: Nobel Symposium No. 84, pp. 111–118. Columbia University Press, New York, NY.
30. Klenk, H.-P., Doolittle, W. F. (1994). Archaea and eukaryotes versus bacteria. Curr. Biol. **4**, 920–922.
31. Kumada, K., Benson, D. R., Hillemann, D., Hosted, T. J., Rochford, D. A., Thompson, C. J., Wohlleben, W., Tateno, T. (1993). Evolution of the glutamine synthase gene, one of the oldest existing and functioning genes. Proc. Natl. Acad. Sci. USA **90**, 3009–3013.
32. Kluyver, A. J., van Niel, C. B. (1936). Prospects for a natural system of classification of bacteria. Zbl. Bakt. (2. Abt.) **94**, 369–393.
33. Labedan, B., Riley, M. (1994). Widespread sequence similarities among *Escherichia coli* proteins. J. Bacteriol. **177**, 1585–1588.
34. Lazcano, A. (1993). Biogenesis: some like it very hot. Science **260**, 1154–1155.
35. Lazcano, A. (1994). The RNA world, its predecessors and descendants. *In* Bengtson, S. (ed.), Early Life on Earth: Nobel Symposium No. 84, pp. 70–80. Columbia University Press, New York, NY.

36. Lazcano, A. (1994). The transition from non-living to living. *In* Bengtson, S. (ed.), Early Life on Earth: Nobel Symposium No. 84, pp. 60–69. Columbia University Press, New York, NY.
37. Lazcano, A., Miller, S. L. (1994). How long did it take for life to begin and evolve to cyanobacteria? J. Mol. Evol. **39**, 546–554.
38. Lazcano, A., Fox, G. E., Oró, J. (1992). Life before DNA: the origin and evolution of early Archean cells. *In* Mortlock, R. P. (ed.), The Evolution of Metabolic Function, pp. 237–295. CRC Press, Boca Raton, FL.
39. Lazcano, A., Guerrero, R., Margulis, L., Oró, J. (1988). The evolutionary transition from RNA to DNA in early cells. J. Mol. Evol. **27**, 283–290.
40. Marguet, E., Forterre, P. (1994). DNA stability at temperatures typical for hyperthermophiles. Nucleic Acid Res. **22**, 1681–1686.
41. Margulis, L. (1993). Symbiosis in Cell Evolution (2nd. ed.). W. H. Freeman and Co., New York, NY.
42. Miller, S. L., Bada, J. L. (1989). Submarine hot springs and the origin of life. Nature **334**, 609–611.
43. Miller, S. L., Lazcano, A. (1994). From the primitive soup to cyanobacteria: it may have taken less than 10 million years. *In* Doyle, L. (ed.), Circumstellar Habitable Zones. California University Press, Berkeley, CA, in press.
44. Nuttall, G. H. F. (1904). Blood Immunity and Blood Relationship: a Demostration of Certain Blood-Relationships amongst Animals by Means of the Precipitin Test for Blood. Cambridge University Press, Cambridge, United Kingdom.
45. Oparin, A. I. (1938).The Origin of Life. MacMillan Co., New York, NY.
46. Oró, J., Lazcano-Araujo, A. (1981). The role of HCN and its derivatives in prebiotic evolution. *In* Vennesland, B., Conn, E. E., Knowles, C. J., Westley, J., Wissing, F. (ed.), Cyanide in Biology, pp. 517–541. Academic Press, New York, NY.
47. Razin, S. (1978). The mycoplasmas. Microbiol. Rev. **42**, 414–470.
48. Reeve, J. N. (1994). Thermophiles in New Zeland. ASM News **60**, 541–545.
49. Reichard, P. (1993). From RNA to DNA, why so many ribonucleotide reductases? Science **260**, 1773–1777.
50. Riley, M., Anilionis, A. (1980). Evolution of the bacterial genome. Annu. Rev. Microbiol. **32**, 519–560.
51. Schimke, R. T., Sherwood, T. W., Hill, A. B. (1986). The rapid generation of genomic change as a result of over-replication. Chemica Scripta **26B**, 305–307.
52. Scher, S. (1959). Thermal factors in archaeometabolism. *In* Oparin, A. I., Pasynskii, A. G., Braunshtein, A. E., Pavlovskaya, T. E. (ed.), The Origin of Life on the Earth, pp. 650–651. Pergamon Press/ MacMillan Co., New York, NY.
53. Schofield, J. P. (1993). Molecular studies on an ancient gene encoding for carbamoyl-phosphate synthethase. Clin. Sci. **84**, 119–128.
54. Schopf, J. W. (1993). Microfossils of the early Archean apex chert: new evidence of the antiquity of life. Science **260**, 640–646.
55. Schwartz, R. M., Dayhoff, M. O. (1978). Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science **199**, 1395–1403.
56. Segerer, A. H., Burograf, S., Fiala, G., Huber, G., Huber, R., Pley, U., Stetter, K. O. (1993). Life in hot springs and hydrothermal vents. Orig. Life Evol. Bios. **23**, 77–90.
57. Sleep, N. H., Zahne, K. J., Kastings, J. F., Morowitz, H. J. (1989). Annihilation of ecosystems by large asteorid impacts on the early Earth. Nature **342**, 139–142.
58. Smith, M. W., Feng, D. F., Doolittle, R. F. (1992). Evolution by acquisition: the case for horizontal gene transfers. Trends Biochem. Sci. **17**, 489–493.
59. Sogin, M. L. (1991). Early evolution and the origin of eukaryotes. Curr. Opin. Gen. Develop. **1**, 457–463.
60. Stetter, K. O. (1994). The lesson of archaebacteria. *In* Bengtson, S. (ed.), Early Life on Earth: Nobel Symposium No. 84, pp. 143–151. Columbia University Press, New York, NY.
61. Stevens, P. F. (1980). Evolutionary polarity of character states. Annu. Rev. Ecol. System. **11**, 333–358.
62. Takao, M., Kobayashi, T., Oikawa, A., Yasui, A. (1989). Tandem arrangements of photolyase and superoxide dismutase genes in *Halobacterium halobium*. J. Bacteriol. **171**, 6323–6329.

63. Takao, M., Oikawa. A., Yasui, A. (1990). Characterization of a superoxide dismutase gene from the archaebacterium *Methanobacterium hermoautotrophicum*. Arch. Biochem. Biophys. **283**, 219–216.
64. Tiboni, O., Cammarano, P., Sanangelantoni, M. A. (1993). Cloning and sequencing of the gene encoding glutamine synthase I from the archaeum *Pyrococcus woesei*: anomalous phylogenies infered from analysis of archeal and bacterial glutamine synthase I sequences. J. Bacteriol. **175**, 2961–2969.
65. Tlsty, T. D., Albertini, A. M., Miller, J. H. (1984). Gene amplification in the *lac* region of *E. coli*. Cell **37**, 217–224.
66. Wächsterhäuser, G. (1990). The case for the chemoautotrophic origins of life in an iron-sulfur world. Orig. Life Evol. Bios. **20**, 173–182.
67. Wallace, D. C., Morowitz, N. H. (1973). Genome size and evolution. Chromosoma **40**, 121–126.
68. Woese, C. (1983). The primary lines of descent and the universal ancestor. *In* Bendall, D. S. (ed.), Evolution from Molecules to Man, pp. 209–233. Cambridge University Press, Cambridge, United Kingdom.
69. Woese, C. R. (1987). Bacterial evolution. Microbiol. Rev. **51**, 221–271.
70. Woese, C. R., Fox, G. E. (1977). The concept of cellular evolution. J. Mol. Evol. **10**, 1–6.
71. Ycas, M. (1974). On the earlier states of the biochemical system. J. Theor. Biol. **44**, 145–160.
72. Ziilig, W., Palm. P., Klenk, H. P. (1992). A model of the early evolution of organisms: the arisal of the three domains of life from the common ancestor. *In* Hartman, H., Matsuno, K. (ed.), The Origin and Evolution of the Cell, pp. 163–182. World Scientific Co., Singapore.
73. Zuckerkandl, E., Pauling, L. (1965). Molecules as documents of evolutionary history. J. Theoret. Biol. **8**, 357–366.

EVOLUTION OF THE BIOSYNTHESIS OF THE BRANCHED-CHAIN AMINO ACIDS

ANTHONY D. KEEFE[1], ANTONIO LAZCANO[2] and STANLEY L. MILLER[1]
[1] *Department of Chemistry, University of California San Diego, La Jolla, CA 92093-0317;*
[2] *Departamento de Biología, Facultad de Ciencias, UNAM, Apdo. Postal 70-407, Cd. Universitaria, México 04510, D.F., México*

(Received December, 1993)

Abstract. The origin of the biosynthetic pathways for the branched-chain amino acids cannot be understood in terms of the backwards development of the present acetolactate pathway because it contains unstable intermediates. We propose that the first biosynthesis of the branched-chain amino acids was by the reductive carboxylation of short branched chain fatty acids giving keto acids which were then transaminated. Similar reaction sequences mediated by nonspecific enzymes would produce serine and threonine from the abundant prebiotic compounds glycolic and lactic acids. The aromatic amino acids may also have first been synthesized in this way, e.g. tryptophan from indole acetic acid. The next step would have been the biosynthesis of leucine from α-ketoisovaleric acid. The acetolactate pathway developed subsequently. The first version of the Krebs cycle, which was used for amino acid biosynthesis, would have been assembled by making use of the reductive carboxylation and leucine biosynthesis enzymes, and completed with the development of a single new enzyme, succinate dehydrogenase. This evolutionary scheme suggests that there may be limitations to inferring the origins of metabolism by a simple back extrapolation of current pathways.

1. Introduction

In the first discussion of the origin of biosynthetic pathways, Horowitz (1945) proposed that biosynthetic pathways arose by backwards development rather than forwards. The basis of the proposed process was the utilization of a prebiotic soup that contained all of the biosynthetic intermediates which constitute the resultant pathway. When a required compound (eg threonine) became exhausted from the environment, the preceding intermediate (homoserine) would have been converted to threonine. The next step arose when homoserine became exhausted from the environment, and an enzyme appeared that could convert aspartic semialdehyde to homoserine. In this manner the pathways evolved in a stepwise fashion.

The discovery of operons led Horowitz (1965) to extend his hypothesis to additionally take account of gene duplications as a source of new enzymes. Hegeman and Rosenberg (1970) suggested that regulation and the recruitment of activities from existing pathways may have been more important than gene duplication. Yčas (1974) and Jensen (1976) emphasized the importance of the broad specificity of early enzymes and the development of more specific enzymes by gene duplication followed by sequence divergence.

Surprisingly there seem to have been no attempts to examine the Horowitz hypothesis in terms of the prebiotic soup. The biosyntheses of threonine and methionine are rationally explained by this backwards stepwise development. The prebiotic synthesis of threonine is by a straightforward electric discharge reaction

Origins of Life and Evolution of the Biosphere **25**: 99–110, 1995.

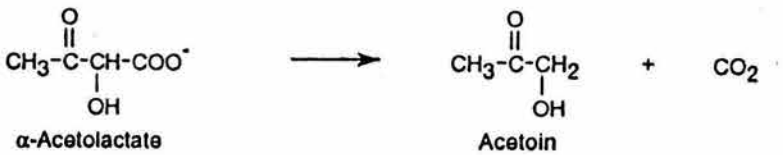
(Ring *et al.*, 1972). The prebiotic synthesis of methionine is from acrolein which also gives homoserine and homocysteine as well as glutamic acid and diamino butyric acid (Van Trump and Miller, 1972). Consistent with this picture is the homology of threonine synthase, threonine dehydratase and serine dehydratase (Parsot, 1986), and the homology of β-cystathionase and cystathionine δ-synthase of the methionine pathway (Belfaiza *et al.*, 1986).

The Horowitz hypothesis does not seem applicable to some biosynthetic pathways, e.g. the biosynthesis of purines from glycine. This pathway clearly cannot be based on components of the prebiotic soup since all the intermediates are ribosides which are unstable to hydrolysis. In addition, ribose is generally considered not to have been a significant component of the prebiotic soup (Shapiro, 1988). If the biosynthesis of purines was by the same pathway as at the present but without the ribose, the sequence would still contain mostly unstable compounds, for example $HCONHCH_2CONH_2$.

Branched-chain Amino Acid Biosynthesis

The contemporary biosyntheses of the branched-chain amino acids valine, isoleucine, and leucine by *Escherichia coli*, are shown in Figure 1. Four steps are common to all three synthetic pathways (Umbarger, 1987) and so the sequence of reactions are considered together. The pathways are the same in eubacteria, archaebacteria, and eukaryotes, and the enzymes are homologous (Xing and Whitman, 1991).

The Horowitz hypothesis cannot apply to the branched-chain amino acids valine, isoleucine and leucine, because their biosynthetic precursors are unstable, e.g. α-acetolactate decarboxylates readily because it is a β-keto acid, and so would not have been in the prebiotic soup. The half-life for decarboxylation is several days at room temperature (Hill *et al.*, 1979), and the anion decarboxylates readily on acidification:



In addition, acetolactate mutase catalyses an alkyl migration by an acyloin rearrangement which is almost unique in biochemistry, and so it may not be a very primitive enzyme.

A variation on the Horowitz proposal assumes that if a compound was present in the prebiotic environment, then so were its decomposition products. Consequentially, a biosynthetic pathway to this compound could have arisen in a stepwise fashion, utilizing the sequence of compounds available in the decomposition path-

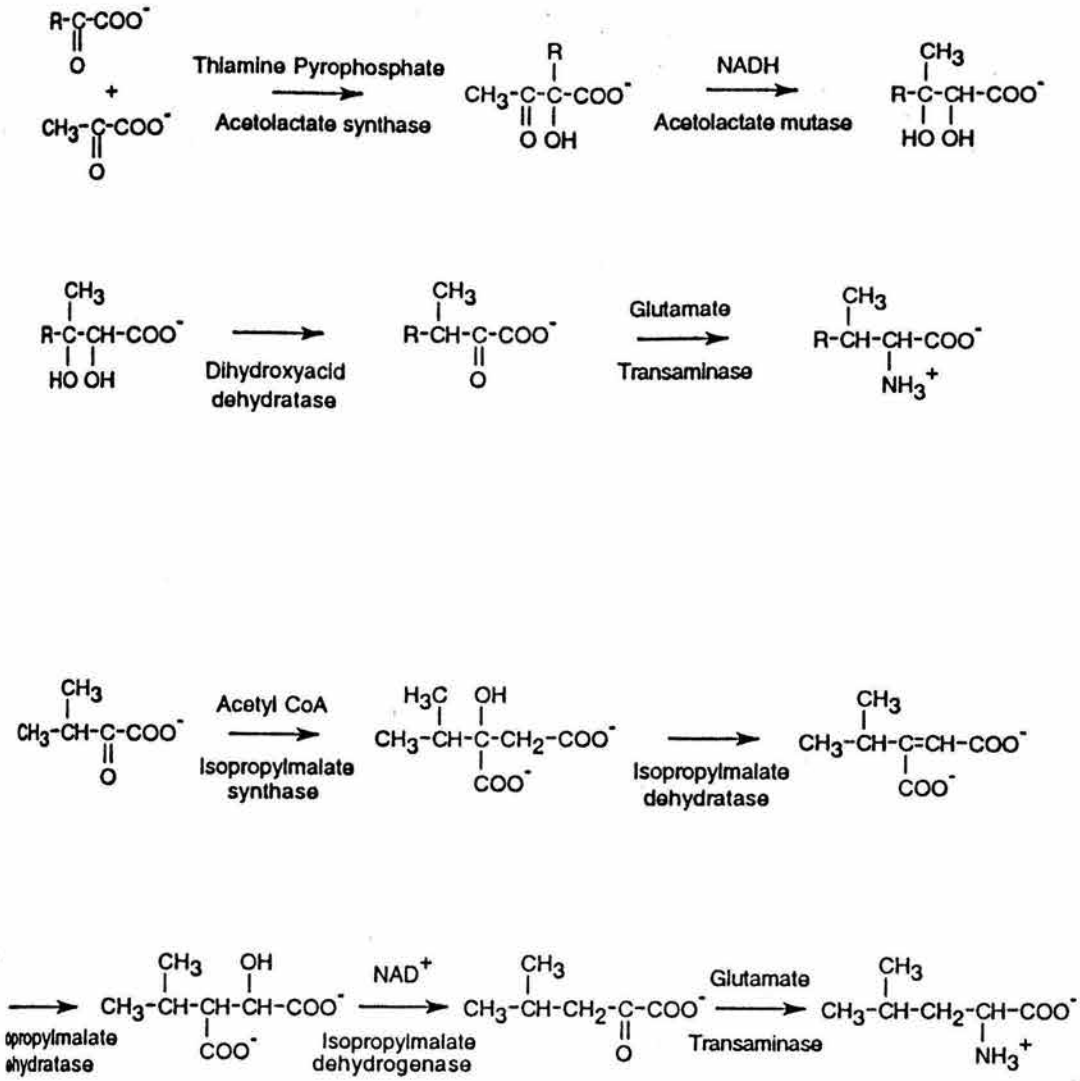
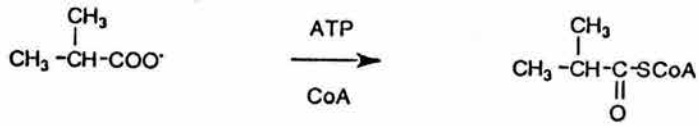
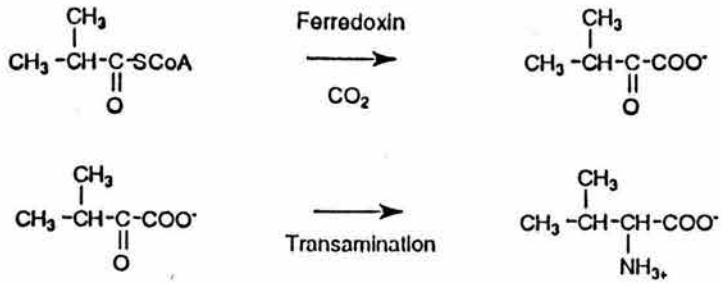


Fig. 1. Contemporary biosyntheses of the branched-chain amino acids. Valine, R = CH_3; Isoleucine R = CH_2CH_3.

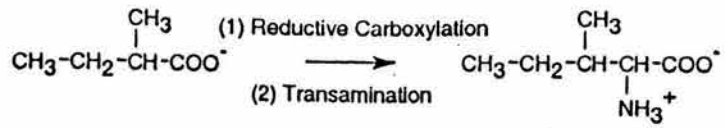
way, rather than the compound's precursors in the contemporary sequence. A modified reverse of the mammalian degradative pathway is, in the case of valine:

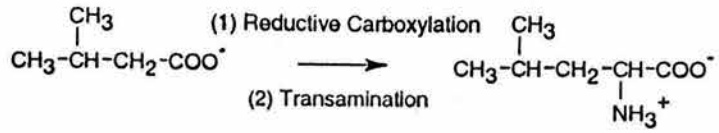




A similar degradative pathway is used in some anaerobic bacteria (Allison, 1978). It should be kept in mind that the acyl CoA could not have been in the prebiotic soup because it is unstable. The utilization of ATP and CoA in this pathway assumes that analogous enzymes were present, and that the enzymes using isobutyric acid arose by gene duplication.

The reactions giving isoleucine and leucine are shown below:





The reductive carboxylation is carried out by low potential ferredoxins, since NADH is not sufficiently reducing for this. Another possible reducing agent is pyrite (i.e., $FeS + H_2S \longrightarrow FeS_2 + H_2$) as proposed by Wächtershäuser (1988). This might work prebiotically or possibly as an early bacterial process.

The pathway discussed above is used by *Methanobacterium ruminantium, Bacteroides rumminicole* and other primitive prokaryotes in ruminants* for valine and isoleucine biosyntheses from isobutyric acid and α-methyl butyric acid, respectively (Robinson and Allison, 1969; Allison and Peel, 1971). The short chain aliphatic acids are likely to have been more abundant than the corresponding amino acids on the primitive Earth, as they are in the Murchison meteorite (Table I). A single enzyme could have produced valine, isoleucine and leucine from isobutyric, α-methylbutyric and isovaleric acids, respectively. When the fatty acid precursors were exhausted it would then have become necessary to develop the acetolactate pathway.

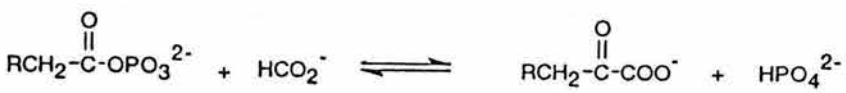
* The cow is modern, but the bacteria are ancient.

TABLE I

Carboxylic acids and corresponding amino acids occurring in the Murchison meteorite [data from Lawless and Yuen (1979) and Cronin and Pizzarello (1983)]

Carboxylic Acid	Abundance $nmol\,g^{-1}$	Corresponding amino acid	Abundance $nmol\,g^{-1}$
Ethanoic	1030	Alanine	44
Propanoic	1830	2-Aminobutyric	18
2-Methylpropanoic	500	Valine	10
Butanoic	380	Norvaline	3
2-Methylbutanoic	120	Isoleucine	4
3-Methylbutanoic	90	Leucine	4
Pentanoic	120	Norleucine	2
4-Methylpentanoic	70		
Hexanoic	60		
Heptanoic	30		
Octanoic	10		

The activation of the short chain fatty acids is a straightforward CoA synthesis which would have been mediated by an enzyme that could easily have been acquired by a gene duplication. The third step would involve a transamination enzyme that could also have arisen by gene duplication. The reductive carboxylation step is the only one requiring a new enzyme not easily obtained from a gene duplication, unless the Krebs cycle biosynthetic pathway is more ancient (see below). An alternative to the reductive carboxylation is to react acyl phosphate with formate (Tanaka and Johnson, 1971). This still requires the activation of the fatty acids, but the reaction may be simpler than the reductive carboxylation:

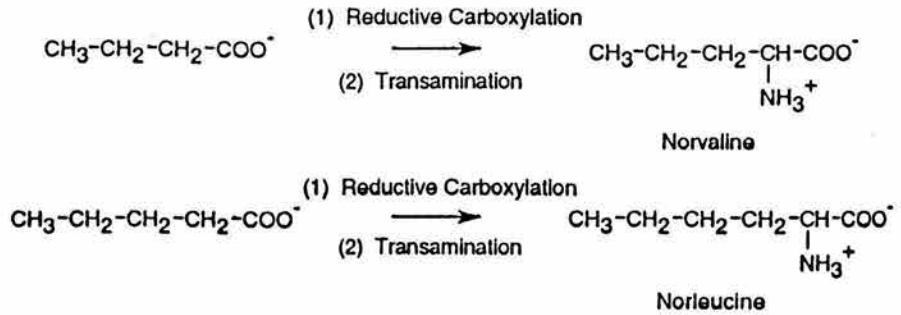


This sequence could be prebiotic. Transaminations are catalyzed non-enzymatically by pyridoxal (Metzler *et al.*, 1954), histidine (Doctor and Oró, 1969) and glyoxalate (Warren, 1971). The activation of the fatty acid would have been the result of the prebiotic activation reactions. A prebiotic version of the carboxylation could be as is shown below (Eggerer *et al.*, 1962):

Acyl cyanides usually hydrolyze to the carboxylic acid and HCN except in strong acid, but soft nucleophiles such as H_2S can react under some conditions to give $CH_3COCSNH_2$ (Hünig and Schaller, 1982), which would hydrolyze to the keto acid. The prebiotic reaction conditions remain to be worked out.

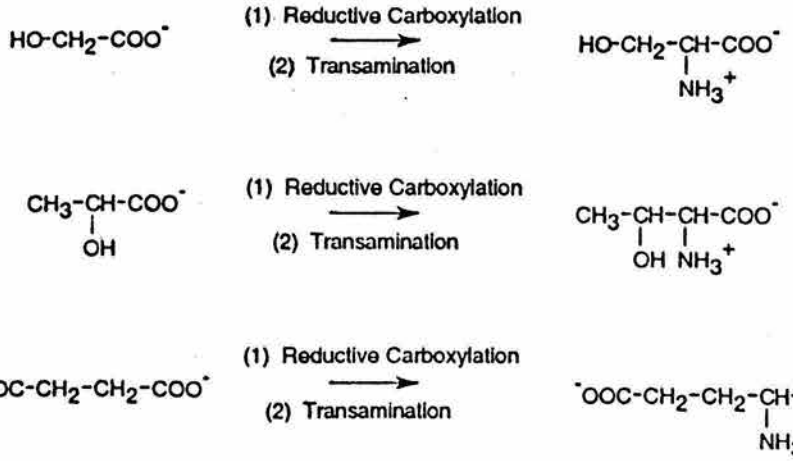
Fatty Acid Metabolism and the Absence of Norvaline and Norleucine from Proteins

If the fatty acid synthesis and degradative pathways developed early, then n-valeric, n-butyric and propionic acids would have been depleted early from the environment. This would have prevented the synthesis of norleucine, norvaline and α-amino-n-butyric acid by reductive carboxylation. This may explain the absence of these straight chain amino acids from proteins, which is otherwise difficult to account for (Weber and Miller, 1981):



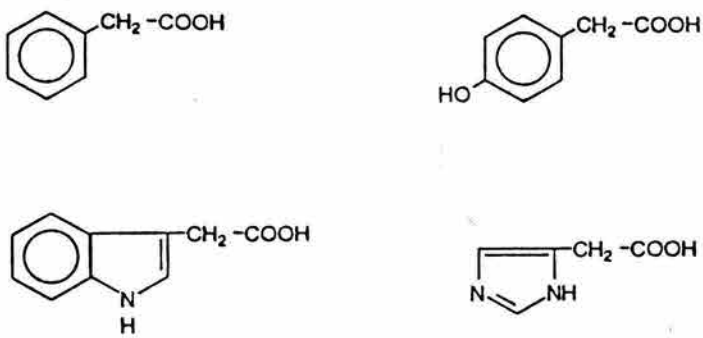
Early Biosyntheses of Serine, Threonine and Glutamic Acid

Similar considerations may be applied to other amino acids with abundant prebiotic precursors. Thus serine could have been made from glycolic acid and threonine from lactic acid. These hydroxy acids are major products of prebiotic syntheses and also occur in the Murchison meteorite (Miller, 1957; Peltzer and Bada, 1978; Peltzer *et al.*, 1984). This scheme would have greatly increased the availability of serine and threonine, as glycolic and lactic acids are more stable than serine and threonine. A slight evolution of the reductive carboxylating enzyme would have allowed the synthesis of glutamic acid from succinic acid and alanine from acetic acid. This would have constituted the beginning of the reverse Krebs cycle used for amino acid synthesis in some anaerobic organisms (see below):



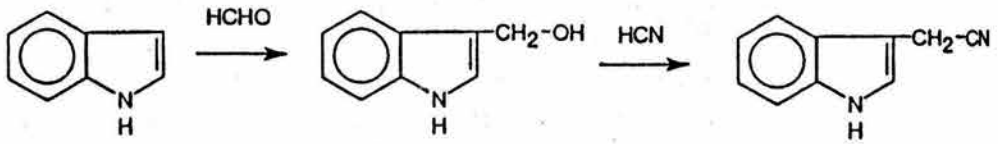
The first biosynthetic Pathway for the Aromatic Amino Acids

Prebiotic syntheses have been demonstrated for phenylalanine and tyrosine (Friedmann and Miller, 1969), trytophan (Friedmann *et al.*, 1971) and histidine (Shen *et al.*, 1987, 1990). These are not particularly efficient syntheses, and the supply of these aromatic amino acids would have been quickly exhausted if they were components of early organisms. The reductive carboxylation/transamination scheme suggested here would have been a prebiotic source of these amino acids. Some primitive archaebacteria synthesize phenylalanine and tyrosine by this pathway (Sauer *et al.*, 1975). By a similar process tryptophan and histidine could be produced from indole acetic acid and imidazole acetic acid. The precursor acids are:



The efficient synthesis of these aromatic acetic acids may be prebiotic. In the

case of indole acetic acid the synthesis is shown below:



This is then followed by hydrolysis. A similar reaction should readily occur with phenol, but imidazole and especially benzene are relatively unreactive. Once these precursors were exhausted from the environment, the development of the complex shikimic acid aromatic biosynthetic pathway would have become necessary.

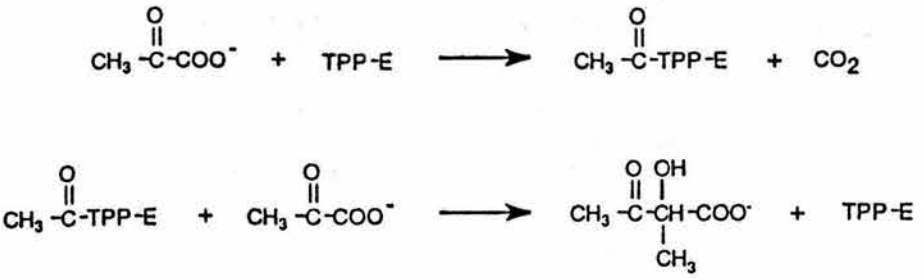
Origin of the Acetolactate Pathway

Although the branched chain fatty acids were more abundant than amino acids in the prebiotic soup, the short branched chain fatty acids quickly would have become exhausted. A reasonable order of development would be to assume that the leucine pathway from α-keto isovaleric acid developed first. The reaction of acetyl CoA with α-keto isovaleric acid is an aldol condensation for which the development of an anzyme, isopropyl malate synthase, may be easily envisioned.

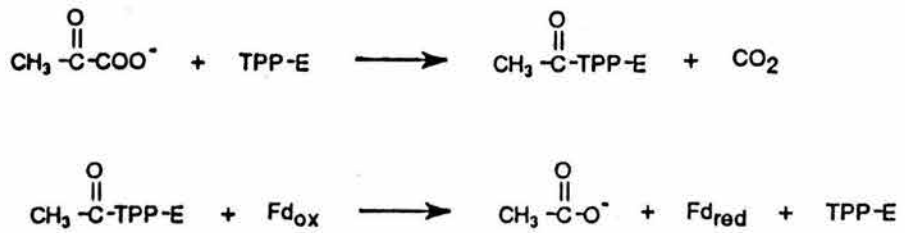
Isopropyl malate isomerase (or dehydratase) catalyses a very similar reaction to that catalyzed by fumarase which proceeds non-enzymatically in acidic (Rozelle and Alberty, 1957), basic (Erikson and Alberty, 1959) and neutral (Bada and Miller, 1969) solutions at elevated temperature, and so development of the enzyme is easily envisaged. If fumarase or crotonase were present in the prokaryotic metabolic apparatus, the development of isopropyl malate dehydrogenase would have rapidly occurred by a gene duplication and subsequent sequence divergence. Isopropyl malate dehydrogenase is a standard NAD^+ alcohol dehydrogenase with a rapid non-enzymatic decarboxylation step.

Acetolactate synthase uses thiamine pyrophosphate. This reaction also occurs non-enzymatically with thiamine. The product is usually acetoin because the acetolactate rapidly decarboxylates (Breslow and McNelis, 1959).

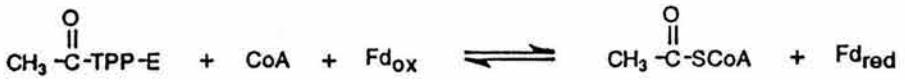
Chang and Cronan (1988) demonstrated the functional and structural homology between pyruvate oxidase and acetohydroxy acid synthase in the branched chain amino acid biosynthetic pathway, and suggested that the synthase was derived from the pyruvate oxidase. The evolutionary conservation of the homology is easy to understand from the standpoint of the chemical reactions involved. Acetolactate synthase catalyzes the reactions



where TPP refers to thiamine pyrophosphate, E is the enzyme, and CH_3-CO-TPP-E is active acetaldehyde [CH_3-CO($^-$)] attached to the thiamine. Pyruvate oxidase catalyzes the reactions



where Fd_{ox} and Fd_{red} are the oxidized and reduced forms of ferredoxin, but other electron acceptors such as NAD^+ and flavins are used with some enzymes. Other pyruvate oxidoreductases carry out the reaction with coenzyme A:



In this case the reaction may be reversible depending on the potential of the electron acceptor.

The homology of acetolactate synthase with pyruvate oxidase presumably extends mostly to the domains involved in the decarboxylation step, and not to other sections of the enzyme. We agree with the proposal of Chang and Cronan (1988) that acetolactate synthase was derived from pyruvate oxidase, but our modification to their scheme is that the pyruvate oxidase was a reversible enzyme operating in the keto acid synthesis direction rather than the irreversible direction. The acetolactate pathway would be completed with the development of the reductoisomerase and the dihydroxy acid dehydratase.

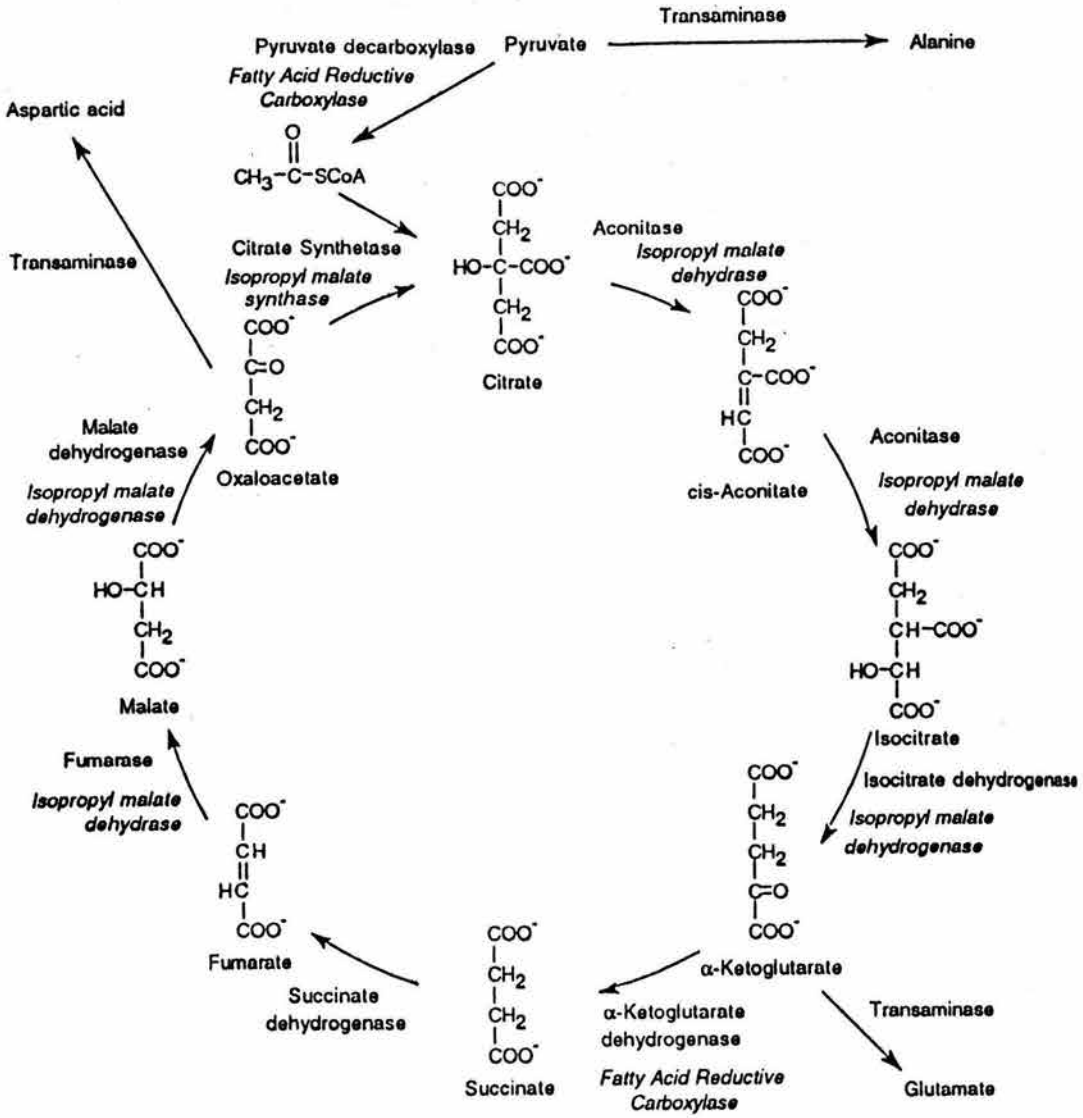


Fig. 2. The Krebs cycle. Also included in italics are the enzymes of the biosynthesis of leucine from α-ketoisovaleric acid and of the branched chain fatty acids to the keto acid (fatty acid reductive carboxylase).

Origin of the Krebs Cycle

The central role of the Krebs cycle in metabolism suggests that it is a very ancient pathway. Its origin is generally accepted as being for the biosynthesis of amino acids rather than ATP production, since no ATP is produced in the absence of an electron acceptor for the NADH produced in the cycle.

It is tempting to suggest that the Krebs cycle came before branched-chain amino acid biosynthesis because most of the twenty protein amino acids are derived from Krebs cycle intermediates. However, the Krebs cycle amino acids are among the most abundant prebiotic amino acids, i.e. alanine, aspartic acid and glutamic acid, and so they would have been depleted later than the less abundant branched-chain amino acids.

In contrast we propose that the Krebs cycle was developed by modification of the leucine biosynthetic pathway from valine. It is assumed that isopropyl malate dehydrogenase, isopropyl malate dehydratase and isopropyl malate synthetase were available, as well as the fatty acid reductive carboxylases and transaminase from the early branched-chain amino acid biosynthetic scheme. Thus only one new enzyme, succinate dehydrogenase, was needed to complete the Krebs cycle. The scheme is shown in Figure 2. The oxidative version of the Krebs cycle would have been established when sufficiently high potential electron acceptors (e.g. O_2) became available.

The counter argument can be made that the biosynthetic Krebs cycle came first and that the branched-chain amino acid pathways were developed from the Krebs cycle enzymes. This would be justified if the branched-chain amino acids were incorporated late into proteins, and the depletion of alanine, aspartic acid, glutamic acid, and related amino acids occured prior to the exhaustion of the branched-chain amino acids from the primitive ocean.

There have been a number of discussions suggesting that the origin of metabolism can be inferred from the backwards extrapolation of the contemporary pathways. Our proposals suggest the present pathways may have replaced even older biosyntheses. Some of the oldest pathways and their enzymes have survived in unusual organisms, but some may have been lost from biology in the same way that the precursor to RNA has disappeared.

Acknowledgments

This paper was presented at the Barcelona ISSOL '93 Meeting, July 4-9, 1993. We thank the NASA Specialized Center of Research and Training (NSCORT) at the University of California San Diego for a post-doctoral fellowship (AK), a visiting professor followship (AL), and grant support (SLM).

References

Allison, M. J.: 1978, *Appl. Environ. Microbiol.* **35**, 872–877.
Allison, M. J. and Peel, J. L.: 1971, *Biochem. J.* **121**, 431–437.
Bada, J. L. and Miller, S. L.: 1969, *J. Am. Chem. Soc.* **91**, 3948–3949.
Belfaiza, J., Parsot, C., Martel, A., Bouthier de la Tour, C., Margarita, D., Cohen, G. N. and Saint-Girons, I.: 1986, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 867–871.
Breslow, R. and McNelis, E.: 1959, *J. Am. Chem. Soc.* **81**, 3080–3082.
Chang, Y. and Cronan, J. E.: 1988, *J. Bacteriol.* **170**, 3937–3945.
Cronin, J. R. and Pizzarello, S.: 1983, *Adv. Space Res.* **3** (9), 5-18.

Doctor, V. M. and Oró, J.: 1969, *Biochem. J.* **112**, 691–697.
Eggerer, H., Stadtman, E. R., and Poston, J. M.: 1962, *Arch. Biochem. Biophys.* **98**, 432–443.
Erickson, L. E. and Alberty, R. A.: 1959, *J. Phys. Chem.* **63**, 705–709.
Friedmann, N. and Miller, S. L.: 1969, *Science* **166**, 766–767.
Friedmann, N., Haverland, W. J., and Miller, S. L.: 1971, in Buvet, R. and Ponnamperuma, C. (Eds.), *Chemical Evolution and the Origin of Life*, North-Holland Publ. Co. Amsterdam, pp. 123–135.
Hegeman, G. D. and Rosenberg, S. L.: 1970, *Ann. Rev. Microbiol.* **24**, 429–462.
Hill, R. K., Sawada, S. and Arfin, S. M.: 1979, *Bioorg. Chem.* **8**, 175–189.
Horowitz, N. H.: 1945, *Proc. Natl. Acad. Sci. U.S.A.* **31**, 153–157.
Horowitz, N. H.: 1965, in Bryson, V. and Vogel, H. J. (Eds.), *Evolving Genes and Proteins*, Academic Press, New York, pp. 15–23.
Hünig, S. and Schaller, R.: 1982, *Angew. Chem. Int. Ed. Engl.* **21**, 36–49.
Jensen, R. A.: 1976, *Ann. Rev. Microbiol.* **30**, 409–425.
Lawless, J. G. and Yuen, G. U.: 1979, *Nature* **282**, 396–398.
Metzler, D. E. Ikawa, M., and Snell, E. E.: 1954, *J. Am. Chem. Soc.* **76**, 648–652.
Miller, S. L.: 1957, *Biochim. Biophys. Acta* **23**, 480–489.
Parsot, C.: 1986, *EMBO Journal* **5**, 3013–3019.
Peltzer, E. T. and Bada, J. L.: 1978.: *Nature* **272**, 443–444.
Peltzer, E. T. Bada, J. L., Schlesinger, G., and Miller, S. L.: 1984, *Adv. Space Res.* **4** (12), 69–74.
Ring, D., Wolman, Y., Friedmann, N., and Miller, S. L.: 1972, *Proc. Natl. Acad. Sci. U.S.A.* **69**, 765–768.
Robinson, I. M. and Allison, M. J.: 1969, *J. Bacteriol.* **97**, 1220–1226.
Rozelle, L. T. and Alberty, R. A.: 1957, *J. Phys, Chem.* **61**, 1637–1640.
Sauer, F. D., Erfle, J. D., and Mahadevan, S.: 1975, *Biochem. J.* **150**, 357–372.
Shapiro, R.: 1988, *Origins of Life* **18**, 71–85.
Shen, C., Yang, L. Miller, S. L., and Oró, J.: 1990, *J. Mol. Evol.* **31**, 167–174.
Shen, C., Yang, L., Miller, S. L., and Oró, J.: 1987, *Origins of Life* **17**, 295–305.
Tanaka, N. and Johnson, M. J.: 1971, *J. Bacteriol.* **108**, 1107–1111.
Umbarger, H. E.: 1987, in Neidhardt, F. C., Ingraham, J. L., Low, K. B., Magasanik, B., Schaechter, M. and Umbarger, H. E. (Eds.), *Escherichia coli and Salmonella typhimurium; Cellular and Molecular Biology*, American Society for Microbiology, Washington, D. C. **1**, 352–367.
Van Trump, J. E. and Miller, S. L.: 1972, *Science* **178**, 859–860.
Wächtershäuser, C.: 1988, *System. Appl. Microbiol.* **10**, 207–210.
Warren, W. A.: 1971, *Arch. Biochem. Biophys.* **143**, 212–217.
Weber, A. L. and Miller, S. L.: 1981, *J. Mol. Evol.* **17**, 273–284.
Xing, R. and Whitman, W. B.: 1991, *J. Bacteriol.* **173**, 2086–2092.
Yčas, M.: 1974, *J. Theor. Biol.* **44**, 145–160.

J Mol Evol (1995) 41:760–774

JOURNAL OF MOLECULAR EVOLUTION

Molecular Evolution of the Histidine Biosynthetic Pathway

Renato Fani,[1] Pietro Liò,[1] Antonio Lazcano[2]

[1] Dipartimento di Biologia Animale e Genetica, Università degli Studi di Firenze, Via Romana 17, I-50125 Firenze, Italy
[2] Facultad de Ciencias, Universidad Nacional Autónoma de México, Apartado Postal 70-407, Cd Universitaria, México 04510, D.F., México

Received: 3 January 1995 / Accepted: 27 July 1995

Abstract. The available sequences of genes encoding the enzymes associated with histidine biosynthesis suggest that this is an ancient metabolic pathway that was assembled prior to the diversification of the Bacteria, Archaea, and Eucarya. Paralogous duplications, gene elongation, and fusion events involving different *his* genes have played a major role in shaping this biosynthetic route. Evidence that the *hisA* and the *hisF* genes and their homologues are the result of two successive duplication events that apparently took place before the separation of the three cellular lineages is extended. These two successive gene duplication events as well as the homology between the *hisH* genes and the sequences encoding the TrpG-type amidotransferases support the idea that during the early stages of metabolic evolution at least parts of the histidine biosynthetic pathway were mediated by enzymes of broader substrate specificities. Maximum likelihood trees calculated for the available sequences of genes encoding these enzymes have been obtained. Their topologies support the possibility of an evolutionary proximity of archaebacteria with low GC Gram-positive bacteria. This observation is consistent with those detected by other workers using the sequences of heat-shock proteins (HSP70), glutamine synthetases, glutamate dehydrogenases, and carbamoylphosphate synthetases.

Key words: Histidine biosynthesis — Evolution of metabolic pathways — Molecular evolution

Introduction

The emergence of basic biosynthetic pathways was one of the major events during the early evolution of life, since their appearance allowed primitive organisms to become increasingly less dependant on exogenous sources of amino acids, purines, and other compounds that may have accumulated in the primitive environment as a result of prebiotic syntheses. How the major biosynthetic pathways actually originated is still an open question, but several different theories have been suggested to account for the establishment of anabolic routes. These explanations include (1) the retrograde hypothesis (Horowitz 1945, 1965); (2) the possibility that at least some biosynthetic routes evolved forward, i.e., from simple precursors to complex end products (Granick 1965); (3) the idea that metabolic pathways appeared as a result of the gradual accumulation of mutant enzymes with minimal structural changes (Waley 1969); and (4) the patchwork theory, according to which metabolic routes are the result of the serial recruitment of relatively small, inefficient enzymes endowed with broad specificity that

Abbreviations: aa = amino acid; ORF = open reading frame; bp = base pair; kb = 10^3 bp; CarA = carbamoyl phosphate synthetase (EC 6.3.5.5); GAT = glutamine amidotransferase; GuaA = GMP synthetase (EC 6.3.4.1); PabA = 4-amino-4-deoxychorismate synthase (EC 4.1.3-); PyrG = GTP synthetase (EC 6.3.4.2); AICAR = 5-aminoimidazole-4-carboxamide-1-β-D ribofuranosyl 5′-monophosphate; HAL = L-histidinal; HOL = L-histidinol; HP = histidinol phosphate; IAP = imidazole acetol-phosphate; IGP = imidazole glycerol phosphate; PR = phosphoribosyl; PRFAR = *N*-[(5′-phosphoribulosyl) formimino]-5-aminoimidazole-4-carboxamide ribonucleotide; 5′-ProFAR = N^1-[(5′-phosphoribosyl) formimino]-5-aminoimidazole-4-carboxamide ribonucleotide; PRPP = phosphoribosyl-pyrophosphate; RFLP = restriction fragment length polymorphism

Correspondence to: R. Fani

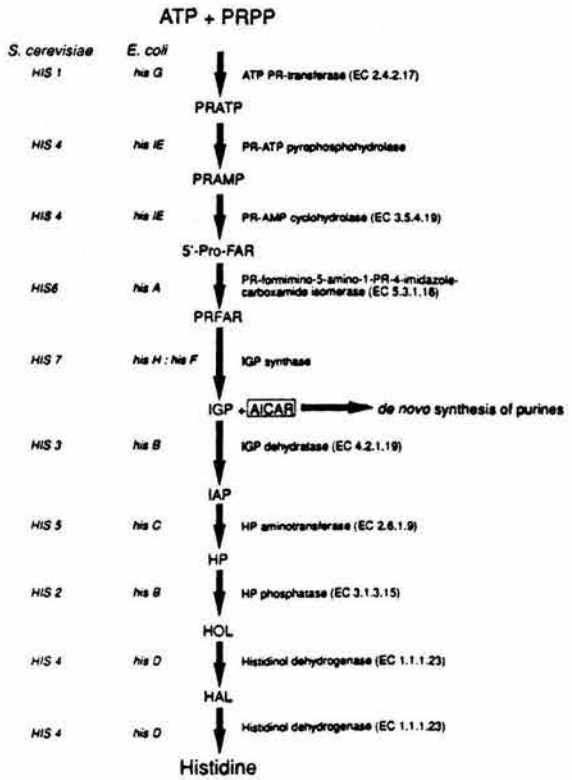


Fig. 1. Gene–enzyme relationships in the histidine biosynthetic pathway both in *E. coli* and *S. cerevisiae*. The schematic representation of the biosynthetic steps from ATP and PRPP to histidine follows the enzyme designations and nomenclature.

could react with a wide range of chemically related substrates (Ycas 1974; Jensen 1976).

Histidine biosynthesis is one of the best-characterized anabolic pathways. There is a large body of genetic and biochemical information, including operon structure, gene expression, and increasingly larger sequence databases which are available. For over 30 years this pathway has been the subject of extensive studies, mainly in the enterobacterium *Escherichia coli* and its close relative *Salmonella typhimurium*, in both of which details of histidine biosynthesis appear to be identical (Winkler 1987). The complete nucleotide sequence of their *his* operons has been determined by Carlomagno et al. (1988). As shown in Fig. 1, in these two enterobacteria the pathway is unbranched and includes a number of complex and unusual biochemical reactions. It consists of nine intermediates, all of which have been described, and of eight distinct enzymes. Three of these enzymes, encoded by the *hisB*, *hisD*, and *hisIE* genes, are bifunctional (Winkler 1987; Carlomagno et al. 1988).

There are several independent indications of the antiquity of the histidine biosynthesis pathway. It is generally accepted that histidine is present in the active sites of enzymes because of the special properties of the imidazole group (Weber and Miller 1981). The apparently universal phylogenetic distribution of the *his* genes (Table 1) suggests that histidine synthesis was already part

Table 1. Organisms where histidine genes have been sequenced[a]

Organism	Genes
Bacteria	
Purple bacteria	
α Subdivision	
Azospirillum brasilense	*Bd H orf168 A F E orf122*
γ Subdivision	
Escherichia coli	*G D C B H A F (IE)*
Salmonella typhimurium	*G D C B H A F (IE)*
Klebsiella pneumoniae	*G*[b]
Gram positive	
Low GC content	
Lactococcus lactis	*C orf3 G D orf6 Bd orf8 H A F (IE) orf13*
Bacillus subtilis	*C*
High GC content	
Streptomyces coelicolor	*D C Bd orf1 H A orf2*[b]
Mycobacterium smegmatis	*D C*[b]
Archaea	
Methanococcus vannielii	*A, I*
Methanococcus voltae	*A*
Methanococcus thermolithotrophicus	*A*
Halobacterium volcanii	*C*
Eucarya	
Saccharomyces cerevisiae	*Bd, C, G, (HF), (IED)*
Saccharomyces kluyveri	*Bd*
Candida maltosa	*C*
Pichia pastoris	*(IED)*
Thricoderma harzianum	*Bd*
Neurospora crassa	*(IED)*
Phytophthora parasitica	*Bd*
Brassica oleracea	*D*

[a] References are given in Materials and Methods. Genes in parentheses are fused; genes coding for bi- or multifunctional enzymes are underlined. Eucaryotic *his* genes are indicated by the name of their prokaryotic counterparts. The name *hisBd* (Iimauro et al. 1990) refers to genes coding an IGP dehydratase corresponding to the distal moiety of the enterobacterial HisB enzyme

[b] Truncated genes or ORF

of the metabolic abilities of the last common ancestor of the three extant cell domains (Lazcano et al. 1992). The chemical syntheses of histidine (Shen et al. 1990b) and prebiotic analogues of histidine (Maurel and Ninio 1987) and of histidyl-histidine under primitive conditions have been reported (Shen et al. 1990a), as well as the role of the latter in the enhancement of some possible prebiotic oligomerization reactions involving amino acids (White and Erickson 1980) and nucleotides (Shen et al. 1990c). Since its biosynthesis requires a carbon and a nitrogen equivalent from the purine ring of ATP, it has also been suggested that histidine may be the molecular vestige of a catalytic ribonucleotide from an earlier biochemical stage in which RNA played a major role in catalysis (White 1976).

Histidine biosynthesis plays an important role in cellular metabolism, since it is interconnected to both the de novo synthesis of purines and to nitrogen metabolism (Fig. 1). The connection with purine biosynthesis results from an enzymatic step catalyzed by imidazole glycerol

phosphate synthase, an enzyme which has recently been shown to be a dimeric protein composed of one subunit each of the *hisH* and *hisF* gene products (Klemm and Davisson 1993). This heterodimeric enzyme catalyzes the transformation of PRFAR into AICAR, which is then recycled into the de novo purine biosynthetic pathway, and imidazole glycerol phosphate (IGP), which in turn is then transformed into histidine (Fig. 1). Histidine biosynthesis is connected to nitrogen metabolism by a glutamine molecule, which is believed to be the source of the final nitrogen atom of the imidazole ring of IGP. The important role played by histidine biosynthesis in cellular metabolism is in fact underscored by the considerable energy (41 ATP molecules) required for the synthesis of each histidine molecule (Brenner and Ames 1971).

The histidine pathway has been also investigated in a number of organisms (Table 1), including the archaea [*Methanococcus vannielii* (Beckler and Reeve 1986), *M. voltae* (Cue et al. 1985) and *M. thermolithotrophicus* (Weil et al. 1987), *Halobacterium volcanii* (Conover and Doolittle 1990)]; the bacteria [*Klebsiella pneumoniae* (Rodriguez et al. 1981; Rodriguez and West 1984), *Bacillus subtilis* (Henner et al. 1986), *Streptomyces coelicolor* (Limauro et al. 1990), *Lactococcus lactis* (Delorme et al. 1992, 1993), *Mycobacterium smegmatis* (Hinshelwood and Stocker 1992) and *Azospirillum brasilense* (Bazzicalupo et al. 1987; Fani et al. 1989, 1993)]; and the eucarya, such as the fungi *Saccharomyces cerevisiae* (Sthrul 1985; Kuenzler et al. 1993), *Neurospora crassa* (Legerton and Yanofsky 1985), *Candida maltosa* (Hikiji et al. 1989), *Candida albicans* (Altboum et al. 1990), *Saccharomyces kluyveri* (Weinstock and Strathern 1993), and *Thricoderma harzianum* (Goldman et al. 1992), as well as the plant *Brassica oleracea* (Nagai et al. 1991).

As discussed below, the study of histidine biosynthesis in these organisms has shown that there are important differences in the way in which their *his* genes are organized. Although in some eubacteria these genes are clustered in a single operon (Carlomagno et al. 1988; Delorme et al. 1992), in other prokaryotes they are scattered throughout the chromosome (Beckler and Reeve 1986; Hopwood et al. 1985; Limauro et al. 1990). In some species more than one enzymatic function is encoded by the same cistron: *hisD, hisB,* and *hisIE* in *E. coli* and *S. typhimurium* (Carlomagno et al. 1988), *hisD* and *hisIE* in *L. lactis* (Delorme et al. 1992), *HIS4* and *HIS7* in *S. cerevisiae* (Donahue et al. 1982; Kuenzler et al. 1993), and *his-3* in *N. crassa* (Legerton and Yanofsky 1985). In eukaryotes the *his* genes appear to be always distributed in different chromosomes (Broach 1981).

It has also been demonstrated that two of the prokaryotic histidine genes, *hisA* and *hisF*, are paralogous in that they have originated from the duplication of an ancestral gene, which in turn resulted from a gene elongation event involving an ancestral module half the size of the extant *hisA* gene (Fani et al. 1994). Furthermore, the homology between the imidazole glycerol-P synthase encoded by the *E. coli hisH* gene and other G-type glutamine amidotransferases (GATs) involved in the formation of GMP, CTP, tryptophan, carbamoyl-phosphate, and other molecules (Zalkin 1985) is well established. These discoveries suggest that several ancient paralogous duplications played a major role in shaping the extant structure of the histidine biosynthetic pathway.

Materials and Methods

Amino acid and nucleotide sequences were retrieved from the GenBank, EMBL, and PIR databases. The *Clustal V* program was used for sequence alignments (Higgins and Sharp 1988). Sequence similarity values among protein sequences were calculated using the S_{AB} coefficient as defined by Fox et al. (1977). Phylogenetic analysis of the available sequences has been performed using the algorithms described by Weir (1990), Li and Graur (1991), and some additional ones prepared by the authors. These are available upon request.

We have used *DNAML* by Felsenstein (1981) to compute maximum likelihood trees for the gene sequences, and the program *PROTML*, developed by Hasegawa and Adaki (personal communication), to obtain maximum likelihood trees for amino acid sequences.

The structure and organization of the *his* genes were deduced from the data available for the following organisms: *A. brasilense* (Fani et al. 1989, 1993); *B. subtilis* (Henner et al. 1986); *B. oleracea* (Nagai et al. 1991); *C. maltosa* (Hikiji et al. 1989); *E. coli* and *S. typhimurium* (Carlomagno et al. 1988); *K. pneumoniae* (Rodriguez and West 1984); *H. volcanii* (Conover and Doolittle 1990); *L. lactis* (Delorme et al. 1992); *M. vannielii* (Cue et al. 1985; Beckler and Reeve 1986); *M. voltae* and *M. thermolithotrophicus* (Weil et al. 1987); *M. smegmatis* (Hinshelwood and Stocker 1992); *N. crassa* (Legerton and Yanofsky 1985); *Phytophthora parasitica* (EMBLGenBank accession number Z11591); *Pichia pastoris* (Crane and Gould 1994); *S. cerevisiae HIS1* (Hinnebusch and Fink 1983), *HIS3* (Sthrul 1985), *HIS4* (Donahue et al. 1982), *HIS5* (Nishiwaki et al. 1987), and *HIS7* (Kuenzler et al. 1993); *S. kluyveri* (Weinstock and Strathern 1993); *S. coelicolor* (Limauro et al. 1990, 1992); and *T. harzianum* (Goldman et al. 1992).

Results and Discussion

The Histidine Genes and Their Organization in the Three Cell Domains

In recent years the use of *E. coli* and *S. cerevisiae* as hosts for the cloning and expression of heterologous genes has opened up the possibility of identification and study of *his* genes from different organisms by complementation of a given *E. coli* and/or *S. cerevisiae his* mutation with cloned genes. It has also allowed the identification of the function coded by the heterologous cloned *his* gene. In other cases, the identification of the homologous *his* gene has been accomplished by comparing amino acid sequences with the already-known *his* gene products. As summarized in Table 1, *his* biosynthetic genes have been cloned and sequenced from eight bacteria, four archaea (three of which belong to the genus *Methanococcus*), and eight eucarya. Other *his* genes have been identified, but not yet cloned and/or se-

quenced. The length of orthologous genes from these different sources is in most cases very similar, as are several features of the proteins they encode for, such as molecular weight, hydropathic profiles, and secondary structure predictions (not shown). The only known exception is the *L. lactis hisG* gene, whose length (624 bp) is about one-third shorter than its homologues (Fig. 2). In spite of this important size difference, this gene can complement the *E. coli hisG* mutation (Delorme et al. 1992). It is noteworthy that the conserved features of the gene products lead to exchangeable functional properties (i.e., complementation) among the different species, although the degree of sequence homology among orthologous genes can show considerable variation (Fani et al. 1993). However, this conservation is not paralleled by the structural organization of histidine genes: many alternative gene rearrangements are found, even for closely related microorganisms (Fig. 2).

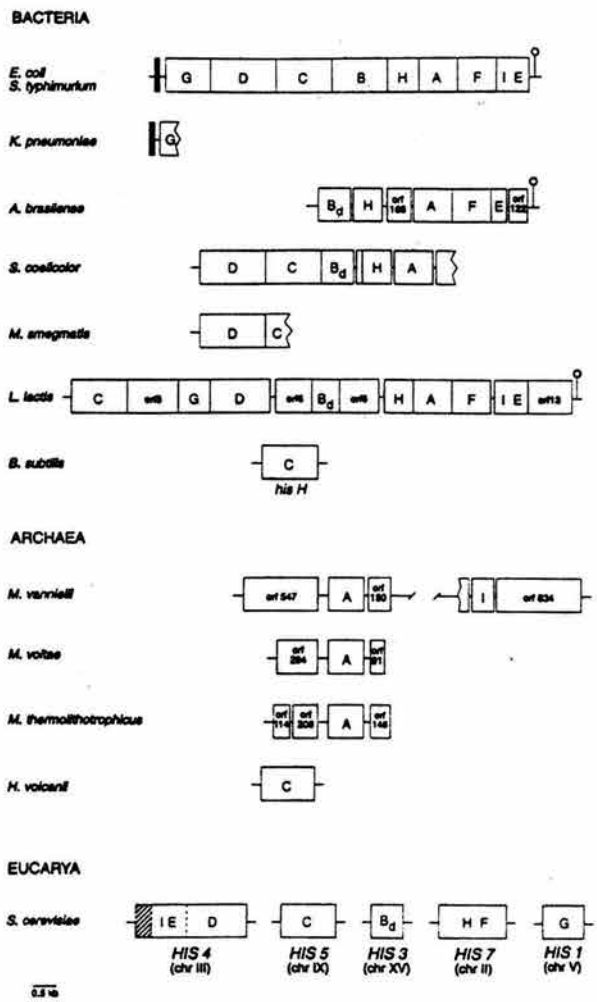


Fig. 2. The organization of some of the histidine biosynthetic genes sequenced to date in different organisms. Symbols: *black boxes*, leader sequences; *stem and loop*, transcription terminators; *open boxes* in the *S. coelicolor* cluster indicate ORFs with unknown functions; *lines* between two genes indicate intergenic regions; the very short intergenic region between the *E. coli hisG* and *hisD* genes is not shown. The *striped box* corresponds to a portion of a gene coding for amino acid sequence not homologous to any other sequence of enzymes involved in histidine pathway identified so far. Since the two activities coded by the *E. coli hisIE* gene are probably coded by two different genes in *A. brasilense* and *M. vannielii*, we have adopted the symbol *hisIE* to identify those genes encoding a bifunctional enzyme endowed with the PR-ATP pyrophosphohydrolase and PR-AMP cyclohydrolase activities, and *hisI* and *hisE* for genes coding a single enzymatic activity. For primary sources from which data on the structure and organization of the *his* genes were deduced, see Materials and Methods.

Bacteria

As shown in Fig. 2, in all the bacteria studied at least some of the *his* biosynthetic genes are clustered in operons of different length. In both *E. coli* and *S. typhimurium* the eight *his* genes are clustered in a compact operon measuring 7,389 and 7,438 bp, respectively, whose order is *hisGDCBHAF(IE)* (Carlomagno et al. 1988). However, in the related nitrogen-fixing eubacterium *K. pneumoniae* only the sequence of a 600-bp fragment containing a truncated open reading frame (ORF) (300 bp) has been reported (Rodriguez and West 1984). Sequence comparison has shown that this truncated ORF encodes a putative protein which is virtually identical (i.e., 88% identity) to the *N*-terminal region of the *E. coli* and *S. typhimurium hisG* gene product, and is preceded by a region that could act as an attenuator (Rodriguez and West 1984). These findings suggest that the structure, organization, and regulation of the *his* genes in *K. pneumoniae* are very similar to *E. coli* and *S. typhimurium*.

In the low G + C Gram-positive bacterium *L. lactis* a *his* operon was recently reported (Delorme et al. 1992, 1993). In this operon 12 different ORFs have been identified, eight of which, *hisCGDBdHAF(IE)*, are homologous to the corresponding *E. coli* genes, whereas the other four encode putative proteins which have no apparent function in histidine biosynthesis. Two of them (ORF6 and ORF13) encode products which have no homology with the proteins present in the available databases. In contrast, ORF8 is homologous to the Alpha-3′ enzymes, which inactivate aminoside antibiotics. Nevertheless, its real function is still unknown (Delorme et al. 1992). ORF3 is homologous to the *E. coli hisS* gene, which encodes the histidyl-tRNA synthetase. It has been postulated (Delorme et al. 1992) that this ORF could play a role in the control of the *L. lactis his* operon.

Gene organization in the high G + C Gram-positive bacterium *S. coelicolor* appears to lie between that of the enterobacteria and *L. lactis*, on the one hand, and that of eukaryotic cells, on the other. In *S. coelicolor* the *his* genes map at three different loci on the chromosome: a cluster maps at 12 o'clock position, one or two genes map at 2 o'clock position, and a single gene, *hisBp.x*, maps close to the 6 o'clock position (Carere et al. 1973; Derkos-Sojak et al. 1985; Hopwood et al. 1985; Russi et al. 1973; Limauro et al. 1992). As shown in Fig. 2, the cloning and sequencing of this *his* cluster (12 o'clock) has shown that it includes at least five genes (*hisDCB dHA*) homologous to the *E. coli his* genes, as well as two

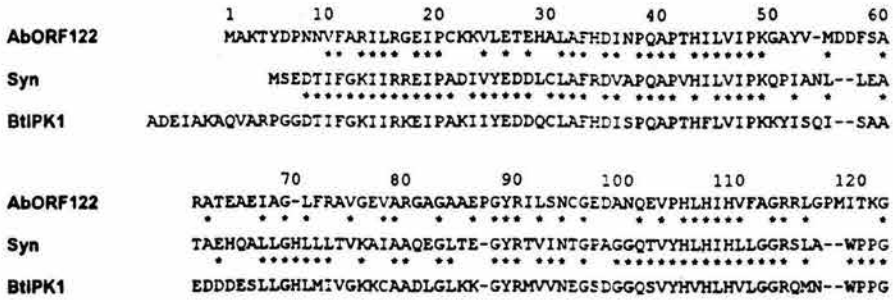


Fig. 3. Alignment of the amino acid sequences of *A. brasilense* ORF 122 (AbORF122), *B. taurus* IPCK-1 (BtIPK1), and *Synhecococcus* sp. ORF1 (Syn). Gaps were introduced for optimal alignment. The numbering does not take into account the gaps in the sequences. *Stars* indicate the position of identical or similar amino acids between two sequences (accepted substitutions: K-R, D-E, S-T, I-L-V-M, F-Y).

additional ORFs. The first of these ORFs lies between the *S. coelicolor hisBd* and the *hisH* genes, and the second one is a truncated (in that only part of it has been sequenced) sequence located downstream from the *hisA* gene (Limauro et al. 1990, 1992).

In the Gram-positive eubacterium *M. smegmatis*, a DNA fragment has been cloned containing two ORFs, the first one of which encodes for a product that shares a high degree of sequence similarity (63% identity; 81% if all conserved residues are included) with the *S. coelicolor hisD* gene product, while the latter is a truncated sequence homologous (60% identity; 89% including conserved residues) to the *hisC* gene (Hinshelwood and Stocker 1992). The absence of a gene homologous to the *E. coli hisG* gene in the 300 bp upstream of *hisD* suggests that the organization of histidine genes in *M. smegmatis* may be similar to *S. coelicolor*, but different from those of enterobacteria and *L. lactis*.

In the α-purple bacterium *A. brasilense* the histidine biosynthetic genes are partially clustered in an operon consisting of five genes encoding proteins whose activity is known (*his BdHAFE*), and two ORFs, ORF168 and ORF122, with no known role in histidine biosynthesis (Fani et al. 1989, 1993). Restriction fragment length polymorphism (RFLP) analysis performed on 19 different *Azospirillum* strains belonging to the three species *A. amazonense*, *A. brasilense*, and *A. lipoferum* has shown that at least three of the *his* genes (*hisBd*, *hisH*, and ORF168) are strongly conserved and arranged in the same way in this genus (Fani et al. 1995). A detailed comparison was undertaken in order to look for homologues of the ORFs associated with the *his* genes. Our search has revealed that the *A. brasilense* ORF122, which is followed downstream by a strong transcription terminator (Fig. 2), shares a considerable sequence similarity with the IPCK-1 *Bos taurus* brain-derived protein inhibitor of protein kinase, (38% identical amino acids, 50% including conserved amino acids) described by Mozier et al. (1991) and with a 342-bp ORF located upstream of a gene of the cyanobacterium *Synechococcus*, which encodes a thylakoid protein that is part of the photosystem II reaction center (34% identity, 48% similarity) (Bustos et al. 1990). The alignment of these three proteins (Fig. 3) shows that, in spite of their different origin, they share two highly conserved domains of about 45 and 35 amino acids, located at the N- and the C-terminus, respectively. Interestingly, the last one contains the sequence His-X-His-X-His, which appears to be a novel zinc binding site, since in the *B. taurus* IPCK-1, where it was first detected, it seems to bind a zinc ion in a 1:1 stechiometric ratio. The homology of these three proteins suggests that they may have similar structure and function in the three organisms. The involvement of ORF122 in histidine biosynthesis or in the regulation of the *A. brasilense his* operon is still to be elucidated. The search for homologues of the deduced amino acid sequence encoded by ORF168 did not reveal any significant degree of sequence similarity with the sequences contained in the EMBLGenBank. Nevertheless, the first 20 amino acids of its putative product share 50% of sequence identity with the *E. coli* and *S. typhimurium hisG* gene product.

In *B. subtilis* (low G + C Gram positive) the *his* genes are separated in at least two chromosomal loci (Piggot and Hoch 1985). Cloning and sequencing of one of these genes, *hisH*, which maps at position 205, have shown that it is homologous to the *E. coli hisC* gene (Henner et al. 1986). All the other *B. subtilis his* genes appear to be grouped in a cluster mapping at a *hisA* chromosomal locus in position 299. Finally, in *Staphylococcus aureus* at least six different histidine genes (*hisE, A, B, C, D,* and *G*) appear to be grouped in a single cluster (Patee et al. 1990).

Archaea

Since the existence of transcriptional units of correlated genes resembling those of bacteria is well established in different archaeal lineages (Zillig et al. 1988; Arndt 1990; Denda et al. 1990; Horne et al. 1991; Auer et al. 1991; Zillig 1991) and includes a tryptophan operon (cf. Doolittle and Brown 1994), there is no a priori reason to doubt the existence of operons which would allow the archaeal *his* genes to be cotranscribed and regulated in a coordinated way. However, little is known about the organization of the histidine biosynthetic genes among the Archaea. Only three different genes, *hisA*, *hisC*, and *hisI*, have been studied in this domain, and two of them, *hisA* and *hisI*, are in the same microorganism,

M. vannielii (Beckler and Reeve 1986). Nonetheless, it is noteworthy that in this archaeon the *hisA* and *hisI* genes are separated by more than 10 kb. In *M. vannielii* the *hisA* gene is transcribed from a promoter located in the intergenic region ORF547-*hisA* (Cue et al. 1985). The *hisI* gene seems to be part of an operon, but neither the upstream ORF nor the one located downstream shares sequence homology with other *his* genes (Beckler and Reeve 1986). Furthermore, the *hisA* genes of the three archaea which have been studied are surrounded by ORFs encoding products of unknown functions and no detectable homology, whereas in eubacteria *hisA* is always preceded by *hisH*, and in most cases followed by *hisF* (Fig. 2).

In spite of the high degree of conservation of the *hisA* gene product in the genus *Methanococcus* (about 67% and 80% of identical and similar amino acids, respectively), large rearrangements have taken place in the flanking regions containing ORFs. The comparison of the nucleotide sequence and the deduced amino acid sequence of these ORFs with the databanks did not reveal a significant degree of similarity with any of the known sequences. Nevertheless, as noted by Weil et al. (1987), some of these ORFs are homologous to one another. For instance, the deduced amino acid sequence of ORF294, ORF114, and ORF547, which is located upstream of the *hisA* gene of *M. voltae*, *M. thermolithotrophicus*, and *M. vannielii*, shares a high degree of sequence similarity (44% and 54%), as do ORF145 and ORF150 (66%), located downstream from the *hisA* gene. On the contrary, ORF91 and ORF206 had no detectable nucleotide or amino acid sequence homology.

Finally, the *H. volcanii hisC* gene is surrounded by DNA sequences that also lack similarity with any of the known *his* genes (Conover and Doolittle 1990).

Thus if the existence of archaeal *his* operons is confirmed, the available data suggest that they may be interrupted, like the *L. lactis* one, by several different ORFs. The apparent large number of archaeal ORFs that may be unrelated to histidine biosynthesis is reminiscent of the existence of large numbers of inserted open reading frames that have been described for the archaebacterial ribosomal protein gene clusters.

Eucarya

Among the eucarya that have been studied the organization of the histidine biosynthetic nuclear genes is completely different from that found in prokaryotes, since no clustered *his* genes have been found in them. Although recently several *his* genes have been cloned from different eukaryotic sources (Table 1), the histidine biosynthetic pathway and its regulation have been extensively studied only in the yeast *S. cerevisiae* (Table 1 and Fig. 2). The enzymatic steps leading to histidine are thought to be identical in *E. coli* and in *S. cerevisiae*, but in the latter the genetic information for the histidine biosynthetic enzymes is encoded by seven genes (Fig. 1), which are located on six different chromosomes. Five of these genes have been cloned and sequenced (*HIS1*, *HIS3*, *HIS4*, *HIS5*, and *HIS7*). Two of them, *HIS7* and *HIS4*, whose structure is discussed in detail below, appear to be the result of fusions of ancestral bacterial cistrons. Histidine biosynthetic genes have also been cloned from *N. crassa (his-3)*, and more recently from other fungi, including *S. kluyveri (K-HIS3)*, *P. parasitica (HIS3)*, *C. maltosa (HIS5)*, *P. pastoris (HIS4)*, and *T. harzianum (igh)*. The only plant histidine gene cloned to date is a cDNA from *B. oleracea* corresponding to the *hisD* gene of enterobacteria, which encodes a bifunctional histidinol-dehydrogenase. The fact that this gene is not fused to a gene coding the HisIE activities as it is in fungi suggests that the organization and structure of the nuclear genes involved in histidine biosynthesis are also variable among eukaryotes (Nagai et al. 1991).

The Structure of his *Genes*

Gene Fusion Events

The ability of both prokaryotic and eukaryotic histidine genes to complement the *E. coli* and the *S. cerevisiae his* mutations (Bazzicalupo et al. 1987; Limauro et al. 1990; Delorme et al. 1992; Goldman et al. 1992; Kuenzler et al. 1993; Weinstock and Strathern 1993) clearly demonstrates that exchangeable functional properties exist among different species, i.e., that the basic structure of the histidine biosynthetic enzymes is highly conserved.

Gene fusion appears to be one of the most important mechanisms of gene evolution in the histidine biosynthetic pathway. Several such events have occurred in both the genomes of bacteria and some eukaryotes, leading to longer genes encoding for bi- or multifunctional enzymes (Figs. 1 and 2 and Table 1). Although gene fusions can be selected for substrate channeling, they also represent an effective mechanism ensuring the coordinate synthesis of two or more enzymatic activities. This may have special significance among nucleated cells, where the absence of operons does not allow coordinate regulation by polycistronic mRNAs (Davidson et al. 1993). In histidine biosynthesis at least five examples of gene fusion and/or bifunctional or multifunctional enzymes can be recognized. As summarized in Figs. 1 and 2, three of the eubacterial *his* genes that have been cloned and sequenced (*hisB*, *hisD*, and *hisIE*) code for bifunctional enzymes.

In the enterobacteria *S. typhimurium* and *E. coli* the fourth gene of the histidine operon, *hisB*, codes for a bifunctional enzyme possessing both IGP dehydratase (EC 4.2.1.19) and HP phosphatase (EC 3.1.3.15) activities. It catalyzes the sixth and the eighth steps of histidine biosynthesis (Winkler 1987). The most widely accepted model for the association of these two enzymatic activities of the *hisB* gene product predicts the existence of

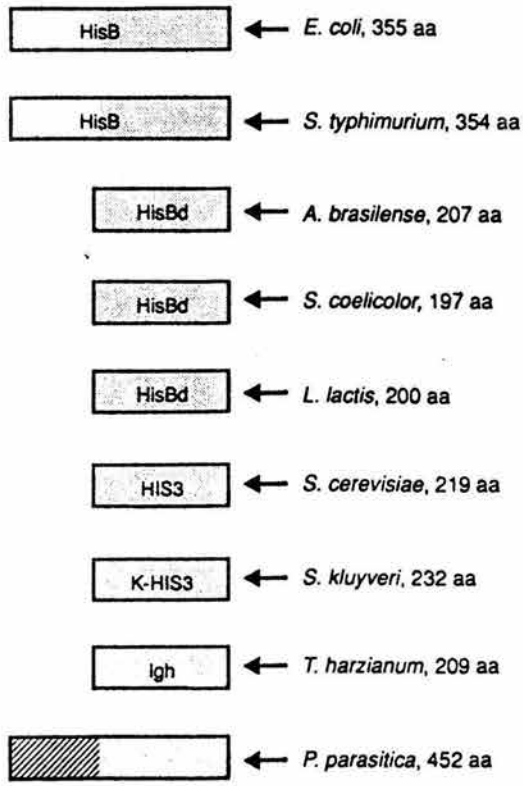


Fig. 4. Structure of the HisB protein from eubacterial and eukaryotic sources.

two independent domains in the gene, i.e., a proximal domain encoding the phosphatase moiety, and a distal one encoding the dehydratase activity. This model is supported by several independent biochemical and genetic lines of evidence (Loper 1961; Brady and Houston 1973; Chumley and Roth 1981).

The structural organization of the two enzymatic activities in some microorganisms supports the two-domain model discussed above (Fig. 4). In *S. cerevisiae* the two activities are encoded by two separate genes, *HIS2* (for phosphatase activity) and *HIS3* (for dehydratase activity) (Broach 1981). The same is true for the *his-1* and *his-4* genes in *N. crassa* (Fink 1964). Genes homologous to the *S. cerevisiae HIS3* gene have been isolated from other eukaryotes. Thus, in the fungus *T. harzianum* an *igh* gene able to complement the *HIS3* mutation of *S. cerevisiae* and encoding an IGP-dehydratase has been cloned (Goldman et al. 1992), and an equivalent gene from *S. kluyveri (K-HIS3)* has also been recently described (Weinstock and Strathern 1993).

In *P. parasitica* a gene encoding a putative IGP dehydratase has also been found. This gene encodes what appears to be a bifunctional protein of 452 amino acids, whose carboxy terminal moiety, spanning from residue 234 to residue 452, exhibits a high degree of sequence similarity with the eubacterial and eukaryotic IGP dehydratases (about 64% and 69% for similar amino acids, respectively). We have found no evidence of sequence similarity between the amino terminal moiety of this putative protein with the amino terminal domain of the enterobacterial HisB enzyme, or with any other His protein. This negative result was confirmed when hydropathic profiles and predicted secondary structures where compared. Assuming that this putative bifunctional protein possesses this form in vivo, it should be concluded that it is encoded by a *hisBd* gene fused to another gene of unknown function. How widespread this situation is among eukaryotes is unknown.

A similar splitting into two genes apparently also took place in some eubacterial branches, as indicated by the gene organization of *S. coelicolor*. In this organism the two activities are encoded by two different genes, *hisBpx* and *hisBd* (Hopwood et al. 1985; Limauro et al. 1990). The same situation may have taken place in the ancestors of *A. brasilense* and *L. lactis*, although the counterpart of the promoter-proximal region (*hisBpx*) of the *E. coli hisB* gene encoding a HP phosphatase has not been identified. In fact, in these two microorganisms only a *hisBd* gene, complementing a mutation in the 3′ region of the *E. coli hisB* gene, has been reported (Fani et al. 1989; Delorme et al. 1992). It has been suggested that in *L. lactis* the gene coding the HP phosphatase could be localized in another region of the chromosome; alternatively, one of the four ORFs with unknown function (ORF8) belonging to the *his* operon could perform the HP dephosphorylation (Delorme et al. 1992).

It has been argued that the *HIS2* and *HIS3* yeast genes evolved through a "split" mechanism from the prokaryotic domains of the *E. coli* bifunctional *hisB* (Glaser and Houston 1974). However, the available data appear to support the alternative view that a bifunctional *hisB* gene is an enterobacterial peculiarity. It seems more likely that the evolution of the *hisB* gene in *E. coli* and *S. typhimurium* could have involved the fusion of two independent cistrons, *hisBpx* and *hisBd*, coding for an IGP dehydratase and a HP phosphatase, respectively (Fani et al. 1989), which may have taken place after the evolutionary split between the α and the γ branches of the purple bacteria.

The second example of a bifunctional enzyme in bacteria is found in the *hisIE* gene product, an enzyme possessing both phosphoribosyl-AMP-cyclohydrolase (PRPC) (HisI) activity, and phosphoribosyl-ATP-pyrophosphohydrolase (PRPI) (HisE) activity, and which catalyzes the second and third steps of histidine biosynthesis. A bifunctional *hisIE* gene has also been identified in *L. lactis* (Delorme et al. 1992), and in the eukaryotes *S. cerevisiae, P. pastoris,* and *N. crassa,* where it appears to be part of larger multifunctional genes. (See below and Figs. 2 and 5). In all of these microorganisms the two moieties maintain the same relative order, with *hisI* always preceding *hisE*. However, in *M. vannielii* a monofunctional *hisI* gene has been isolated that codes for PRPC activity; likewise, in *A. brasilense* the PRPI activity is also encoded by a monofunctional gene (*hisE*). It

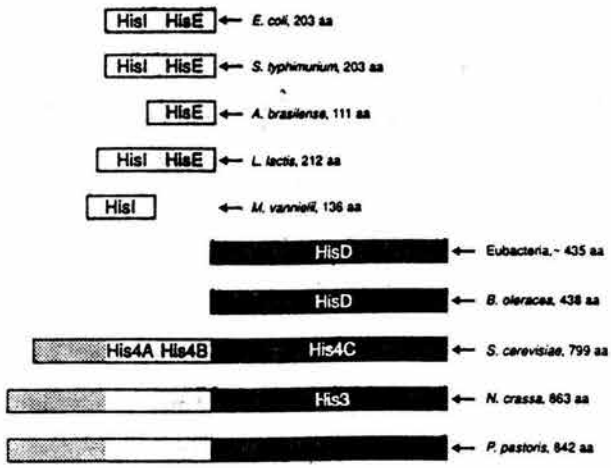


Fig. 5. Structure and comparison of the orthologous proteins with PR-AMP cyclohydrolase (HisI), PR-ATP pyrophosphohydrolase (HisE), and/or histidinol dehydrogenase (HisD) activity. Homologous regions are represented by the same *hatching*.

is thus possible that in these two widely separated microorganisms the two enzymatic activities are encoded by different genes, although the counterparts of the *hisE* and *hisI* genes have not been yet identified. On the basis of the available data, at least two different hypotheses can be proposed. In one of them the existence of an ancestral *hisIE* bifunctional gene which gave rise in some prokaryotes to monofunctional genes by a splitting mechanism should be advocated. An alternative explanation requires the existence of two ancestral genes, each of which encoded a monofunctional enzyme catalyzing sequential steps in histidine biosynthesis. These ancestral cistrons, either adjacent or scattered on the chromosome, could have then undergone independent fusion events in different cell lineages.

The third example of a gene encoding a bifunctional enzyme, that maintains the same structure in all organisms where it has been identified and cloned, is *hisD*. This gene codes for a histidinol dehydrogenase (EC 1.1.1.23), which catalyzes the two final, consecutive steps in histidine biosynthesis. As suggested elsewhere, the N-terminal part of the protein could catalyze the first oxidation step, whereas the second enzymatic function, i.e., the aldehyde dehydrogenase (which catalyzes the oxidation of the aldehyde histidinal to the corresponding amino acid, histidine), may reside in its C-terminal segment (Bruni et al. 1986). The presence of this gene in prokaryotes, fungi, and plants (Table 1 and Fig. 2) also implies that the final steps of the histidine biosynthesis proceed in plants as they do in prokaryotes and fungi, and may indeed be universal among all organisms. Since all known prokaryotic and eukaryotic *hisD* genes share the same unsplit structure, it is reasonable to assume that if this gene arose from the fusion of two independent ancestral cistrons, this event probably took place well before the divergence of the last common ancestor into the three domains. Analysis of the *E. coli hisD* gene product did not reveal any evidence of internal sequence homology.

With the exceptions of the rRNA gene cluster, no operons or transcriptional units of correlated sequences have been found in eukaryotes. However, in nucleated cells coordinate regulation of two or more enzymatic activities can be achieved by gene fusion. In the yeast *S. cerevisiae* at least two different *his* genes, *HIS4* and *HIS7*, could be the result of one (or more) gene fusion event(s). The *HIS4* gene codes for a relatively long polypeptide of 799 amino acids and is able to complement the mutation of *E. coli* or *S. typhimurium* strains altered in the *hisIE* and *hisD* genes. Thus, this enzyme possesses at least four different enzymatic activities. As shown in Figs. 2 and 5, the *his-4* gene from *N. crassa* and the *HIS4* genes from *S. cerevisiae* and *P. pastoris* all share the same internal organization. In fact, in the *HIS4*-encoded protein three functional domains can be recognized, each of which is encoded by a subregion of the *HIS4* gene.

The most significant difference between these three genes is the presence of a 59-bp intron in the *N. crassa his-3* gene in the region encoding the HisD activities. This intron prevents the complementation to the *E. coli hisD* mutation, but not to the *E. coli hisI* mutation (Legerton and Yanofsky 1985). As noted above, this eukaryotic multifunctional enzyme is encoded by a gene that apparently originated from the fusion of bacterial separated cistrons (Bruni et al. 1986), but just how many gene fusion and/or gene elongation events have led to the extant *HIS4* and *his-3* genes is still an open question. Since *HIS4* and *his-3* genes share the same internal organization, it is possible that the putative fusion event(s) leading to the extant genes took place before the separation of these fungi. Nevertheless, the fact that the *hisD* gene from *B. oleracea* is not fused to the *hisIE* gene demonstrates that the structure and organization of the *his* genes could be very different among eukaryotes.

The *S. cerevisiae HIS7* gene also appears to be the result of a fusion event (Kuenzler et al. 1993). This gene codes for a polypeptide able to complement both the *hisH* and *hisF* mutations of *E. coli*. Analysis of the putative encoded enzyme has demonstrated that it shares sequence similarity with the eubacterial *hisH* and *hisF* gene products in both their amino- and carboxy-terminal regions, respectively (Fig. 7). It has also been shown (Klemm and Davisson 1993) that the *E. coli hisF* and *hisH* genes code for proteins that associate to form a heterodimeric enzyme (IGP synthase) catalyzing the transformation of PRFAR into IGP and AICAR via a glutamine molecule. This represent the central step in the pathway that interconnects histidine biosynthesis with nitrogen metabolism and the de novo synthesis of purines (Fig. 1). Perhaps not surprisingly, fusion of different cistrons encoding proteins involved in the same enzymatic step appears to be a frequent strategy among

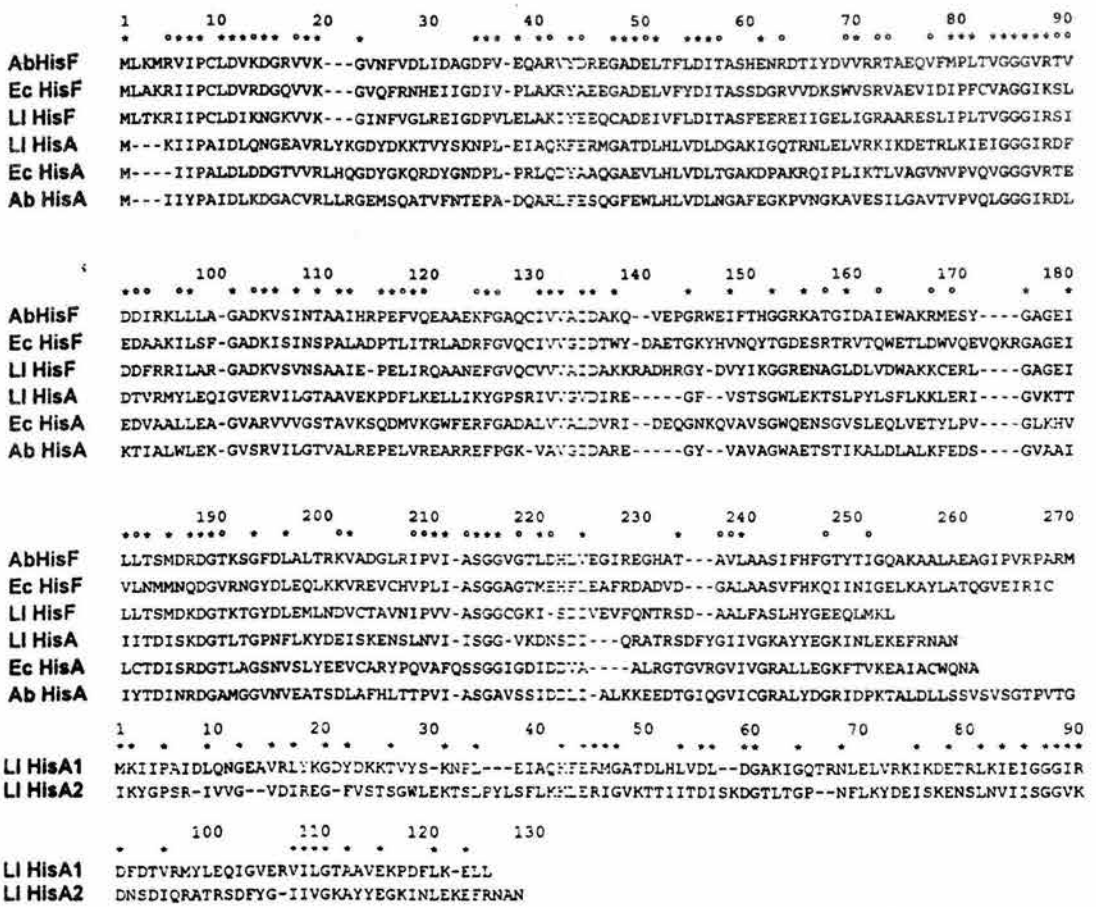


Fig. 6. **Upper** Alignment of the amino acid sequences deduced from the *L. lactis*, *E. coli*, and *A. brasilense hisA* and *hisF* genes. The amino acids are indicated by the *single-letter code*. Gaps were introduced for optimal alignment. *Symbols above* the sequences indicate the position of identical or similar amino acids (accepted substitutions: K-R, D-E, S-T, I-L-V-M, F-Y) in four (*dots*) or at least in five sequences (*stars*). **Lower** Alignment of the amino acid sequences deduced from the 5′-terminal domain (HisA1) and the 3′-terminal domain (HisA2) of the *L. lactis hisA* gene. *Stars above* sequences indicate the position of identical or similar amino acids.

eukaryotes. It is noteworthy that despite the fact that the two bacterial proteins interact to form an active IGP synthase, the genes encoding them are not adjacent in any of the studied microorganisms. It is possible that the yeast gene resulted from domain shuffling. Alternatively, the *HIS7* gene could have originated by the fusion of two genes via the deletion of the intervening region.

Some *his* Genes are the Result of Ancient Paralogous Duplications

Sequence similarity between the *E. coli* HisF protein and the 5′-ProFAR isomerase had been recognized by Sheridan and Venkataraghavan (1992). Further and detailed analysis of the *hisA* and *hisF* genes that code for these two enzymes in *E. coli*, *S. typhimurium*, *A. brasilense*, *S. coelicolor*, *M. voltae*, *M. vannielii*, and *M. thermolithotrophicus* showed that they share a high degree of sequence similarity and may be thus the result of an ancient duplication event that may have taken place prior to the divergence of the last common ancestor of archaebacteria and eubacteria (Fani et al. 1994).

It has also been shown that these two genes share a similar internal organization into two homologous modules half the size of the entire genes (Fani et al. 1994). The comparison of these modules led to the suggestion that both genes are the result of two ancient successive duplication events, the first one involving the *hisA1* module and leading to the extant *hisA* gene, which in turn duplicated and gave rise to the *hisF* gene. This suggests that the two genes must encode proteins with similar activities and/or catalyzing comparable chemical reactions.

We have extended the above analysis to the *hisA* and the *hisF* genes from *L. lactis* and to the *HIS7* gene from *S. cerevisiae*. This comparison has revealed that the two *L. lactis* gene products are also homologous among themselves (29% identity, 47% similarity) and exhibit the same internal two-module organization (36% and 32% similarity for A1 vs A2, and F1 vs F2, respectively) (Fig. 6). The identification of the *hisF* gene from *L. lactis* also suggests that the second duplication giving the *hisF* gene took place before the divergence of the Gram-positive from Gram-negative bacteria.

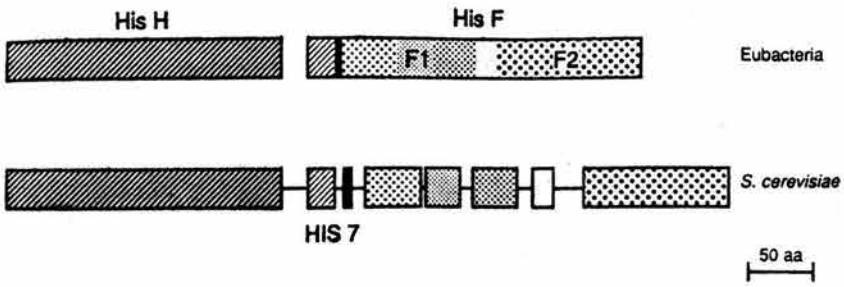


Fig. 7. Comparison of the eubacterial *hisH* and *hisF* gene products with the HIS7 protein from *S. cerevisiae*. *F1* and *F2* represent the two modules of eubacterial HisF protein according to Fani et al. (1994). Homologous regions are represented by the same *hatching*. *Lines* in the HIS7 protein represent regions not homologous to the eubacterial counterparts

The molecular structure of the *S. cerevisiae HIS7* gene is rather interesting. The two moieties of the gene that correspond to the *E. coli hisH* and *hisF* genes, respectively, are linked by a short DNA stretch of 66 bp. Moreover, the *hisH* moiety of the *HIS7* gene has almost the same length as eubacterial homologues sequenced so far (about 600 bp), whereas the *hisF* moiety is much longer (951 bp) than its eubacterial counterparts, which have on the average 770 bps. This difference in size is due to the insertion of six DNA stretches of different length in the yeast gene (Fig. 7). All of these insertions are located in the first half of the *S. cerevisiae HIS7* gene *hisF*-like moiety, which corresponds to the eubacterial *hisF1* module. Analysis of the deduced amino acid sequence of this moiety revealed a high degree of sequence similarity with the *M. voltae* HisA protein (25% identity and 39% similarity if the insertions mentioned above are not considered) (Fig. 8, upper part). Sequence comparisons have also shown that the HisF moiety of the *S. cerevisiae* HIS7 protein is formed by two homologous modules (38% similar amino acids) (Fig. 8 lower), which are half the size of the entire moiety. As previously reported (Klemm and Davidsson 1993), the *E. coli* HisF protein must interact with HisH to catalyze the transformation of PRFAR into IGP (Fig. 1). Thus, it is possible that one of the two HisF modules interacts with HisH, whereas the other one is directly involved in the catalytic site. In *S. cerevisiae*, where the *hisH* and *hisF* bacterial counterparts are fused into the *HIS7* gene, the second half of the HisF (i.e., HisF2) moiety of the product of this gene is strongly conserved, whereas F1 showed a strong rearrangement with six different insertions (Fig. 7). This appears to favor the possibility that the F2 moiety is directly involved in catalysis. Unfortunately, although it has already been identified as *HIS6*, the *S. cerevisiae* gene homologous to *hisA* has not been cloned yet. However, the available data suggest that two successive paralogous gene duplications took place long before the diversification of the three domains.

Sequence similarity between the imidazole glycerol phosphate synthases (IGP synthases) with the so-called G-type amidotransferases (GAT) shows that the genes encoding these different enzymes are the result of ancient paralogous duplications. Figure 9 shows an unrooted tree depicting the phylogenetic relationship between the IGP synthases encoded by the enterobacterial *hisH* genes, and the anthranilate synthase (Nichols et al. 1980), 4-amino-4-deoxychorismate synthase (Kaplan and Nichols 1983), carbamoyl-P synthase (Piette et al. 1984), GMP synthase (Tiedeman et al. 1985), CTP synthase (Weng et al. 1986), and formylglycinamide synthase (Schendel et al. 1989; Sampei and Mizobuchi 1989).

Structural Organization of the *his* Operons

As shown in Fig. 2, differences in the relative gene order may be observed in those prokaryotes in which at least some of the histidine biosynthetic genes are clustered, such as *E. coli*, *S. typhimurium*, *A. brasilense*, *S. coelicolor*, and *L. lactis*. Nevertheless, three of the clustered genes (*hisBd*, *hisH*, and *hisA* and also *hisF*, except for *S. coelicolor*) are always present and appear in the same relative order. Although the gene order is not colinear to the enzymatic reactions in the histidine pathway, these three genes encode enzymes involved in the central, sequential enzymatic steps of the pathway. Indeed, the *hisH* and *hisF* genes encode two polypeptides which interact forming a heterodimer that catalyzes the reaction connecting histidine biosynthesis with nitrogen metabolism and with the de novo synthesis of purines. As shown in Fig. 1, the substrate of IGP synthase is PRFAR, the product of the enzymatic step catalyzed by the HisA protein, whereas IGP, which is one of the products of the reaction catalized by IGP synthase, is the substrate of the IGP dehydratase encoded by the *hisBd* gene. Thus it is possible that the four genes *hisBdHAF* could represent the core of the histidine biosynthesis.

Molecular Phylogenies of the his *Genes*

A molecular phylogenetic analysis of the histidine genes was done in order to compare the possible evolutionary relationships among the organisms from which the different genes involved in the pathway have been sequenced. We have performed a phylogenetic analysis for each set of homologous gene and protein sequences available, by using both parsimony and maximum likelihood methods, as described in the Materials and Methods section. The unrooted maximum likelihood phyloge-

```
                    10        20        30        40        50        60        70        80        90
                    ** ****   *   **  *     *  **    *** * * *   *   *   * *     * * *  *****
Sce HisF            GLTRRIIACLDVRTNDVVTK--GDKYKGVRNLGKPVQLAQKYYQQGADEVTFLNITSFRDCP-LKDTEVLKQA-AKTVFVPLTVGGGIKD
Mvo HisA            MY---IIPAVDMKEGKCVQLIQGDPTK-RHEYDNPDEIAKMWIENGAEMLHLVDLDGAIDGERVNLPCVKKIIQESKVPVQM--GGGIRT

                    100       110       120       130       140       150       160       170       180
                    * *       *  ** *** **   * * * *  * *   *** ** * * *         *       ** * *   *   *     *
Sce HisF            IVDASLYFRSGADKVSIGTDAVYAAEKYYE-LGNRYGAQAVVISVDPKNGEKYCWYQCTIKGGRESRDLG-VWELTRACEALGA---GEI
Mvo HisA            IQDVEELVDLGINKVIIGTVAVQNPD-FVEQLAKKVGSDKIMVALDAKDG-KVVIKGWKEKTEYTPVQMGKILEEKGAGSILFTNVDSEG

                    190       200       210       220       230       240       250       260       270
                    ***     * *     ***    * * ****** *    *   * * *   *   **  ** *  * * *     *     **
Sce HisF            LLNCIDKDGSNSGYDLELIEHVKDAVKIPVIASSGAGVPEHFEEAFLKT-RADACLGAGMFHRGEFTVNDVKEYLLEHGLKVRMDEE
Mvo HisA            LLN-----GINITPTKELV----DNLNIPIIASGGVTTIEDLIE-FKKIGVAGVVVGSALY-KNNFKLQDAINTVNNF--KI
```

```
                    10        20        30        40        50        60        70        80        90
                    **  ** * **  *  ** *   * **   * *      ** **  **        *  *  *   *  * * * * ***   *
Sce HisF1           GLTRRIIACLDVRTNDVVTK-GDKYKGVRNLGK-PVQLAQKYYQQGADEVTPLN------ITSFRDCPLKDTEVLKQAAKTVFPVLT-VG
Sce HisF2           -----VVISVDPK-NEQCTIKGGRES--RDLGVWELTRACEAL--GAGEIL-LNCIDKDGSNSGYDLELI--EHVKDAVK-I-PVIASSG

                100       110       120       130       140       150
                    ** *  *   * ** **   **   **   *    **
Sce HisF1           GGIKDIVDASLYFRSGADKVSIGTDAVY-AAEKYY-ELGNRYGAQA
Sce HisF2           AGVPEHFE-EAFLKTRADA-CLGAGMMFHRGEFTVNDVKEYLLEHGLKVRMDEE
```

Fig. 8. Alignment of the amino acid sequences deduced from the *S. cerevisiae* 3′ moiety of *HIS7* gene (Sce HisF) and the *M. voltae hisA* gene (Mvo HisA) **(upper)** and from the 5′-terminal domain (HisF1) and the 3′-terminal domain (HisF2) of the 3′ moiety of the *S. cerevisiae* *HIS7* gene **(lower)**. The amino acids are indicated by the *single-letter code*. Gaps were introduced for optimal alignment. *Stars above* the sequences indicate the position of identical or similar amino acids (accepted substitutions: K-R. D-E. S-T. I-L-V-M. F-Y).

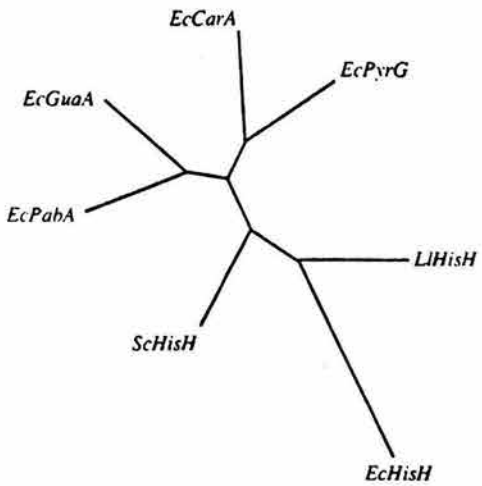


Fig. 9. Unrooted phylogenetic tree constructed using the maximum likelihood method based on the following protein sequences: *E. coli* CarA. GuaA. HisH. PabA. and PyrG (EcCarA. EcGuaA. EcHisH. EcPabA and EcPyrG). *L. lactis* HisH (LlHisH). *S. cerevisiae* HisH moiety of HIS7 protein (ScHisH).

netic trees calculated for the *his* gene products are shown in Fig. 10.

Conclusions

Histidine biosynthesis is an ancient metabolic route that it is tightly connected with several other fundamental pathways, such as those leading to the de novo synthesis of purines and nitrogen metabolism. Its degree of integration and connection with overall cellular metabolism must have involved not only a rather large number of steps involving duplication, fusion, and elongation events involving different *his* genes, but also the many constraints imposed by its coevolution with the components of other pathways and processes. If histidine was required by primitive catalysts, then the eventual exhaustion of the prebiotic supply of histidine and histidine-containing peptides (Shen et al. 1990a–c) must have imposed a selective pressure favoring those organisms capable of synthesizing imidazole-containing compounds. How the biosynthesis of histidine actually emerged can only be surmised, but the phylogenetic distribution of the genes involved in synthesis strongly suggests that the entire pathway existed in the last common ancestor of the three extant domains. This conclusion is supported by the fact that at least two different *his* genes (*hisC* and *hisI*) have been identified in each of the three cell lineages, suggesting that they may have been part of the genome of the last common ancestor. This possibility is strongly supported by the robustness of the trees depicting the evolutionary distances between different *his* genes, which display the same general topology and similar interdomain relationships (Fig. 10). This suggests that the evolution of the 16S-like rRNA and the *his* genes was roughly parallel.

The evolutionary comparison of the *his* genes in the

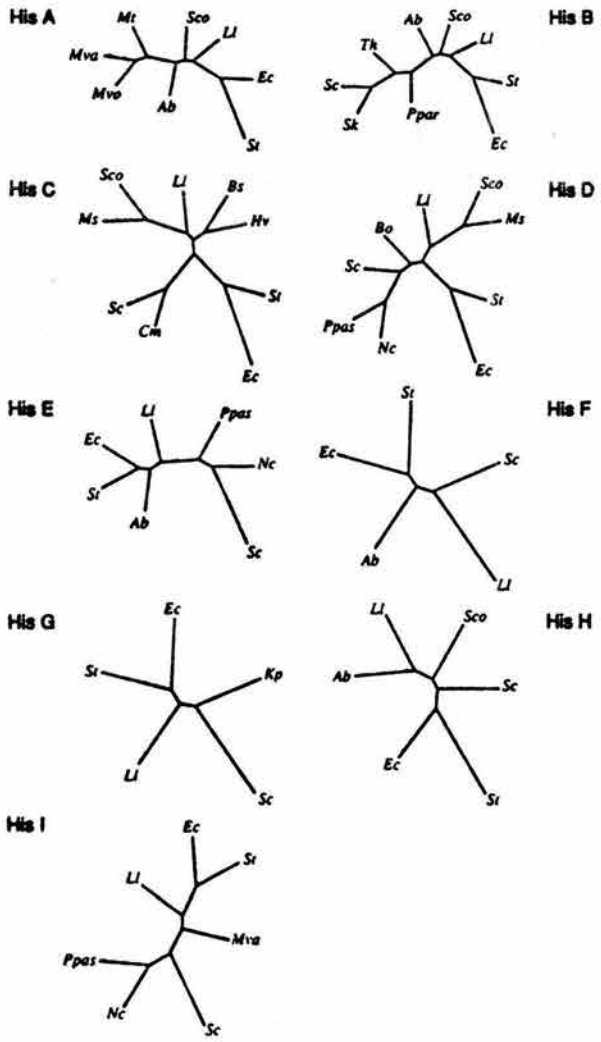


Fig. 10. Phylogenetic trees constructed using the maximum likelihood method based on histidine proteins sequence. Abbreviations: *Ab = Azospirillum brasilense. Bo = Brassica oleracea. Bs = Bacillus subtilis. Cm = Candida maltosa. Ec = Escherichia coli. Hv = Halobacterium volcanii. Kp = Klebsiella pneumoniae. Ll = Lactococcus lactis. Ms = Mycobacterium smegmatis. Mt = Methanococcus thermolithotrophicus. Mva = Methanococcus vannielii. Mvo = Methanococcus voltae. Nc = Neurospora crassa; Ppas = Pichia pastoris: Ppar = Phytophthora parasitica; Sc = Saccharomyces cerevisiae; Sco = Streptomyces coelicolor: Sk = Saccharomyces kluyveri: St = Salmonella typhimurium: Th = Thricoderma harzianum.*

different cell lineages clearly indicates that their structure, organization, and order has undergone several major rearrangements. In particular, many fusion events have occurred and the organization of *his* genes appears to be rather different in the different microorganisms. Nevertheless, the *hisBd, H, A* (and often *F*) eubacterial genes were always found to be part of an operon (Fig. 2), where they are contiguous and arranged in the same order in all eubacteria from which they have been sequenced. The proteins encoded by these genes catalyze the sequential reaction steps and probably represent the core of the histidine biosynthetic pathway.

The question of whether these four genes could have acted as a unitary block in the evolution of eubacterial *his* genes will be solved when additional *his* operons from other different eubacteria are described in detail.

Perhaps one of the most important aspects revealed by the cladistic analysis of the available *his* sequences is the role that paralogous gene duplications have played in shaping the pathway. This is indicated by the evidence of two successive duplications involving an ancestral module which eventually led to the *hisA* and *hisF* genes and their homologues (Fani et al. 1994). These events not only point to the significance of duplication events in shaping the encoding abilities of ancient genomes (Labedan and Riley 1995; Lazcano and Miller 1994), but also have been interpreted (Fani et al. 1994) as supporting the hypothesis of a retrograde evolution of metabolic pathways, according to which biosynthetic abilities are the result of the stepwise, sequential acquisition of enzymes in reverse order, as found in extant pathways (Horowitz 1945).

The only available examples of homologous enzymes catalyzing sequential steps in the same biosynthetic route of which we are aware include (1) eubacterial β-cystathionase, which is homologous to cystathione γ-synthase, the enzyme that catalyzes the preceding step in methionine biosynthesis (Belfaiza et al. 1986); (2) protochlorophyllide reductase and chlorin reductase, which are involved in bacteriochlorophyll synthesis (Burke et al. 1993); and (3) the products of the *hisA* and *hisF* genes (Fani et al. 1994). It is quite possible that this list will increase as more sequences became available. However, the existence of contiguous duplicated genes does not constitute by itself conclusive evidence of the retrograde hypothesis. Alternative explanations, based on the so-called patchwork hypothesis (Ycas 1974; Jensen 1976), are also feasible. According to this idea, primitive metabolic routes were mediated by enzymes of low substrate specificity that were eventually recruited into different pathways (Jensen 1976). This hypothesis is also consistent, however, with the possibility that an ancestral pathway may have had a primitive enzyme catalyzing two or more similar reactions on related substrates of the same metabolic route and whose substrate specificity was refined as a result of later duplication events.

The possibility that histidine biosynthesis was originally mediated by less specific enzymes is in fact strongly supported by common origin of the imidazole glycerol-P synthase, encoded by the enterobacterial *hisH* gene, with other *E. coli* G-type glutamine amidotransferases which participate in the biosynthesis of purines, pyrimidines, arginine, tryptophan, and other ancient pathways. Although the present lack of sequences of G-type glutamine amidotransferases from the three domains limits a complete evolutionary analysis, the phylogenetic distribution of these different G-type GATs suggests that they are the result of several duplications that took place long before the divergence of the three domains and that they may be the descendants of an ancient, less-specific glutamine amidotransferase that

mediated the transfer of the amide group of glutamine to a wide range of substrates.

Although the phylogenetic comparison of the available *his* genes sequences is consistent with the existence of three cell domains, a detailed analysis of the trees drawn in Fig. 10 also shows that the nitrogen-fixing α-purple Gram-negative bacterium *A. brasilense* appears in most cases to be nearer to Gram-positive bacteria than to its close relatives, the γ purple enterobacteria *E. coli* and *S. typhimurium*. The same is true for the *K. pneumoniae*, when the first 100 amino acids that are available for its HisG gene product are compared. Whether the peculiar position of these two nitrogen-fixing eubacteria reflects an ancient lateral gene transfer event is not clear for the time being.

Finally, it should be added that the trees depicting the phylogenies of the *hisC* and *hisI* gene products point toward the evolutionary proximity of the low GC Gram-positive branch including *B. subtilis* and *L. lactis* to the Archaea. This observation is consistent with the phylogenetic distribution of multimetabolite control by feedback inhibition of prephenate dehydratase in aromatic acid biosynthesis (Fischer et al. 1993), as well as with results achieved by the comparison of glutamate dehydrogenase (Benachenhou-Lafha et al. 1993), heat-shock proteins (Gupta and Golding 1993), glutamine synthetase (Kumada et al. 1993; Tiboni et al. 1993; Brown et al. 1994), and carbamoyl synthetases (Lazcano, Puente, and Gogarten, unpublished results), all of which have been interpreted as indicating an early massive lateral gene transfer event between the ancestors of both Archaea and Gram-positive bacteria (Gogarten 1994). If this event actually took place, there may be an evolutionary correlation between the ORFs that appear to surround the methanogenic archaeal *his* genes and the inserted ORFs that have been detected in the low GC Gram-positive *L. lactis his* operon (Fig. 2).

Acknowledgments. We are grateful to Professors M. Polsinelli (R.F. and P.L.), M. Riley, and S. L. Miller (A.L.) for many useful discussions on the issue of evolution of metabolic pathways. We are deeply indebted to John Orò, Emile Zuckerkandl, and an anonymous reviewer for their many useful suggestions and assistance in improving the manuscript.

References

Altboum Z, Gottlieb S, Lebens GA, Polacheck I, Segal E (1990) Isolation of the *Candida albicans* histidinol dehydrogenase (*HIS4*) gene and characterization of a histidine auxotroph. J Bacteriol 172: 3898–3904

Arndt E (1990) Nucleotide sequence of four genes encoding ribosomal proteins from the "S10 and Spectinomycin" operon equivalent region in the archaebacterium *Halobacterium marismortui*. FEBS Lett 267:193–198

Auer J, Spicker G, Mayerhoff L, Puhler G, Bock A (1991) Organization and nucleotide sequence of a gene cluster comprising the translation elongation factor 1α from *Sulfolobus acidocaldarius*. System Appl Microbiol 14:14–22

Bazzicalupo M, Fani R, Gallori E, Turbanti L, Polsinelli M (1987) Cloning of the histidine, pyrimidine and cysteine genes of *Azospirillum brasilense:* expression of pyrimidine and three clustered histidine genes in *Escherichia coli*. Mol Gen Genet 206:76–80

Beckler GS, Reeve JN (1986) Conservation of primary structure in the *hisI* gene of the archaebacterium *Methanococcus vannielii*, the eubacterium *Escherichia coli* and the eucaryote *Saccharomyces cerevisiae*. Mol Gen Genet 204:133–140

Belfaiza J, Parsot C, Martel A, Bouthier de la Tour C, Maragarita D, Cohen GN, Saint-Girons I (1986) Evolution in biosynthetic pathways: two enzymes catalyzing consecutive steps in methionine biosynthesis originate from a common ancestor and possess a similar regulatory region. Proc Natl Acad Sci USA 83:867–871

Benachenhou-Lafha N, Forterre, P, Labedan B (1993) Evolution of glutamate dehydrogenase genes: evidence for two paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life. J Mol Evol 36:335–346

Brady DR, Houston LL (1973) Some properties of the catalytic sites of imidazoleglycerolephosphate dehydratase-histidinol phosphate phosphatase, a bifunctional enzyme from *Salmonella typhimurium*. J Biol Chem 248:2588–2592

Brenner M, Ames BN (1971) The histidine operon and its regulation. In: Vogel HS (ed) *Metabolic pathways*, vol 5. Academic Press, New York, pp 349–387

Broach JR (1981) Genes of *Saccharomyces cerevisiae*. In: Strathern JN, Jones EW, Broach JR (eds) The molecular biology of the yeast *Saccharomyces:* life cycle and inheritance. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp 653–727

Brown JR, Masuchi Y, Robb FT, Doolittle WF (1994) Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. J Mol Evol 38:566–576

Bruni CB, Carlomagno MS, Formisano S, Paolella G (1986) Primary and secondary structural homologies between the *HIS4* gene product of *Saccharomyces cerevisiae* and the *hisIE* and *hisD* gene products of *Escherichia coli* and *Salmonella typhimurium*. Mol Gen Genet 203:389–396

Burke DH, Hearst JE, Sidow A (1993) Early evolution of photosynthesis: clues from nitrogenase and chlorophyll proteins. Proc Natl Acad Sci USA 90:7134–7138

Bustos SA, Schaefer MR, Golden, SS (1990) Different and rapid responses of four cyanobacterial transcripts to changes in light intensity. J Bacteriol 172:1998–2004

Carere A, Russi S, Bignami M, Sermonti G (1973) An operon for histidine biosynthesis in *Streptomyces coelicolor* I. Genetic evidence. Mol Gen Genet 123:219–224

Carlomagno MS, Chiarotti L, Alifano P, Nappo AG, Bruni CB (1988) Structure of the *Salmonella typhimurium* and *Escherichia coli* K-12 histidine operons. J Mol Biol 203:585–606

Chumley FG, Roth JR (1981) Genetic fusions that place the lactose genes under histidine operon control. J Mol Biol 145:697–712

Conover RK, Doolittle WF (1990) Characterization of a gene involved in histidine biosynthesis in *Halobacterium (Haloferax)volcanii:* isolation and rapid mapping by transformation of an auxotroph with cosmid DNA. J Bacteriol 172:3244–3249

Crane DJ, Gould SJ (1994) The *Pichia pastoris HIS4* gene: nucleotide sequence, creation of a non-reverting *his4* mutant, and development of *HIS4*-based replicating and integrating plasmids. Curr Genet 26:443–450

Cue D, Bekler G, Reeve J, Konisky J (1985) Structure and sequence divergence of two archaebacterial genes. Proc Natl Acad Sci U S A 82:4207–4211

Davidson JN, Chen KC, Jamison RS, Musmanno LA, Kern CB (1993) The evolutionary history of the first three enzymes in pyrimidine biosynthesis. Bioessays 15:157–164

Delorme C, Ehrlich SD, Renault P (1992) Histidine biosynthesis genes in *Lactococcus lactis* subsp. *lactis*. J Bacteriol 174:6571–6579

Delorme C, Godon JJ, Ehrlich SD, Renault P (1993) Gene inactivation

in *Lactococcus lactis*: histidine biosynthesis. J Bacteriol 175:4391–4399

Denda K, Konishi J, Hajiro K, Oshima T, Dale T, Yosshida M (1990) Structure of an ATPase operon of an acidotermophilic archaebacterium, *Sulfolobus acidocaldarius*. J Biol Chem 265:21509–21513

Derkos-Sojak V, Pigac J, Delic V (1985) Biochemical and genetic studies of a histidine regulatory mutant of *Streptomyces coelicolor* A3(2). J Basic Microbiol 25:479–485

Donahue TF, Farabaugh PJ, Fink GR (1982) The nucleotide sequence of the *His4* region of yeast. Gene 18:47–59

Doolittle FW, Brown JR (1994) Tempo, mode, the progenote, and the universal root. Proc Natl Acad Sci USA 91:6721–6728

Fani R, Bazzicalupo M, Damiani G, Bianchi A, Schipani C, Sgaramella V, Polsinelli M (1989) Cloning of the histidine genes of *Azospirillum brasilense*: organization of the *ABFH* gene cluster and nucleotide sequence of the *hisB* gene. Mol Gen Genet 216:224–229

Fani R, Alifano P, Allotta G, Bazzicalupo M, Carlomagno MS, Gallori E, Rivellini F, Polsinelli M (1993) The histidine operon in *Azospirillum brasilense*: organization, nucleotide sequence and functional analysis. Res Microbiol 144:187–200

Fani R, Liò P, Chiarelli I and Bazzicalupo M (1994) The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the *hisA* and *hisF* genes. J Mol Evol 38:489–495

Fani R, Bandi C, Bazzicalupo M, Damiani G, Di Cello F, Fancelli S, Gallori E, Gerace L, Grifoni A, Liò P, Mori E (1995) Phylogenetic studies of the genus *Azospirillum*. In Proceedings of the NATO Advanced Research Workshop on *Azospirillum* and Related Microorganisms. Sarvar, Hungary, September 4–7, 1994 (in press)

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Fink GR (1964) Gene-enzyme relations in histidine biosynthesis in yeast. Science 146:525–527

Fischer RS, Bonner CA, Boone DR, Jensen RA (1993) Clues from a halophilic methanogen about aromatic amino acid biosynthesis in archaebacteria. Arch Microbiol 160:440–446

Fox GE, Pechmann KR, Woese CR (1977) Comparative cataloging of 16S rRNA: a molecular approach to prokaryotic systematics. Int J Syst Bacteriol 27:44–57

Glaser RD, Houston LL (1974) Subunit structure and photoxidation of yeast imidazole glycerolphosphate dehydratase. Biochemistry 13: 5145–5152

Gogarten JP (1994) Which is the most conserved group of proteins? Homology, orthology, paralogy and the fusion of independent lineages. J Mol Evol 39:541–543

Goldman GH, Demolder J, Dewaele S, Herrera-Estrella A, Geremia RA, Van Montagu M, Contreras R (1992) Molecular cloning of the imidazoleglycerolphosphate dehydratase gene of *Trichoderma harzianum* by genetic complementation in *Saccharomyces cerevisiae* using a direct expression vector. Mol Gen Genet 234:481–488

Granick S (1965) Evolution of heme and chlorophyll. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 67–88

Gupta RS, Golding GB (1993) Evolution of the HSP70 Gene and its implication regarding relationships between Archaebacteria, Eubacteria, and Eukaryotes. J Mol Evol 37:573–582

Henner DJ, Band L, Flaggs G, Chen E (1986) The organization and nucleotide sequence of the *Bacillus subtilis hisH, tyrA* and *aroE* genes. Gene 49:147–152

Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. Gene 73:237–244

Hikiji T, Okhuma M, Takagi M, Yano K (1989) An improved host-vector system for *Candida maltosa* using a gene isolated from its genome that complements the *his5* mutation of *Saccharomyces cerevisiae*. Curr Genet 16:261–266

Hinnebusch AG, Fink GR (1983) Repeated DNA sequences upstream from *HIS1* also occur at several other co-regulated genes in *Saccharomyces cerevisiae*. J Biol Chem 258:5238–5247

Hinshelwood S, Stoker NG (1992) Cloning of mycobacterial histidine synthesis genes by complementation of a *Mycobacterium smegmatis* auxotroph. Mol Microbiol 6:2887–2895

Hopwood DA, Bibb M, Chater KF, Kieser T, Bruton CJ, Kieser HN, Lydiate D, Smith C, Ward JM, Schrempf H (1985) Genetic manipulation of Streptomyces. A laboratory manual. The John Innes Foundation, Norwich, p 356

Horne M, Englert C, Wimmer C, Pfeifer F (1991) A DNA region of 9 Kbp contains all genes necessary for gas vesicle synthesis in halophilic archaebacteria. Mol Microbiol 5:1159–1174

Horowitz NJ (1945) On the evolution of biochemical synthesis. Proc Natl Acad Sci USA 31:153–157

Horowitz NJ (1965) The evolution of biochemical synthesis-retrospect and prospect. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 15–23

Jensen RA (1976) Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409–425

Kaplan JB, Nichols BP (1983) Nucleotide sequence of *Escherichia coli pabA* and its evolutionary relationships to *trp(G)D*. J Mol Biol 168:451–468

Klemm T, Davisson VJ (1993) Imidazole glycerol phosphate synthase: the glutamine amidotransferase in histidine biosynthesis. Biochemistry 32:5177–5186

Kuenzler M, Balmelli T, Egli CM, Paravicini G, Braus GH (1993) Cloning, primary structure, and regulation of the *HIS7* gene encoding a bifunctional glutamine amidotransferase: cyclase from *Saccharomyces cerevisiae*. J Bacteriol 175:5548–5558

Kumada Y, Benson DR, Hillemann D, Hosted TJ, Rochford DA, Thompson CJ, Wohlleben W, Tateno Y (1993) Evolution of the glutamine synthase gene, one of the oldest and functioning genes. Proc Natl Acad Sci USA 90:3009–3013

Labedan B, Riley M (1995) Widespread protein sequence similarities: origins of *Escherichia coli* genes. J Bacteriol 177:1585–1588

Lazcano A, Fox GE, Otó J (1992) Life before DNA: the origin and evolution of early Archean cells. In: Mortlock RP (ed) The evolution of metabolic function. CRC Press, Boca Raton, FL, pp 237–339

Lazcano A, Miller SL (1994) How long did it take for life to appear and evolve to cyanobacteria? J Mol Evol 39:546–554

Legerton TL, Yanofsky C (1985) Cloning and characterization of the multifunctional *his-3* gene of *Neurospora crassa*. Gene 39:129–140

Li WH, Graur D (1991) Fundamentals of molecular evolution. Sinauer, Sunderland, MA

Limauro D, Avitabile A, Cappellano C, Puglia AM, Bruni CB (1990) Cloning and characterization of the histidine biosynthetic gene cluster of *Streptomyces coelicolor* A3(2). Gene 90:31–41

Limauro D, Avitabile A, Puglia AM, Bruni CB (1992) Further characterization of the histidine gene cluster of *Streptomyces coelicolor* A3(2): nucleotide sequence and transcriptional analysis of *hisD*. Res Microbiol 143:683–693

Loper JC (1961) Enzyme complementation in mixed extracts of mutants from the *Salmonella* histidine B locus. Proc Natl Acad Sci USA 47:1440–1450

Maurel MC, Ninio J (1987) Catalysis by a prebiotic nucleotide analog of histidine. Biochimie 69:551–553

Mozier NM, Walsh MP, Pearson JD (1991) Characterization of a novel zinc-binding site of protein kinase C inhibitor-1. FEBS Lett 279: 14–18

Nagai A, Ward E, Beck J, Tada S, Chang JY, Scheidegger A, Ryals J (1991) Structural and functional conservation of histidinol dehydrogenase between plants and microbe. Proc Natl Acad Sci USA 88:4133–4137

Nichols BP, Miozzari GF, van Cleemput M, Bennett GN, Yanofsky C (1980) Nucleotide sequence of the *trpG* regions of *Escherichia coli, Shigella dysenteriae, Salmonella typhimurium* and *Serratia marcescens*. J Mol Biol 142:503–517

Nishiwaki K, Hayashi N, Irie S, Chung DH, Harashima S, Oshima Y

(1987) Structure of the yeast *HIS5* gene responsive to general control of amino acid biosynthesis. Mol Gen Genet 208:159–167

Patee PA, Lee HC, Bannantine JP (1990) Genetic and physical mapping of *Staphylococcus aureus*. In: Novick RP (ed) Molecular biology of the *Staphylococci*. VCH Publishers, New York, pp 42–56

Piette J, Nyunoya H, Lusty CJ, Cunin R, Weyens G, Crabeel M, Charlier D, Glansdorf N, Pierard A (1984) DNA sequences of the *carA* gene and the control region of *carAB*: tandem promoters, respectively controlled by arginine and the pyrimidines, regulate the synthesis of carbamoyl-phosphate synthetase in *Escherichia coli* K-12. Proc Natl Acad Sci USA 81:4134–4138

Piggot PJ, Hoch JA (1985) Revised genetic linkage map of *Bacillus subtilis*. Microbiol Rev 49:158–179

Rodriguez RL, West RW, Tait RC, Jaynes JM, Shanmugam KT (1981) Isolation and characterization of the *hisG* and *hisD* genes of *Klebsiella pneumoniae*. Gene 16:317–320

Rodriguez RL, West RW (1984) Histidine operon control region of *Klebsiella pneumoniae*: analysis with an *Escherichia coli* promoter-probe plasmid vector. J Bacteriol 157:764–771

Russi S, Carere A, Siracusano A, Ballio A (1973) An operon for histidine biosynthesis in *Streptomyces coelicolor*. II. Biochemical evidence. Mol Gen Genet 123:225–232

Sampei G, Mizobuchi K (1989) The organization of *purL* gene encoding 5′ phosphoribosyl-formyl-glycinamide amidotransferase of *Escherichia coli*. J Biol Chem 264:21230–21238

Schendel FJ, Mueller E, Stubbe J, Shiau A, Smith JM (1989) Formylglycinamide ribonucleotide synthetase from *Escherichia coli*: cloning, sequencing, overproduction, isolation and characterization. Biochemistry 28:2459–2471

Shen C, Mills T, Oro J (1990a) Prebiotic synthesis of histidyl-histidine. J Mol Evol 31:175–179

Shen C, Yang L, Miller SL, Oro J (1990b) Prebiotic synthesis of histidine. J Mol Evol 31:167–174

Shen C, Lazcano A, Oro J (1990c) The enhancement activities of histidyl-histidine in some prebiotic reactions. J Mol Evol 31:445–452

Sheridan RP, Venkataraghavan R (1992) A systematic search for protein signature sequences. Proteins 14:16–28

Sthrul K (1985) Nucleotide sequence and transcriptional mapping of the yeast *pet56-his3-ded1* gene region. Nucleic Acids Res 13:8587–8601

Tiboni O, Cammarano P, Sanangelantoni AM (1993) Cloning and sequencing of the gene encoding glutamine synthase I from the archaeum *Pyrococcus woesei*: anomalous phylogenies inferred from analysis of archaeal and bacterial glutamine synthase I sequences. J Bacteriol 175:2961–2969

Tiedeman AA, Smith JM, Zalkin H (1985) Nucleotide sequence of the *guaA* gene encoding GMP synthetase of *Escherichia coli* K12. J Biol Chem 260:8676–8679

Waley SG (1969) Some aspects of the evolution of metabolic pathways. Comp Biochem Physiol 30:1–7

Weber AL, Miller SL (1981) Reasons for the occurrence of the twenty coded protein amino acids. J Mol Evol 17:273–284

Weil C, Bekler G, Reeve J (1987) Structure and organization of the *hisA* gene of the thermophilic archaebacterium *Methanococcus thermolithotrophicus*. J Bacteriol 169:4857–4859

Weinstock K, Strathern JN (1993) Molecular genetics in *Saccharomyces kluyveri*: The *HIS3* homolog and its use as a selectable marker gene in *S. kluyveri* and *Saccharomyces cerevisiae*. Yeast 9:351–361

Weir B (1990) Genetic data analysis. Sinauer Press, Sunderland.

Weng M, Makaroff CA, Zalkin H (1986) Nucleotide sequence of *Escherichia coli pyrG* encoding CTP synthetase. J Biol Chem 261: 5568–5574

White HB (1976) Coenzymes as fossils of an earlier metabolic state. J Mol Evol 7:101–117

White DH, Erickson JC (1980) Catalysis of peptide bond formation by histidyl-histidine in a fluctuating clay environment. J Mol Evol 16:279–290

Winkler ME (1987) Biosynthesis of histidine. In: Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaechter M, Humbarger HE (eds) *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, vol 1 American Society for Microbiology, Washington, DC, pp 395–411

Ycas M (1974) On the earlier states of the biochemical system. J Theor Biol 44:145–160

Zalkin H (1985) Glutamine amidotransferases. Methods Enzymol 113: 263–264

Zillig W (1991) Comparative biochemistry of Archaea and Bacteria. Curr Opin Genet Dev 1:544–551

Zillig W, Palm P, Reiter WD, Gropp F, Pulher G, Klenk HP (1988) Comparative evaluation of gene expression in archaebacteria. Eur J Biochem 173:473–482

MICROBIOLOGICAL REVIEWS

A publication of the American Society for Microbiology

Editor in Chief
CATHERINE L. SQUIRES, PhD, Chair
Department of Molecular Biology & Microbiology
Tufts University School of Medicine
136 Harrison Avenue
Boston, MA 02111
Phone: (617)636-6047
Fax: (617)636-0337
Email: csquires_sib@opal.tufts.edu

Journals Division
American Society for Microbiology
Journals Division
1325 Massachusetts Avenue NW
Washington, DC 20005-4171
Phone: (202)737-3600

October 30, 1995

Dr. Carmelo B. Bruni
Dipartimento di Biologia e
Patologia-Cellulare e Moleculare
Via S. Pansini 5
80131 Napoli, ITALY

Dear Dr. Bruni:

I apologize for the long review time, but enclosed please find your interesting manuscript, "Histidine Biosynthetic Pathway and Genes: Structure, Regulation, and Evolution." It has finally been reviewed by two experts in the field and both are very enthusiastic about the review. The delay was caused when one reviewer did not return the manuscript in a timely fashion, and a third reviewer then had to be found and given the materials. I am certainly sorry for the delay.

I have enclosed the comments of one reviewer, the other gave no specific comments other than to say it is an excellent review. I have also quickly read the manuscript and found it to be exceedingly interesting. You and your colleagues have produced a valuable resource for the scientific community and you are to be congratulated on doing such a fine job.

The three comments by reviewer one should be easy to incorporate into a revised version, and I will accept the manuscript for publication as soon as I received the revised version. It should appear in the March, 1996 issue of Microbiological Reviews.

Again, thank you for your hard work. We are very pleased to have this excellent manuscript for publication in Microbiological Reviews and I look forward to soon receiving the revised version.

Sincerely,

Catherine L. Squires

Catherine L. Squires
Editor-in-Chief

Histidine Biosynthetic Pathway and Genes: Structure, Regulation and Evolution.

PIETRO ALIFANO,[1] RENATO FANI,[2] PIETRO LIO',[2] ANTONIO LAZCANO,[3] MARCO BAZZICALUPO,[2] M. STELLA CARLOMAGNO[1] AND CARMELO B. BRUNI[1]*

Centro di Endocrinologia ed Oncologia Sperimentale del Consiglio Nazionale delle Ricerche, Dipartimento di Biologia e Patologia Cellulare e Molecolare "L. Califano", Università degli Studi di Napoli, Via S. Pansini 5, I-80131 Napoli, Italy,[1] and Dipartimento di Biologia Animale e Genetica, Università degli Studi di Firenze, Via Romana 17, I-50125 Firenze, Italy,[2] and Departamento de Biologia, Faculdad de Ciencias, UNAM, México 04510, D.F., México[3]

Correspondent Footnote

Dipartimento di Biologia e Patologia Cellulare e Molecolare

Via S. Pansini 5, 80131 Napoli, Italy

Phone: + 39 - 81 - 7462047

Fax: + 39 - 81 - 7703285

e-mail Bruni@MVX36.CSATA.IT

TABLE OF CONTENTS

INTRODUCTION

The study of the biosynthetic pathway leading to synthesis of the amino acid histidine in prokaryotes and lower eukaryotes was begun over 40 years ago (Haas et al., 1952) and has resulted in the unraveling of many fundamental mechanisms of biology. Together with a few other systems, it can be considered a cornerstone in the foundation and evolving concepts of modern cell biology. For some of us who have been involved with its beauties and intricacies for over twenty years it is particularly important to remember just a few of the accomplishments that have been obtained and of the scientists who tackled those problems.

The histidine system was of the utmost importance in the definition and refinement of the operon theory. A genetic and biochemical analysis of thousands of mutants in the *his* operon of *Salmonella typhimurium* was performed in the late fifties and early sixties in the laboratories of Bruce Ames and Phil Hartman (Hartman, 1956; Hartman et al., 1960a; Hartman et al., 1960b; Ames et al., 1960; Ames et al., 1961). These studies showed that, at variance with the yeast systems (Haas et al., 1952; Leupold, 1958), the bacterial *his* genes were tightly clustered. Demonstration of coordinate expression of this cluster led to the suggestion that this group of genes might function as a single unit of expression and regulation (Ames and Garry, 1959). After the formal enunciation of the operon concept (Jacob and Monod, 1961a), Ames, Hartman and Jacob analyzed 5'-proximal deletions of the regulatory region resulting in a completely non-functional operon and revertants in which expression of the individual genes was restored, to obtain additional evidences of the operon structure (Ames et al., 1963).

Biochemical studies on the *his* mRNA species synthesized in bacteria were performed by Robert Martin in 1963 (Martin, 1963a). Double labeling experiments of constitutive and deletion mutants, RNA chromatographic

fractionation and sucrose gradient centrifugation analysis showed that *his* mRNA is polycistronic and substantiated the one operon-one messenger theory of transcription.

Together with *lac* (Newton et al., 1965) and *trp* (Imamoto et al., 1966), the *his* operon was used as a model system (Martin et al., 1966a; Fink and Martin, 1967) to study the phenomenon of polarity (Ames and Hartman, 1963; Franklin and Luria, 1961; Jacob and Monod, 1961b). The often hot debate on the mechanisms, translational or transcriptional (Martin et al., 1966b; Morse and Yanofsky, 1969), governing polarity and its implications in general operon function (Zipser, 1969) run into the seventies (Imamoto, 1970; Imamoto and Kano, 1971; Morse and Guertin, 1971; Morse and Primakoff, 1970). Polar mutants in the *his* operon are still used in the present days to study fundamental aspects of transcription (Alifano et al., 1988; 1991; Ciampi et al., 1989; Ciampi and Roth, 1988).

Another area in which studies of *his* operon expression were of fundamental help was the study of regulatory mutants and of the mechanisms governing operon expression. Work performed mostly in the laboratories of Ames and Hartman by John Roth identified the different classes of regulatory mutants and showed that, aside from the operator ones, all caused direct or indirect impairment of the histidyl-tRNA^{His} molecule (Anton, 1968; Fink et al., 1967; Roth and Ames, 1966; Roth et al., 1966; Silbert et al., 1966). These findings, in turn, were the basis, together with early studies on the *trp* operon (Jackson and Yanofsky, 1973), for the identification and elucidation of a novel regulatory mechanism of gene expression, namely attenuation (Blasi and Bruni, 1981; Yanofsky, 1981). This term was proposed by Takashi Kasai, then in Phil Hartman laboratory in 1974 (Kasai, 1974). By performing in vivo and in vitro transcription studies with *his* transducing phages and by measuring *his*-specific RNA in wild-type and operator constitutive mutants, Kasai identified a transcriptional barrier (the attenuator) the deletion of which in the mutants was responsible for efficient synthesis of downstream mRNA molecules. Although a positive factor was believed to be required in the process it was also clearly stated by Kasai that the DNA sequence of the attenuator could by itself be responsible for the constitutive expression.

In addition to attenuation the synthesis of histidine in the cells is also regulated by feedback inhibition (Umbarger, 1956). Studies of the mechanisms by which the first enzyme in the pathway was inhibited by the end product histidine and by some analogs provided important insights in this field of enzymology and regulation of biochemical reactions (Ames et al., 1961; Martin, 1963b; Sheppard, 1964).

These are just some examples of the importance of the histidine system in the evolution of the modern concepts of biology. Many more can be found in the classic and lovely book "Gene action" written by Hartman and Suskind and published in 1965 in the Prentice-Hall "Foundations of Modern Genetics" series (Hartman and Suskind, 1965).

Many excellent reviews dealing with several aspects of the histidine pathway have appeared in the course of the years and the readers are referred to them for early aspects (Ames and Hartman, 1963; Ames et al., 1967; Artz and Holzschu, 1983; Brenner and Ames, 1971; Martin et al., 1971). In particular, the last comprehensive review on histidine biosynthesis was published in 1987 by Malcom Winkler (Winkler, 1987) and a revised version of that review will appear in the second edition of the "*Escherichia coli* and *Salmonella typhimurium*. Cellular and molecular biology " book by the American Society for Microbiology currently in press (M. Winkler, personal communication). In the last ten years many studies on the histidine biosynthetic pathway have appeared dealing with several aspects such as gene structure and regulation, transcription initiation and termination, RNA processing, enzymology. In addition, the system has been extensively investigated not only in enterobacteria but also in many other species (Gram-positive and Gram-negative bacteria, archaebacteria and

eukaryotic organisms, affording a unique opportunity to study the evolution of this fundamental pathway. We will try in this review to summarize and describe all these findings in a broader context, as well as to indicate future perspectives and still unanswered questions.

THE HISTIDINE BIOSYNTHETIC PATHWAY

The biosynthesis of histidine has been studied extensively in *S. typhimurium* and *E. coli*. The pathway is described in great detail in the excellent review by Brenner and Ames (1971) with particular emphasis on the physiological implications. The pathway is also accurately presented in the review by Marion Winkler (1987). Very recently important studies have appeared which require to partially modify previous beliefs. In the original studies the pathway was believed to be composed of eleven enzymatic reactions since two of the nine genes (*hisD* and *hisB*) encoded bifunctional proteins (Ames and Hartman, 1963; Hartman and Suskind, 1965). In later reviews for unexplained reasons the dehydrogenase encoded by the *hisD* gene was no longer considered bifunctional and the steps became ten (Brenner and Ames, 1971). The demonstration that *hisI* and *hisE* are in fact a single gene (now *hisI*) (Chiariotti et al., 1986) brought the genes to eight and the steps to ten (Winkler, 1987). Since then the bifunctional nature of the *hisD* gene product has been reaffirmed (Bruni et al., 1986) but at the same time it has been discovered that the *hisH* and *hisF* gene products form an heterodimer and both catalyze the same step. The postulated unknown intermediate (Brenner and Ames, 1971) in fact does not exist (Klem and Davisson, 1993; Rieder and Kleiner, 1993; Rieder et al., 1994). In conclusion, three (*hisD*, *hisB* and *hisI*) of the eight genes of the operon encode bifunctional enzymes and two (*hisH* and *hisF*) encode polypeptide chains which form an enzyme molecule catalyzing a single step for a total of ten enzymatic reactions (Fig. 1).

The Reactions and Enzymes in Histidine Biosynthesis

The first reaction in histidine biosynthesis (Fig. 1) is the condensation of ATP and 5-phosphoribosyl 1-pyrophosphate (PRPP) to form N-1-(5'-phosphoribosyl)-ATP (PR-ATP). This enzymatic reaction has been studied in detail by Martin (1963b) and is the one involved in feedback inhibition. It is catalyzed by the N-1-(5'-phosphoribosyl)-ATP transferase, the product of the *hisG* gene. Most informations about structure and regulation of the activity of the transferase come from the *S. typhimurium* and *E. coli* homologous enzymes. In both microorganisms the purified enzyme is a hexamer composed of identical subunits of 34 kDa (Klungsøyr and Kryvi, 1971; Parsons and Koshland, 1974a; Voll et al., 1967; Whitfield, 1971). Multiple aggregation states have been evidenced under different assay conditions. There is an equilibrium between various oligomers, such as dimers, tetramers, hexamers and high aggregates; dimer is the basic oligomeric unit (Klungsøyr, 1971; Parsons and Koshland, 1974b). Dimer is the most active species of the enzyme isolated from *E. coli* (Dall-Larsen 1988a; 1988b). The equilibrium between the aggregation states is shifted toward the hexameric form by histidine alone, AMP alone and by combination of these ligands with synergistic effects (Klungsøyr and Kryvi, 1971; Parsons and Koshland, 1974b; Dall-Larsen and Klungsøyr, 1976). However, the most powerful ligand for stabilizing the hexameric form is the product PR-ATP. Using the purified transferase from *E. coli*, Tebar et al. (1973; 1975) demonstrated that one of the two substrates of the transferase, the PRPP, brings about a dissociation of hexamers and higher aggregates with a resulting increase in the concentration of dimers. On the other hand, ATP counteracts PRPP in this respect. This finding is in apparent conflict with data indicating that the hexameric form of the homologous enzyme of *S. typhimurium* is stabilized by the substrates (ATP and PRPP) (Bell et al., 1974). The different behaviors observed

for the two transferases may simply reflect differences in the experimental conditions.

The aggregation state of the transferase of *E. coli* has been, at least in part, related to regulation of its activity: ligands that stabilize the hexameric form of the transferase also play an inhibitory role on its activity (Klungsøyr and Kryvi, 1971; Dall-Larsen, 1988a; 1988b). The feedback control of the transferase by histidine was documented since 1961 (Ames et al., 1961). Klungsøyr et al. (1968) found that the transferase from *E. coli* is per se insensitive to histidine inhibition; the histidine effect becomes apparent in the presence of the product of the reaction PR-ATP and it is further increased by the AMP. The synergistic inhibition by the product of the reaction and the end product of the pathway represents a sophistication of the general principle of feedback control and it has been documented by other few examples (Hubbard and Stadtman, 1967). The inhibitory effect of the AMP supports the energy charge theory proposed by Atkinson (Klungsøyr et al., 1968) and seems logical if we consider the high metabolic cost required for the histidine biosynthesis.

The product of the reaction of the transferase, PR-ATP, is hydrolyzed to N-1-(5'-phosphoribosyl)-AMP (PR-AMP). This irreversible hydrolysis is catalyzed by one of the two activities, corresponding to the carboxyl-terminal domain of the enzyme coded for by the *hisI* gene (Smith and Ames, 1965), formerly known as *hisIE* (Fig. 1). The other activity which is localized in the amino-terminal domain of the bifunctional enzyme (Chiariotti et al., 1986; Carlomagno et al., 1988) is a cyclohydrolase which opens the purine ring of PR-AMP, leading to the production of an imidazole intermediate, the N'-[(5'-phosphoribosyl)-formimino]-5 aminoimidazole-4 carboxamide-ribonucleotide (abbreviated 5'-ProFAR or BBMII) (Smith and Ames, 1965).

The fourth step of the pathway is an internal redox reaction, also known as "Amadori rearrangement", involving the isomerization of the aminoaldose 5'-ProFAR (or BBMII) to the aminoketose N'-[(5'-phosphoribulosyl)-formimino]-5 aminoimidazole-4 carboxamide-ribonucleotide (5'-PRFAR or BBMIII). The reaction is catalyzed by the *hisA* gene product (Smith and Ames, 1964; Margolies and Goldberger, 1966; 1967).

Although the pathway of histidine biosynthesis was almost completely characterized since 1965 an ambiguity remained for a long time concerning the biochemical event leading to the synthesis of imidazole-glycerol-phosphate (IGP) and 5'-phosphoribosyl- 4 carboxamide-5 aminoimidazole (PRAIC, AICAR, or ZMP) from the 5'-PRFAR (or BBMIII). The functions of the *hisH* and *hisF* genes were known to be involved in the overall process in eubacteria (Smith and Ames, 1964), but the catalytic properties of each protein had not been completely characterized. Smith and Ames (1964) demonstrated that this process required the presence of glutamine as a source of amide nitrogen; however high concentration of ammonia could be substituted for glutamine and the *hisH* enzyme at alkaline pH values in an in vitro system. These authors suggested that the *hisH* gene product added the amide nitrogen of glutamine to that portion of 5'-PRFAR (or BBMIII) which cyclized to form IGP at the intracellular pH values. The *hisF* gene product (the "cyclase") cleaved the side chain of 5'-PRFAR (or BBMIII); the cyclization of the moiety cleaved away was supposed to occur spontaneously. It was hypothesized that the *hisF* and *hisH* gene products controlled two separate steps whose order in the pathway was difficult to establish. As a consequence, the structure of the intermediate could not be predicted (Smith and Ames, 1964; Martin et al., 1971). The last blind spot of the histidine biosynthesis has been recently clarified by analyzing the catalytic properties of the *hisH* and *hisF* gene products from *E. coli* (Klem and Davisson, 1993; Rieder and Kleiner, 1993; Rieder et al., 1994). Klem and Davisson (1993) found that the protein encoded by the *hisF* gene has an ammonia-dependent activity that is responsible for the conversion of PRFAR (or BBMIII) to PRAIC (AICAR or ZMP) and IGP, while the product of the *hisH* gene had no detectable catalytic properties. However, in combination, the two proteins were able to

carry out the reaction in the presence of glutamine as a nitrogen donor without releasing any free metabolic intermediate. The *hisH* and *hisF* gene products formed a stable 1:1 dimeric complex that constituted the IGP synthase holoenzyme. The existence of this functional dimeric complex has been proved genetically by isolation and characterization of a mutated *hisF* gene product in *Klebsiella pneumoniae* which catalyzes the reaction utilizing free ammonia but not the ammonia moiety from glutamine bound to *hisH* gene product. The mutation, which results in the replacement of aspartic acid by asparagine, has been speculated to affect the interaction between the *hisH* and *hisF* gene products (Rieder et al., 1994). A striking feature of the protein coded for by the *hisH* gene is that, despite the high degree of active site sequence homology with several amidotransferases which also exhibit glutaminase activity in the absence of their respective substrates (Goto et al., 1976; Trotta et al., 1974), its glutamine-dependent catalytic properties require the presence of the *hisF* gene product (Klem and Davisson, 1993). Another heterodimeric glutamine amidotransferase, the enzyme aminodeoxychorismate synthase which is composed of the proteins encoded by the *pabA* and *pabB* genes, has similar features (Green and Nichols, 1991; Ye et al., 1990).

The PRAIC (AICAR or ZMP) which is produced in the reaction catalyzed by the IGP synthase is recycled into de novo purine biosynthetic pathway (see below). The other product, the IGP, is dehydrated by one of the activities corresponding to the carboxyl-terminal domain of the bifunctional enzyme encoded by the *hisB* gene (Brenner and Ames, 1971). The resulting enol is ketonized non-enzymatically to imidazole-acetol-phosphate (IAP).

The seventh step of the pathway consists of a reversible transamination involving the IAP and a nitrogen atom from glutamate. The reaction leads to the production of α-ketoglutarate and L-histidinol-phosphate (HOL-P), and is catalyzed by a pyridoxal-P-dependent aminotransferase encoded by the *hisC* gene (Brenner and Ames, 1971). This enzyme shares certain mechanistic

features with other pyridoxal-P-dependent aminotransferases: i) covalent binding of pyridoxal-P to an active site lysine residue; ii) the formation of an aldimine between pyridoxal-P and the amino acid substrate as the first intermediate; a Ping-Pong Bi-Bi mechanism of catalysis (Hsu et al., 1989; Mehta et al., 1989).

The HOL-P is converted to L-histidinol (HOL) by the phosphatase activity localized in the amino-terminal domain of the bifunctional enzyme encoded by the *hisB* gene (Brenner and Ames, 1971).

During the last two steps of the histidine biosynthesis HOL is oxidized to the corresponding amino acid L-histidine (HIS) (Adams, 1954). This irreversible four-electron oxidation proceeds via the unstable amino aldehyde L-histidinal (HAL) which is not found as a free intermediate (Adams, 1954; Görisch and Hölke, 1985). A single enzyme, the L-histidinol dehydrogenase encoded by the *hisD* gene, catalyzes both oxidation steps probably to prevent decomposition of the unstable intermediate (Loper and Adams, 1965). This enzyme represents one of the first established examples of bifunctional NAD^+-linked dehydrogenase (Bürger and Görisch, 1981a; Kirschner and Bisswanger, 1976). Most information about the L-histidinol dehydrogenase comes from the enzyme purified from *S. typhimurium*. It is an homodimeric Zn^{2+} metalloenzyme which functions by carrying out the first oxidation step at an active site on one subunit and then moving the intermediate to a vicinal site on an adjacent subunit (Bürger and Görisch, 1981a; Bürger et al., 1979; Eccleston et al., 1979; Görisch and Hölke, 1985; Grubmeyer et al., 1989). Steady-state kinetic patterns demonstrated that the enzyme acts via a Bi-Uni Uni-Bi Ping-Pong mechanism: HOL binds first to the enzyme, followed by the binding of NAD^+; HIS is the last product to dissociate (Bürger and Görisch, 1981a; Görisch, 1979).

A comparative analysis of the *E. coli* and *S. typhimurium hisD* gene products and the homologous region (*HIS4*C) of the multifunctional product of the *Saccharomyces cerevisiae HIS4* gene (Donahue et al., 1982) showed the

presence of two long regions characterized by highly conserved amino acid sequences (Bruni et al., 1986). These regions were supposed to represent functional domains of the enzyme: the amino-terminal region responsible for the first oxidation step, and the carboxyl-terminal region responsible for the second one. Genetic data on the existence of two groups of mutations in *hisD* which exhibit intracistronic complementation further supported this belief (Hartman et al., 1971).

An essential lysine residue appears to participate in the reversible oxidation/reduction converting the alcohol HOL to the aldehyde HAL during the first step of the reaction (Bürger and Görisch, 1981b). Based on the well elucidated mechanism of catalysis of the glyceraldehyde-3-phosphate dehydrogenase, in which an active site cysteine adds to form a thiohemiacetal intermediate (Harris and Waters, 1976), it was assumed that the thiohemiacetal/thiolester pair represented the route for the aldehyde oxidation in the second step catalyzed by the histidinol-dehydrogenase. This hypothesis was initially supported by the evidences that the *S. typhimurium* enzyme contains two conserved cysteine residues, Cys-116 and Cys-153, and is inactivated by active site modification of Cys-116 by the reagent 4-nitro-7-chlorobenzadioxazole (Grubmeyer and G. y, 1986). However, the recent result that mutant enzymes with either alanine or serine substitution of Cys-116 and Cys-153 are active with kinetic properties resembling the wild-type enzyme has ruled out this hypothesis thus suggesting that the reaction might proceed through a scheme different from those common to most aldehyde dehydrogenases (Teng et al., 1993).

Metabolic Links between the Histidine and the Purine Biosynthetic Pathways: the "AICAR Cycle".

Mutants bearing non-functional enzymatic activities, which are required for histidine biosynthesis, grow normally in minimal medium when supplied with exogenous histidine. On the basis of this evidence, the pathway was supposed to lack any branch point leading to other metabolites required for growth (Ames et al., 1967; Brenner and Ames, 1971). Nevertheless, the two initial substrates of histidine biosynthesis, PRPP and ATP, play a key role in the intermediate and energetic metabolism, and link this pathway to the biosynthesis of purines, pyrimidines, pyridine nucleotides, folates and tryptophan (Brown and Williamson, 1987; Neuhard and Nygaard, 1987; Pittard, 1987). These metabolic links would, at least in part, account for the pleiotropic effects generated by the derepressed synthesis of the enzymes coded for by the histidine operon.

It has been calculated that 41 ATP molecules are sacrificed for each histidine molecule made (Brenner and Ames, 1971). The considerable metabolic cost justifies the finding that histidine regulatory mutants which have constitutive expression of the histidine operon and lack feedback control of the biosynthetic pathway require adenine for growth at 42°C (Johnston and Roth, 1979; Shioi et al., 1982; Stougaard and Kennedy, 1988).

The purine and histidine biosynthetic pathways are connected through the "AICAR cycle" (Fig. 1). The AICAR (PRAIC or ZMP), a by-product of the histidine biosynthesis, is also a purine precursor being converted to IMP. This conversion involves a folic acid mediated one-carbon (C-1) transfer (Neuhard and Nygaard, 1987). Bochner and Ames (1982) reported that the unusual nucleotide 5 aminoimidazole-4 carboxamide-riboside-5'-triphosphate (ZTP) accumulated in *S. typhimurium* cells following treatment thought to lower the folic acid pool and that strains unable to make Z-ribotides were hypersensitive to antifolate drugs. Under the same treatment ZTP production was also detected in *E. coli* cells even though the accumulation of the unusual nucleotide did not correlate strictly with folate deficiency (Rohlman and Matthews, 1990). Bochner and Ames (1982)

proposed that ZTP is an alarmone signaling C-1-folate deficiency and mediating a physiologically beneficial response to folate stress. This ~~belief~~ is supported by several findings concerning the ZTP synthesis. At variance with other triphosphate ribotides whose synthesis involves a two-step process controlled by specific monophosphate kinases and a non-specific diphosphate kinase, ZTP is made by pyrophosphate transfer onto ZTP in a single enzymatic reaction catalyzed by PRPP synthetase (Bochner and Ames, 1982; Sabina et al., 1984; 1985). The specificity of this conversion is also supported by results from kinetic studies indicating that the increase in ZTP pool does not track the increase in ZMP pool (Bochner and Ames, 1982).

Inhibition of folate metabolism leads to alterations of intracellular processes which may involve ZTP-mediated responses. Interesting effects on gene expression are observed in folate-deficient cells. For example, the availability of 10-formyl-tetrahydrofolate influences the mode of derepression - sequential or simultaneous - of the genes clustered in the *his* operon of *S. typhimurium* possibly by affecting the mechanism of translation coupling at the intercistronic barriers (Berberich et al., 1966; Brenner and Ames, 1971; Petersen et al., 1976a; 1976b). Addition of inhibitors of folate metabolism induces polarity in *E. coli* and *S. typhimurium* (Alifano et al., 1994; Petersen et al., 1978) and affects the rate of decay (Joseph et al., 1978) or processing (Alifano et al., 1994) of several polycistronic mRNAs. Other bacteria, such as *Bacillus subtilis*, respond to folate shortage by initiating sporulation (Freese et al., 1979; Heinze et al., 1978; Milani et al., 1977). However, the involvement of ZTP in these processes is only speculative and the evidence for a folate stress regulon controlled by ZTP remains elusive up to date.

It has been known since a long time that the constitutive expression of the *his* operon of *S. typhimurium* results in a number of phenotypic changes (i.) growth inhibition at 42°C somehow relieved by methionine (Fink et al., 1967) (ii.) wrinkled morphology of colonies grown in either 2% glucose or "green plates" (Gibert and Casadesús, 1990; Murray and Hartman, 1972; Roth and Hartman, 1965) (iii.) growth inhibition in high-salt media (Casadesús and Roth, 1989). A similar pleiotropic response is also observed in *E. coli* (Frandsen and D'Ari, 1993). The wrinkled morphology of the *his*-constitutive strains is due to filament formation as a consequence of cell division inhibition (Gibert and Casadesús, 1990; Murray and Hartman, 1972). Murray and Hartman (1972) demonstrated that the pleiotropic response is caused by the overproduction of the *hisH* and *hisF* gene products acting in a concerted fashion. Now we know that these two proteins associate to form the heterodimeric IGP synthase complex which catalyzes the closure of the imidazole ring of histidine, thereby releasing the by-product ~~AICAR~~ (PRAIC or ZMP). Since AICAR is a potential precursor of the unorthodox nucleotide ZTP, a proposed alarmone, it ~~was~~ hypothesized that the pleiotropic response might be caused by the enzymatic activity of IGP synthase leading to AICAR accumulation.

~~As~~ the base 5-amino-4-imidazole carboxamide is a mutagen in *E. coli* (Murray, 1987), mutagenic properties of the corresponding riboside AICAR or its derivatives ~~were~~ also invoked to ~~account for~~ the wrinkled phenotype by analogy with SOS-induced filamentation upon perturbation of DNA synthesis (Donachie and Robinson, 1987). Isolation of an antimutator strain of *E. coli* carrying a $purB_{ts}$ mutation (Geiger and Speyer, 1977) was in agreement with this proposal, since *purB* catalyzes the synthesis of AICAR in the purine biosynthetic pathway (Neuhard and Nygaard, 1987).

However, more recent evidences ~~have~~ ruled out this hypothesis. First, it has been reported that AICAR is not an endogenous mutagen in *E. coli* (Fox et al., 1993). Second, overexpression of the *his* operon in *S. typhimurium* cells does not result in increased incidence of spontaneous mutations as a consequence of AICAR accumulation thereby confirming that the pleiotropic response does not involve DNA damage (Flores et al., 1993). This conclusion agrees with the observations that overexpression of the *his* operon does not induce SOS response

(Gibert and Casadesús, 1990) and that filamentation in *his*-constitutive strains is independent of the two SOS-associated division inhibitors, SfiA and SfiC (Flores et al., 1993). Moreover, the filaments observed in *his*-constitutive strains, unlike the aseptate filaments formed after SOS induction, present smooth partial constrictions in DNA-free regions (Frandsen and D'Ari, 1993) and are reminiscent of those observed in $ftsI_{ts}$ mutants bearing a temperature sensitive penicillin-binding-protein 3 (PBP 3) (Taschner et al., 1988). Third, elevated levels of IGP synthase cause inhibition of cell division by themselves and not via AICAR production. Filamentation, as well as the other pleiotropic effects associated with *his* overexpression, were shown to occur even in *E. coli* and *S. typhimurium* strains devoid of AICAR as a consequence of interrupting the carbon flow through the histidine and purine pathways in *his pur* double mutants (Flores et al., 1993; Frandsen and D'Ari, 1993).

Based on intergenic suppression analysis, it has been suggested that elevated levels of *hisH* and *hisF* gene products induce filamentation by interfering somehow with synthesis of cell wall (Anton, 1978; 1979). Several of these suppressors result in spherical cells with increased autolysis and sensitivity to penicillin or related antibiotics. One such mutant was affected at the *envB* locus, also known as *mre* in *E. coli* (Anton, 1979). Interestingly, an *E. coli mreB* mutant, which also has spherical shape and is hypersensitive to mecillinam, overproduces PBP 3 (Wachi et al., 1987). Altogether, these findings suggest that division inhibition in *his*-constitutive strains may result from a shortage of septal murein synthesis catalyzed by PBP 3 (Frandsen and D'Ari, 1993). The classical approach of studying intergenic suppression has recently allowed to identify novel loci on the *S. typhimurium* chromosome. These suppressors behave either as "general" or "partial" suppressors of the pleiotropic response; epistatic effects among suppressors have been also documented (Flores and Casadesús, personal communication).

It is conceivable that further study of the pleiotropic response triggered by the overexpression of *hisH* and *hisF* gene products will facilitate the discovery of genes controlling the metabolic pathway(s) leading to cell division in *E. coli* and *S. typhimurium*.

ORGANIZATION OF THE HISTIDINE GENES

Since the beginning of the microbial genetic studies histidine requiring mutants were frequently isolated and characterized. These works led in *E. coli* and *S. typhimurium* to the identification of all the genes necessary for histidine biosynthesis and to their localization on the genetic map (Winkler, 1987). Similar work on the yeast *S. cerevisiae* also allowed the identification and the genetic mapping of several genes involved in histidine biosynthesis (Mortimer, 1994). After the introduction of DNA sequencing techniques and genetic engineering in microbial genetics, the data on *his* genes accumulated rapidly in model microorganisms and in several other bacterial and fungal species providing a wealth of information regarding at least 14 bacterial (including Archaea), 5 fungi and 3 plant species with more than 60 genes sequenced. Table 1 reports the complete list of all these genes grouped according to their homology with the well characterized genes of the *his* operon of *E. coli*. Many of these genes/operons were first identified by the ability of cloned DNA fragments to complement histidine auxotrophic mutations either in *E. coli* or in homologous hosts. Final identification was generally achieved from DNA/protein sequence comparison with *E. coli* counterpart, assuming, as it is widely accepted, that the biosynthetic pathway is fundamentally the same in all the organisms.

As indicated by the frequent occurrence of genetic complementation, homologous genes (that are presumed to code for proteins performing the same function in the biosynthetic pathway) in different species are generally similar to

each other. The similarity from overall length of the gene extends to the molecular weight and secondary structure of the protein they code for.

In many of the species where *his* genes were identified and characterized they were not found alone but clustered with other genes to constitute complete operons or at least part of them. The same is in part true for operonless fungi, where some of the *his* genes resulted from the fusion of different parts, each of which is homologous to different bacterial genes. The organization of genes in the *his* operons or clusters is variable among the different species indicating that, during the evolution, genes were separated or linked apparently without severe constrains (see below). In other bacterial operons, which have been characterized in several species, such as the tryptophan operon, gene order was mostly found to be invariant (Crawford and Milkman, 1991). The organization of the known *his* gene clusters in microorganisms is presented in Fig. 2.

Eubacteria

Gram negative. In the Enterobacteriaceae *E. coli* and *S. typhimurium* a single operon composed of eight genes very tightly linked to each other encodes all the enzymes required for the biosynthesis of histidine. The complete genetic structure of the *E. coli* and *S. typhimurium* operons was reported by Carlomagno et al. (1988), with minor differences found in genes *hisG* and *hisD* by Jovanovic et al. (1994). The operons measure 7389 and 7438 bp in *E. coli* and *S. typhimurium*, respectively with an overall homology of 81%. The order of the genes in the operon does not match the sequence by which the enzymes they code for take part in the synthesis of histidine; it is possible that the particular gene order resulted from both regulatory and metabolic constrains (see below). In the other Enterobacteriacea *K. pneumoniae* (Rieder et al., 1994; Rodriguez et al., 1981), the DNA sequence available, consisting of part of the *hisG*, the complete *hisF* and a very little part of *hisI*, suggests a general organization similar to that of *E. coli* and *S. typhimurium*.

In the Gram negative proteobacterium *Azospirillum brasilense*, belonging to the α-purple subdivision, an operon structure was found with a transcription initiation, seven ORFs and a transcription terminator (Fani et al., 1989, 1993). Only five of these ORFs, however, are homologous to *E. coli* genes, the remaining two code for proteins without any homology to other *his* genes. Assuming that the histidine biosynthetic pathway of *A. brasilense* is identical to that of *E. coli*, there are at least three genes missing in the operon: *hisGDC*. A fourth missing gene could be the sequence coding for the PR-ATP pyrophosphohydrolase/ PR-AMP cyclohydrolase, *hisI* of *E. coli*, whose corresponding ORF in *A. brasilense* (*hisE*) is just a part of it. The same deduction can be applied to *hisB*px moiety of the *E. coli hisB* (see below) that is also missing in the *A. brasilense* operon. The presence of other *his* genes in *A. brasilense* genome, unlinked to the main operon, is further suggested by the inability of the genes of the identified cluster to complement some uncharacterized histidine requiring mutants of *A. brasilense* (Fani et al., unpublished a). The order of genes in the *A. brasilense his* operon, *hisBdHAFE*, is the same as in *E. coli* . The same gene order does not mean however the same gene organization: in fact one of the two unidentified ORFs lays between *hisH* and *hisA*, another unidentified ORF at the end of the operon separates *hisE* from the terminator, and finally *hisBd* and *hisE*, as already mentioned, correspond to just a portion of the *E. coli* genes (see below).

Gram positive. A single operon organization was also found in the Gram positive bacterium *Lactococcus lactis* where 12 ORFs have been identified (Delorme et al., 1992). Eight ORFs are homologous to the *E. coli his* genes, indicating that all the enzymes of the pathway are coded for by this operon. The exception is the product of *hisB* gene that, like in other organisms (see below), is only present with the dehydratase activity coded by the *hisBd* portion of the gene, while *hisB*px, the phosphatase domain, is apparently absent. Four ORFs

code for unidentified proteins with unknown function. The homology with *E. coli* genes, however, is not followed by similar gene order that is in fact maintained only in the 3' portion of the operon for the *hisHAFI* genes. The 5' portion of the operon apparently underwent a translocation with respect to the *E. coli* arrangement, moving *hisC* from immediately downstream *hisD* to upstream *hisG*, at the very beginning of the operon. The general organization of the *his* operon of *L. lactis* reflects the features of other genes for amino acid biosynthesis in lactic acid bacteria, such as the single chromosomal location and the presence of unrelated ORFs (Chopin, 1993).

Clustered genes for the histidine biosynthesis were also found in the Gram positive *Streptomyces coelicolor* in which the *his* genes were mapped in three unlinked loci (Hopwood et al., 1993). Two minor loci, each containing one gene, are located at 2 o' clock and 6 o' clock on the strain A3(2) chromosome (the latter identified as *hisB*px). The third locus, mapping at 12 o' clock, contains the main *his* cluster with seven ORFs five of which are homologous to the *E. coli hisDCBdHA* (Limauro et al., 1990, 1992). Also in *S. coelicolor* the general gene order is the same as in *E. coli*, with the insertion of unknown ORFs changing the organization of the operon. In this case however the sequence of the last ORF at the 3' end of the cluster was not completely determined and it is possible that other ORFs will be included in the same operon. It is interesting to note that three out of the seven ORFs sequenced start with unusual codons, *hisC* and *hisD* with GUG, and *hisH* with UUG. However, at last GUG was found to be rather frequent as start codon in *Streptomyces* (Hopwood, 1986).

In two other Gram positive bacteria, *Mycobacterium smegmatis* and *B. subtilis*, only limited sequences of *his* genes are available, however it is possible to infer that *M. smegmatis his* genes share the same organization of the more closely related *S. coelicolor* (Hinshelwood and Stoker, 1992) with an operon structure starting with *hisD* followed by *hisC*. *B. subtilis hisH* gene, that is homologous to *E. coli hisC*, seems, on the contrary, to be an isolated gene (Henner et al., 1986). Another *his* locus, containing all other *his* genes, was mapped in the *B. subtilis* chromosome but none of these genes was sequenced up to now (Anagnostopulos et al., 1993). A genetic organization similar to that of *B. subtilis* was postulated for the *his* genes of the related Gram positive *Staphylococcus aureus* (Pattee, 1993).

Archaebacteria

his genes in archaebacteria are less known than in eubacteria; only three *his* genes have been recognized in just four species, three of them belonging to the genus *Methanococcus* and one to *Halobacterium* (Fig. 2). In *M. vannielii* (Beckler and Reeve, 1986), *M. voltae* (Cue et al., 1985) and *M. thermolithotrophicus* (Weil et al., 1987) a gene highly homologous to *E. coli hisA* has been sequenced together with its flanking regions; the analysis of these sequences failed to demonstrate the occurrence of an operon-like structure. The ORFs surrounding *hisA* are in fact apparently unrelated to the histidine biosynthesis, on the other hand no promoter-like sequence was recognizable upstream *hisA* but only a putative ribosome binding site (Weil et al., 1987). A second *his* gene, *hisI* sequenced in *M. vannielii* (Bekler and Reeve, 1986) was found unlinked to *hisA*. If further studies will eventually demonstrate an operon structure for archaebacterial *his* genes, it will result interrupted by apparently unrelated ORFs, like the operons found in *L. lactis* and *A. brasilense*. A conserved genetic organization, if not an operon, among archaebacterial *his* genes, is suggested by the high level of homology found between the peptides coded for by ORF547, ORF294 and ORF114 preceding *hisA* and between ORF150 and ORF125 following the same gene in the three *Methanococci* (Fig. 2). The remaining ORFs do not show any kind of homology to each other. The *H. volcanii hisC* gene is probably a single gene, as the flanking sequences do not indicate the presence of other *his* genes (Conover and Doolittle, 1990).

Eukaryotes

In eukaryotes, the general rule that no operon structures were found also apply to *his* genes. In particular, in the well known yeast *S. cerevisiae*, the seven genes responsible for the biosynthesis of histidine are located on six different chromosomes (Mortimer, 1994). Despite the spreading of the *his* genes, a particular sort of gene organization could be seen on two of the six yeast *his* genes sequenced so far. *HIS4* (Donahue et al., 1982) and *HIS7* (Kuenzler et al., 1993) are in fact very large genes whose sequence is homologous to two pairs of *E. coli* genes: *hisI* and *hisD* for *HIS4* and *hisH* and *hisF* for *HIS7* (Fig. 2 and Table 1). The same multifunctional structure of yeast *HIS4* is also found in *Neurospora crassa his-3* (Legerton and Yanofsky, 1985) and *Pichia pastoris* (unpublished) as reported below. A gene that complements *S. cerevisiae HIS4* deletion was also found in *Candida albicans* but not sequenced up to now (Altbzum et al., 1990). Several other *his* genes have been cloned and sequenced in eukaryotes other than yeast (Table 1), however they are single genes and it is not possible to make inferences about their organization.

Structure of *his* Genes and Their Products

The high level of homology between corresponding genes for histidine biosynthesis in different organisms indicated that this metabolic pathway was fundamentally conserved during evolution (see below). Looking at the structure of *his* genes, however, some interesting peculiarities easily become evident.

hisG. *hisG* genes code for the first enzyme in the biosynthesis of histidine, a protein of about 300 amino acids in *E. coli*, *S. typhimurium* and *S. cerevisiae* but shorter (about 200 amino acids) in *L. lactis*. Because the short *L. lactis* gene is able to complement *E. coli hisG* mutants (Delorme et al., 1992), it is likely that

a considerable portion of the enterobacterial and yeast protein is dispensable for the enzymatic activity.

hisB. Two enzymatic activities have been recognized to be coded by the *hisB* gene of *E. coli* and *S. typhimurium* (see above). In the other microorganisms studied, on the contrary, the two activities appear to be coded for by two independent genes: *hisB*d coding for IGP dehydratase and *hisB*px, coding for HOL-P phosphatase. *hisB*d genes, with different levels of homology to each other and with protein products ranging from 195 to 270 amino acids, were found in almost all the species, including cyanobacteria, fungi and plants (Table 1); on the other hand *hisB*px homologous gene is still to be found in any other microorganism except *S. coelicolor* (Hopwood et al., 1993), where it is called *hisD* and was mapped outside from the main gene cluster but not sequenced yet, and *S. cerevisiae* (Malone et al., 1994) whose *HIS2* was found on chromosome VI and to be very close to a site of a ricombinational hot-spot. Surprisingly we did not find significant degree of similarity with the enterobacterial *hisB* 5' proximal domain. This finding suggests that different genes, possibly not clustered with the other *his* genes, or that (Delorme et al., 1992) some of the uncharacterized ORFs frequently found in the *his* operons, can perform in different organisms the HOL-P dephosphorylation step of histidine biosynthesis.

A different structure of *hisB*d gene was found in the fungus *Phytophtora parasitica* where the sequence coding for IGP dehydratase, recognized by the high level of homology with the corresponding protein of *E. coli*, is fused with a DNA fragment whose translation gives an amino acid sequence that is not homologous to any of the known histidine biosynthetic enzymes (Fani et al., 1995).

hisI and ***hisE***. *E. coli*, *S. typhimurium* and *L. lactis hisI* genes (Carlomagno et al., 1988, Delorme et al., 1992), mapping at the end of their respective operons, also code for a bifunctional enzyme responsible, with its amino terminal domain, for the third step and, with the carboxyl terminal one,

for the second step of histidine pathway. This is one example of the lack of correspondence between the order of genes (and coding sequences) in the *his* operon and the order of the biosynthetic steps performed by the enzymes encoded by these genes.

The same enzymatic steps are achieved in fungi, *S. cerevisiae*, *N. crassa* and *P. parasitica* (Legerton and Yanofsky, 1985, Donahue et al., 1982; unpublished) by a large multifunctional protein whose amino-terminal domain shares good homology with the bifunctional *hisI* gene product. Pyrophosphohydrolase and cyclohydrolase activities being in the same polyfunctional enzyme is not however an universal phenomenon. In the Gram negative *A. brasilense* (Fani et al., 1993) and in the archaebacterium *M. vannielii* (Beckler and Reeve, 1985) the two enzymatic activities seem to be separated in two proteins, coded for by different genes, although in *A. brasilense* only *hisE* (pyrophosphohydrolase) and in *M. vannielii* only *hisI* (cyclohydrolase) were found up to now.

hisD. The product of *hisD* gene is likewise a protein with bifunctional activity, histidinol dehydrogenase, performing the last two steps of the histidine biosynthesis. Contrary to the examples of *hisB* and *hisI*, *hisD* is a distinct unsplitted gene in the nine species in which it has been found and sequenced, including the plant *Brassica oleracea*. In this plant the gene product, which complements the corresponding *E. coli* mutation, is probably located in the plastid as indicated by a 31 amino acid signal peptide typical of the proteins transferred to plastid compartment, absent in the mature enzyme (Nagai et al., 1991). In the fungi *S. cerevisiae*, *N. crassa* and *P. pastoris* (Legerton and Yanofsky, 1985; Donahue et al., 1982; ~~unpublished~~ Gene and Genet.) the two enzymatic activities are carried out by the carboxyl-terminal domain of the same huge protein that also comprehends the homologue of *hisI* gene product. The bifunctional enzymatic activity of *hisD* gene product seems to be an universal property,

suggesting that a specific constraint, related to the mechanism of reaction, favored the persistence of the single protein arrangement.

The four enzymatic activities, pyrophosphohydrolase and cyclohydrolase and HOL-dehydrogenase (the 2 steps), are thus performed in fungi by a large super-enzyme of 800-850 amino acids whose genes appear highly homologous to each other in *S. cerevisiae*, *P. pastoris* and *N. crassa* with the only difference in this latter species of a 59 bp intron in the region coding for the domain corresponding to *hisD* product. This intron makes the *N. crassa* gene unable to complement the *E. coli hisD* mutations.

***hisA-hisF* and *hisH. hisA* and *hisF* genes**, as reported below, share high sequence homology and have probably arisen from a gene duplication event, their gene products participate in two successive steps of the biosynthetic pathway. The two genes are also adjacent in almost all eubacterial operons.

As reported before, the products of *E. coli hisH* and *hisF* genes form in vivo an heterodimeric enzyme, IGP synthetase, that accomplishes the fifth step of histidine synthesis (Klem and Davisson, 1993). Their genes in prokaryotic operons however are always kept apart by at least a third ORF (Fig. 2). In *S. cerevisiae*, on the contrary, there is a single enzyme, coded for by a single gene, *HIS7*, fulfilling IGP synthetase activity (Kuenzler et al., 1993). The amino acid sequence of this protein shows high level of homology, in its amino-terminal domain, with the product of *E. coli hisH* and, in the carboxyl- terminal domain, with the product of *E. coli hisF*. These two domains of *HIS7* are separated by a short sequence of 22 amino acids without homology to *hisH* or *hisF* gene products. The portion of *S. cerevisiae HIS7* sequence corresponding to eubacterial *hisF* also showed a peculiar structure in its extreme amino terminal. In this part of the protein in fact, the homology with *hisF* product is partially lost because of six short non-homologous sequences scattered in the domain rendering this portion of the yeast protein almost one third longer than the bacterial counterpart. If the catalytic activity of HisF resides in the carboxyl-part of the

polypeptide while the interaction with HisH is brought about by the amino-part, that will account for the number of sequence alterations in the yeast homologue, where this interaction is no more necessary since *hisH* and *hisF* are fused in a single gene, *HIS7* (Fani et al., 1995). This hypothesis is strongly supported by the recent finding of Rieder et al. (1994), who suggested that a mutation falling in the F1 module of the *K. pneumoniae hisF* gene probably affected the interaction between HisH and HisF.

ORFs. Fig. 2 reports the structure of DNA regions of different organisms where histidine biosynthetic genes were found more or less grouped together. A remarkable feature of these (bacterial and archaebacterial) clusters is the presence of several ORFs with unknown function: seventeen ORFs have been in fact identified between or flanking *his* genes in *A. brasilense*, *S. coelicolor*, *L. lactis*, *M. vannielii*, *M. voltae* and *M. thermolithotrophicus*. The involvement of these ORFs in the biosynthesis of histidine is an open question as their function was not determined up to now and mutants for studying the phenotype are not generally available. The only connection of these ORFs with histidine biosynthesis is their linkage with *his* genes. On the other hand information derived from sequence analysis is not of much help. In *L. lactis* four unknown ORFs, scattered in the operon, were detected. One of them, ORF3, is homologous to a known gene, the *E. coli hisS*, coding for histidyl-tRNA synthetase (Delorme et al., 1992), however the *L. lactis* ORF lacks an essential motif implicated in synthetase activity (Fridman et al.,), suggesting that it has a role different from tRNA aminoacylation. What could be the function, if any, of this gene is not clear although a regulatory role in histidine biosynthesis was postulated (Delorme et al., 1992). ORF8 is partially homologous to Apha3 enzyme which inactivates aminoglycoside antibiotics (Threu-Cluot et al.,) but does not seem to be active for this function. Delorme et al. (1992) hypothesize that ORF8 could carry out dephosphorylation of HOL, the function of missing *hisB*px gene product. The remaining two ORFs were not homologous to any gene sequenced so far.

Two unknown ORFs were found in the *his* operon of *S. coelicolor* (Limauro et al., 1990), one very short and the other truncated at the end of the sequenced region. None of these ORFs shows homology with known genes.

Two ORFs were likewise detected in the *his* operon of *A. brasilense*. The protein coded for by the first ORF showed a significant homology with the product of *E. coli hisG* gene, although the homology is limited to the 20 amino-terminal amino acids out of the 168 forming the putative protein. Since the homologue of *E. coli hisG* was not found yet in *A. brasilense*, the possibility that the product of ORF168 actually accomplishes the first step of histidine biosynthesis cannot be excluded. The second *A. brasilense* unknown ORF (ORF122) is localized at the end of the operon, just before the transcription terminator. The gene product of this ORF was found to be homologous with two apparently unrelated proteins: the cow IPCK-1, inhibitor of protein kinase (Mozier et al., 1991), and a cyanobacterial protein with unknown function (?). These homologies however do not provide any convincing explanation for the function of the putative protein coded for by ORF122.

In the three *Methanococcus* species where *his* genes have been sequenced, they are surrounded by nine unknown ORFs. None of these ORFs, or their products, show homology with known genes. Some of them however seem to be partially homologous to each other (Weil et al., 1987): these are the cases of ORF547, ORF294 and ORF114 whose sequence lay upstream of *hisA*, and of ORF150 and ORF145 downstream *hisA*. ORF206 appear to be deleted from the genome of the two mesophilic *Methanococcus* species, as only few nucleotides in the intergenic region upstream *hisA* of *M. voltae* and *M. vannielii* were left to witness for the deletion event (Weil et al., 1987). Whether the abundance of unknown ORFs among *Archaeabacteria* is just the consequence of a different, and still unexplained, histidine biosynthetic pathway or it is a common feature of archaebacterial clusters (operons) to be interrupted by unrelated genes, it is not known. Many more data must be gathered to answer the question, in particular

major insights on the function of unknown ORFs will probably be obtained from the addition of new DNA sequences to the gene-bank.

Overlapping genes. The analysis of *E. coli* and *S. typhimurium* sequences revealed that almost all the *his* cistrons of the operons have overlapping termination and initiation signals for translation (Carlomagno et al., 1988). In particular the stop codons of *hisD*, *hisC*, *hisB*, and *hisH*, overlap by one or four bases the ATG of the following gene, while the two last cistrons (*hisA*/*hisF* and *hisF*/*hisI*) overlap by 19 and 7 bases. Therefore the ribosome binding sites of these genes lie inside the preceding cistron. This kind of organization, that was found in other polycistronic operons (...?...), is supposed to be associated with the occurrence of "translational coupling", a mechanism by which ribosomes start translating a new gene without moving away from mRNA after terminating the translation of the preceding one (.....?....).

The exceptions to overlapping are represented by the regions between the leader peptide and *hisG* and between *hisG* and *hisD*. This latter intercistronic sequence is composed of five nucleotides in *E. coli* and of 102 nucleotides in *S. typhimurium*. The unusually long *S. typhimurium* sequence was analyzed in detail (Carlomagno et al., 1988) and shown to contain a kind of REP (repetitive extragenic palindromic) sequence, whose function is possibly connected to the regulation of gene expression at the post-transcriptional level (Belasco and Higgins, 1988).

Overlapping cistrons were also detected in the *his* operons of *A. brasilense* (Fani et al., 1993), *S. coelicolor* (Limauro et al., 1992) and *L. lactis* (Delorme et al., 1992). In these organisms however the overlapping of stop and start signals is not a general phenomenon like in the *Enterobacteriaceae*; in particular intergenic regions were mostly detected where ORFs with unknown function are present (Table 2). The meaning of this discrepancy, of either regulatory or evolutionary origin, is not known at the present.

The compactness of *his* operons structure, particularly those of *E. coli* and *S. typhimurium*, could somehow be related to the regulation of operon expression at the translational level through translational coupling (Oppenheim and Yanofsky, 1980), experimental evidences however did not support this hypothesis.

Gene order. As already remarked, the order of the genes in the *his* clusters or operons does not coincide with the order of the enzymatic steps carried out by the proteins they code for, moreover the order of the genes is not rigorously comparable. Observing however the eubacterial clusters it is possible to recognize that some genes tend to remain together in the same order. This is particularly apparent for the four genes *hisBHAF* and partially also for the pair *hisDC*. The tendency of these genes to form stable clusters can be related with the particular role of their products in the biosynthetic pathway. The evolutionary history of these genes and of these particular enzymes and their significance in the evolution of *his* operon will be discussed in a following chapter.

REGULATION OF HISTIDINE BIOSYNTHESIS

The considerable metabolic cost required for histidine biosynthesis accounts for the evolution, in different organisms, of multiple and complex strategies to tune finely the rate of synthesis of this amino acid to the changeable environment.

In *S. typhimurium* and in *E. coli* the biosynthetic pathway is under control of distinct regulatory mechanisms which operate at different levels. Feed-back inhibition by histidine of the activity of the first enzyme of the pathway (see above) almost instantaneously adjust the flow of intermediates along the pathway to the availability of exogenous histidine. Transcription attenuation at a regulatory element, located upstream of the first structural gene of the cluster,

allows coordinate regulation of the amounts of the histidine biosynthetic enzymes in response to the levels of charged histidyl-tRNA (Artz and Holzschu, 1983; Blasi and Bruni, 1981; Winkler, 1987).

In addition to histidine, the system is also regulated by other molecules whose levels are indicative of the energetic and metabolic state of the cell. It has been previously mentioned that PRPP and ATP stimulate the activity of the first enzyme of the pathway, whereas AMP enhances the inhibitory effect of the histidine on this enzyme. Moreover, the alarmone guanosine 5'-diphosphate 3'-diphosphate (ppGpp) which is the effector of the stringent response (Cashel and Rudd, 1987) regulates positively *his* operon expression by stimulating transcription initiation under condition of moderate amino acid starvation or in minimal media growing cells (Artz and Holzschu, 1983; Cashel and Rudd, 1987; Winkler, 1987).

Transcription of the *his* operon is also modulated by a non-specific mechanism operating during the elongation step at the level of intracistronic Rho-dependent terminators (Alifano et al., 1988; Alifano et al., 1994a). These regulatory elements account for the polar phenotype exhibited by several nonsense and frame-shift mutations (Adhya and Gottesman, 1978; Alifano et al., 1988; 1991; Ciampi et al., 1989; Ciampi and Roth, 1988). Their physiological significance should be to prevent further elongation of infrequently translated transcripts (Alifano et al., 1988; Alifano et al., 1994a; Richardson, 1991).

Finally, it has been recently documented that post-transcriptional events contribute substantially to *his* operon expression. The native polycistronic *his* message is degraded with a net 5' to 3' directionality originating processing products which decay with differential rates (Alifano et al., 1992; 1994a; 1994b).

The differential levels of regulation of *his* operon expression will be discussed in more detail in the next sections. Particular emphasis will be given to the contribute of this experimental system as a powerful tool to the solution of general aspects of control of gene expression in prokaryotes.

Regulation of Transcription Initiation

The primary *hisp1* promoter. The *his* operon of *S. typhimurium* and *E. coli* (Fig. 3) is transcribed into a polycistronic mRNA extending from the primary promoter, *hisp1*, to the bidirectional Rho-independent terminator located at the end of the gene cluster (Carlomagno et al., 1983; 1985; 1988; Freedman and Schimmel, 1981; Frunzio et al., 1981; Verde et al., 1981).

The primary *hisp1* promoter in *S. typhimurium* and in *E. coli* was initially defined by DNA sequence analysis and by location of transcription start point (Barnes, 1978; Carlomagno et al., 1988; Di Nocera et al., 1978; Freedman and Schimmel, 1981; Frunzio et al., 1981; Verde et al., 1981). By inspection of the nucleotide sequence it was enclosed among the $E\sigma^{70}$ class of promoters (Harley and Reynolds, 1987; Mulligan et al., 1984). In the two closely related microorganisms the structure of this genetic element is almost identical and it is characterized by the presence of three out of six matches to consensus base pairs in both the - 10 and - 35 hexamers with a less conserved suboptimal 18 base pair separation between the hexamers. More recently, detailed mutation analysis has provided a rigorous evidence that the *hisp1* promoter of *S. typhimurium* is a $E\sigma^{70}$ promoter: all the base pair substitution mutations altered position in either the - 10 or - 35 hexamer (Shand et al., 1989a); conversely, mutations that improved the match to consensus of the - 10 hexamer increased its intrinsic strength (Riggs et al., 1986). The *hisp1* promoter is considerably strong both in vivo and in vitro being about four times stronger than the *gal* promoter in vivo (Verde et al., 1981; Riccio et al., 1985). In vitro its activity is about 20-fold higher on supercoiled than on linear templates (Verde et al., 1981). However, it has been recently shown that in vivo treatment of cells with the DNA gyrase inhibitor novobiocin enhanced the expression of a *his-lac* fusion bearing the complete *his* control region and did not modify substantially the expression of a

his-lac fusion bearing a 35 base pairs deletion of the *his* attenuator (O'Bryne et al., 1992). Moreover, repression of *his* expression by anaerobiosis or high osmolarity, two environmental parameters which increase the negative supercoiling of bacterial DNA, required the intact *his* attenuator sequence (O'Bryne et al., 1992). These findings are consistent with previous observations suggesting that the well documented modulation by supercoiling of *his* expression is mostly exerted at the unlinked *hisR* locus codifying for the only $tRNA^{His}$ in the cell (Brenner and Ames, 1971; Davis and Williams, 1982; Figueroa et al., 1991; Rudd and Menzel, 1987; Toone et al., 1992).

In 1968 Venetianer was the first to propose the existence of positively regulated genes during the stringent response noting that the *his* mRNA accumulated in the cell following amino acid starvation (Venetianer, 1968; Venetianer, 1969). Up to now, most information about the positive control derive from the *his* operon of *S. typhimurium*. In 1975 Stephens et al. () formally demonstrated that the expression of the *his* operon is subjected to metabolic regulation being the biosynthetic enzyme levels lower in amino acid-rich than in minimal-glucose media supplemented with histidine. They concluded that the *relA* gene product is required for the maximal *his* operon expression in vivo and observed that a strong stimulation of transcription occurred when ppGpp was added to cell-free S30 extracts. The stimulatory effect of ppGpp is believed to be on transcription and not on translation, and is promoter-specific as it occurs even when the *his* attenuator is deleted (Stephens et al., 1975; Winkler et al., 1978).

his operon transcription is maximally stimulated at lower than the maximum intracellular ppGpp concentration (Stephen et al., 1975; Winkler et al., 1978). The saturating concentration of ppGpp is thought to be slightly more than the basal level of minimal-glucose grown *S. typhimurium* cells. This would account for the scarce increase in *his* expression following amino acid downshift in a stringent strain grown in minimal media. In contrast, the attenuator-independent *his* expression in vitro has been shown to vary over a 10- to 20-fold

range in correlation with ppGpp levels (Stephen et al., 1975). In the attempt to define a precise correlation between *his* operon expression and ppGpp levels in vivo, Shand *et al.* (1989b) have developed a mild starvation method by using the serine analog serine hydroxamate, which increases ppGpp levels in a *relA*$^+$ strain and decreases ppGpp levels in a *relA* mutant. This method allowed to show that the full range of regulation of attenuator-independent expression in vivo is 20- to 40-fold and well correlates with the intracellular ppGpp concentrations.

Although mechanisms operating at the levels of transcription elongation, mRNA decay or translation cannot be rigorously excluded (Faxén and Isaksson, 1994; Nègre et al., 1989), several lines of evidences support the conclusion that ppGpp modulates *his* operon expression at the level of transcription initiation. First, the range of stimulation of *hisp1* activity by physiological concentrations of ppGpp in in vitro transcription systems (Stephens et al., 1975; Riggs et al., 1986) is very close to the overall range of stimulation of *his* operon expression observed in vivo (Shand et al., 1989b). Second, mutations that increased the homology of the -10 hexamer of *hisp1* to the consensus sequence of the $E\sigma^{70}$ promoters or that altered the sequence between the -10 hexamer and the start point dramatically enhanced *his* operon transcription in vitro in the absence of ppGpp, and reduced the stimulation of this alarmone to less than a factor of 2 (Riggs et al., 1986). On the basis of these results, Riggs et al. (1986) argued that promoters presumed to be positively regulated by ppGpp are partially defective in open complex formation bearing a suboptimal Pribnow box which does not contain an A residue in the fourth position, as is characteristic of negatively controlled promoters, in addition to differences in the region between the - 10 hexamer and the start point (Travers, 1984).

The internal *hisp2* and *hisp3* promoters. In addition to *hisp1*, two weak internal promoters, designated *hisp2* and *hisp3*, have been mapped both genetically and physically proximal to the *hisB* and *hisI* cistrons, respectively

(Atkins and Loper, 1970; Ely and Ciesla, 1974; Grisolia et al., 1983; Schmid and Roth, 1983). Although quite common in large bacterial operons, the physiological significance of these genetic elements is controversial. Even though it is possible that these promoters are physiologically unimportant and their presence merely fortuitous, their maintenance by selective pressure in homologous genomic regions of related microorganisms (Bauerle and Margolin, 1967; Jackson and Yanofsky, 1972) supports their physiological relevance. They could reinforce the expression of distal cistrons of large operons thereby alleviating the effects of natural polarity. Alternatively, they could allow regulation of an operon in a non-coordinate fashion and cause temporally different expression of certain genes under specific growth conditions (see below).

According to several features of the nucleotide sequence, *hisp2* promoter belongs to the $E\sigma^{70}$ class of promoters. It is subjected to metabolic regulation although to a less extent than the primary *hisp1* promoter, being its activity only two-fold lower in rich than in minimal-glucose media (Grisolia et al., 1983; Winkler et al., 1978).

The overall contribution of this internal promoter to the expression of the distal genes of the *his* operon is negligible when transcription proceeds from *hisp1* in wild-type cells growing in minimal-glucose media (Ely and Ciesla, 1974). However, its activity increases about three-fold when transcription from *hisp1* in wild-type is abolished, whereas it is almost completely inhibited when transcription from the upstream promoter is very efficient, as in constitutively derepressed mutants (Ely and Ciesla, 1974; Alifano et al., 1992). Such an inhibition of promoter activity by transcription read-through has been called "promoter occlusion" and might result from direct steric hindrance of an internal initiation site by RNA polymerase molecules initiating upstream or from distortion of DNA structure (Adhya and Gottesman, 1982; Bateman and Paule, 1988; Cole and Honoré, 1989; Jink-Robertson and Nomura, 1987).

Regulation of Transcription Elongation and Termination

The attenuation control. In both *S. typhimurium* and *E. coli* expression of the *his* structural genes is coordinately modulated in response to the availability of charged histidyl-tRNA by an attenuation mechanism of transcription at the level of the leader region preceding the first structural gene (Artz and Broach, 1975; Barnes, 1978; Di Nocera et al., 1978; Johnston et al., 1980; Kasai, 1974; Kolter and Yanosfky, 1982). Since regulation by attenuation has been reviewed in the past more than once, the readers are referred back to more exhaustive articles for an historical picture of the fundamental steps leading to the discovery of this control mechanism (Artz and Holzschu, 1983; Blasi and Bruni, 1981; Winkler, 1987).

The purpose of this section is to present the current mechanistic model of attenuation as formulated on the basis of the numerous evidences accumulated up so far, and to report more recent insights concerning particular aspects of this phenomenon which may have a relevance in understanding fundamental mechanisms governing transcription elongation-termination.

Two more prominent features characterize the leader region of the *his* operon which may account for *his*-specific translational control of transcription termination which is the essence of the attenuation control: i. a short coding region that includes numerous tandem codons specifying histidine (seven histidine codons in a row of sixteen); ii. overlapping regions of dyad symmetry that may fold into alternative secondary structures, one of which includes a ρ-independent terminator. In the "termination" configuration, base pairing involves regions A and B, C and D, and E and F (Fig. 4). The stable stem-loop structure E:F followed by a run of uridylate residues constitutes a strong intrinsic terminator. In the "antitermination" configuration, base pairing between B and C, and D and E prevents formation of the terminator thus allowing readthrough transcription. The equilibrium between these alternative

configurations is determined by the ribosome occupancy of the leader region which in turn depends on the availability of charged histidyl-tRNA.

Low levels of the specific charged tRNA will cause ribosome to stall on the leader region in correspondence of the histidine codons, and to disrupt A:B pairing by masking the region A. Under these circumstances, the "antitermination" configuration will be favored. Conversely, in the presence of high levels of charged histidyl-tRNA, ribosome will move faster away from the histidine regulatory codons thereby occupying both A and B regions. Pairing between C and D, and E and F will result in premature transcription termination.

Impairment of translation of the leader region as a consequence of severe limitation of the intracellular pool of all charged tRNAs will result in strong transcription termination ("superattenuation"). Under these conditions, A:B, C:D and E:F stem-loop structures will form sequentially without interference by active translating ribosomes.

A characteristic features common to all known examples of attenuation is a significant RNA polymerase pausing after synthesis of the first leader transcript RNA hairpin. This pause event is believed to synchronize transcription and translation in the leader region by halting the transcribing RNA polymerase until a ribosome commences synthesis of the leader peptide. Absence of synchronization would not allow complete relief of termination in the presence of low levels of charged histidyl-tRNA (Landick, 1987; Landick et al., 1987; Landick and Yanofsky, 1987; Winkler and Yanofsky, 1981).

In the *his* leader region RNA polymerase pausing occurs after synthesis of the first secondary structure (A:B) and immediately prior to addition of a G residue (Chan and Landick, 1989). The position of the pause site would allow ribosome initiating synthesis of the leader peptide to release the paused transcription complex by disrupting the pause hairpin.

Analysis of the effects of base substitutions upstream from the pause site revealed that the pausing signal is multipartite and consists of at least four distinct components: i. a 5 bp nascent transcript stem-loop structure (the pause hairpin); ii. the 11 nucleotides 3'-proximal segment of transcript or DNA template; iii. the 3'-terminal nucleotide; iv. the immediate downstream DNA sequence (Chan and Landick, 1989; 1993; Lee et al., 1990; Lee and Landick, 1992). The behavior of compensatory substitutions in the A:B hairpin region suggested that the *his* pause hairpin corresponds to only the upper portion of the larger A:B secondary structure (Chan and Landick, 1993). Such a finding is consistent with the deduced structure of RNA and DNA chains in purified transcription complexes paused at the *his* leader region which has been determined by analyzing the reactivity of specific residues on DNA to chemical modifying agents and the sensitivity of the nascent RNA molecules to ribonuclease A (Lee and Landick, 1992). This analysis evidenced that, in spite of a considerable variation of dimension of the transcription bubble during elongation, the 3'-proximal nucleotides of transcript constantly pair with the DNA template and that the DNA:RNA hybrid is not disrupted by hairpin formation at the pause site. This finding ruled out the possibility that extensive secondary structures halt elongation by removing the 3' end of the transcript from the catalytic site of RNA polymerase, or by disrupting pairing between the 3'-proximal segments of transcript and DNA template (Landick and Yanofsky, 1984; Landick, 1987). Chan and Landick (1993) suggested that pausing is mediated in part by non-sequence specific, electrostatic interactions between the phosphate backbone of the pause RNA hairpin and a positively charged region on RNA polymerase. NusA elongation factor enhances pausing possibly by directly contacting the pause hairpin or by increasing its interaction with RNA polymerase (Chan and Landick, 1989; 1993).

Polarity. In polycistronic operons certain mutations which cause premature arrest of translation not only affect the gene in which they occur, but

expression of downstream genes. This phenomenon is commonly
Albeit it were first described in the lactose system (Franklin
Jacob and Monod, 1961b), the coordinate effect of polar
pression of downstream cistrons and the existence of "polarity
better defined in the *his* system by using a large collection of
widespread in different cistrons (Ames and Hartman, 1963;
, 1971; Fink and Martin, 1967; Hoppe et al., 1979; Johnston
Martin and Talal, 1968; Martin et al., 1966a; Rechler et al.,
and the physiological significance of polarity has been
long time (see Introduction). The phenomenon has been
lating the existence of intracistronic cryptic Rho-dependent
and Gottesman, 1978; de Crombrugghe et al., 1973).
model of transcriptional polarity (Adhya and Gottesman,
translation would favor binding of Rho to the nascent
this terminator factor with elongating RNA polymerase
transcript.
pported by more recent studies on polarity in the
Several cryptic Rho-dependent terminators which
ranscription and translation have been identified
ing the origin of truncated *his*-specific transcripts
oring polar mutations scattered in four cistrons
ranscription studies (Alifano et al., 1988;
1994a; Ciampi and Roth, 1988; Ciampi et al.,

the intracistronic terminators in the *his*
features of these signals and to better
m by which Rho causes termination of
tif consisting of a cytosine-rich and
upstream of the heterogeneous 3'-end

points of the prematurely terminated transcripts has been detected in all Rho-dependent terminators analyzed in the *his* operon as well as in other systems (Alifano et al., 1991; Rivellini et al., 1991). This region (TTE = transcription termination element) is believed to be the binding-activation site of Rho protein on nascent RNA (Bear et al., 1988; McSwiggen et al., 1988; Yager and von Hippel, 1987). Deletions which eliminate or reduce the extent of these TTEs impair transcription termination (Chen and Richardson, 1987; Galloway and Platt, 1988) and relieve polarity (Ciampi and Roth, 1988; Ciampi et al., 1989). Release of transcripts occurs downstream to TTEs at multiple sites (TS = termination sites) in coincidence with RNA polymerase pause sites (Alifano et al., 1991; Rivellini et al., 1991).

The occurrence of more than one TTE in a given cistron accounts for the higher degree of polarity generated by promoter-proximal than distal mutations. Moreover, the variable number and spatial distribution of TTEs within different cistrons explains the differences in shapes of polarity gradients. The existence of two closely spaced TTEs in the proximal part of *hisG* accounts for the discontinuity of the polar effects (Fink and Martin, 1967). Location of a single TTE toward the distal end of *hisA* results in a similar degree of polarity exhibited by all mutations irrespective of their relative position in this cistron (Fink and Martin, 1967). The sequential use of more than one TTE results in an additive effect for the more proximal mutations in *hisD* and generates a real gradient in the cistron (Fink and Martin, 1967). The bimodal shape of the gradient in *hisC* (Martin et al., 1966) is due both to the presence of several TTEs tightly clustered in the proximal and the distal regions of the cistron and to the occurrence of concomitant processing events (Alifano et al., 1991).

The unusual features of a class of polar and prototrophic mutations which map in the intercistronic *hisD-hisC* region has contributed to understand the physiological significance of intracistronic termination (Rechler et al., 1972; Alifano et al., 1988). In the wild-type *his* operon, the *hisD* and *hisC* cistrons are

also reduce the expression of downstream genes. This phenomenon is commonly called "polarity". Albeit it were first described in the lactose system (Franklin and Luria, 1961; Jacob and Monod, 1961b), the coordinate effect of polar mutations on expression of downstream cistrons and the existence of "polarity gradients" were better defined in the *his* system by using a large collection of polar mutations widespread in different cistrons (Ames and Hartman, 1963; Brenner and Ames, 1971; Fink and Martin, 1967; Hoppe et al., 1979; Johnston and Roth, 1979; Martin and Talal, 1968; Martin et al., 1966a; Rechler et al., 1972). The nature and the physiological significance of polarity has been controversial for a long time (see Introduction). The phenomenon has been explained by postulating the existence of intracistronic cryptic Rho-dependent terminators (Adhya and Gottesman, 1978; de Crombrugghe et al., 1973). According to a general model of transcriptional polarity (Adhya and Gottesman, 1978) premature arrest of translation would favor binding of Rho to the nascent transcript, interaction of this terminator factor with elongating RNA polymerase and subsequent release of transcript.

This model has been supported by more recent studies on polarity in the *his* operon of *S. typhimurium*. Several cryptic Rho-dependent terminators which are activated by uncoupling of transcription and translation have been identified within the *his* operon by analyzing the origin of truncated *his*-specific transcripts produced in vivo in strains harboring polar mutations scattered in four cistrons of the *his* operon and by in vitro transcription studies (Alifano et al., 1988; Alifano et al., 1991; Alifano et al., 1994a; Ciampi and Roth, 1988; Ciampi et al., 1989; Rivellini et al., 1991).

The comparative analysis of the intracistronic terminators in the *his* operon allowed to identify common features of these signals and to better understand the molecular mechanism by which Rho causes termination of nascent transcripts. A consensus motif consisting of a cytosine-rich and guanosine-poor region that is located upstream of the heterogeneous 3'-end

points of the prematurely terminated transcripts has been detected in all Rho-dependent terminators analyzed in the *his* operon as well as in other systems (Alifano et al., 1991; Rivellini et al., 1991). This region (TTE = transcription termination element) is believed to be the binding-activation site of Rho protein on nascent RNA (Bear et al., 1988; McSwiggen et al., 1988; Yager and von Hippel, 1987). Deletions which eliminate or reduce the extent of these TTEs impair transcription termination (Chen and Richardson, 1987; Galloway and Platt, 1988) and relieve polarity (Ciampi and Roth, 1988; Ciampi et al., 1989). Release of transcripts occurs downstream to TTEs at multiple sites (TS = termination sites) in coincidence with RNA polymerase pause sites (Alifano et al., 1991; Rivellini et al., 1991).

The occurrence of more than one TTE in a given cistron accounts for the higher degree of polarity generated by promoter-proximal than distal mutations. Moreover, the variable number and spatial distribution of TTEs within different cistrons explains the differences in shapes of polarity gradients. The existence of two closely spaced TTEs in the proximal part of *hisG* accounts for the discontinuity of the polar effects (Fink and Martin, 1967). Location of a single TTE toward the distal end of *hisA* results in a similar degree of polarity exhibited by all mutations irrespective of their relative position in this cistron (Fink and Martin, 1967). The sequential use of more than one TTE results in an additive effect for the more proximal mutations in *hisD* and generates a real gradient in the cistron (Fink and Martin, 1967). The bimodal shape of the gradient in *hisC* (Martin et al., 1966) is due both to the presence of several TTEs tightly clustered in the proximal and the distal regions of the cistron and to the occurrence of concomitant processing events (Alifano et al., 1991).

The unusual features of a class of polar and prototrophic mutations which map in the intercistronic *hisD-hisC* region has contributed to understand the physiological significance of intracistronic termination (Rechler et al., 1972; Alifano et al., 1988). In the wild-type *his* operon, the *hisD* and *hisC* cistrons are

overlapped in that the termination UGA codon of the proximal *hisD* cistron and the start AUG codon of the distal *hisC* cistron share the central UG dinucleotide (AUGA; Riggs and Artz, 1984; Carlomagno et al., 1988). Mutation *hisD2352*, which maps in this region, is a G-C/T-A transversion, which changes the stop codon UGA of *hisD* cistron to a sense codon thus allowing the ribosome to continue translation until a new stop codon is encountered 30 nucleotides downstream. At the same time, the transversion changes the initiation codon of the *hisC* cistron from AUG to the triplet AUU which can still be used as an initiation codon, albeit with a reduced efficiency. This genetic mutation generates a strong polar phenotype by inducing premature transcription termination in *hisC* in spite of the persistence of residual levels of translation (Alifano et al., 1988). At a certain extent premature transcription termination has been detected in *hisC* in wild-type strains grown in the presence of antibiotics which are believed to lower the intracellular pool of fMet-tRNAf thereby impairing efficient translation initiation (Alifano et al., 1994a). These evidences support the proposal that transcriptional polarity does not require complete arrest of translation (Stanssens et al., 1986) and that the degree of premature Rho-dependent termination is inversely proportional to the efficiency of translation (Alifano et al., 1988; Alifano et al., 1994a; Richardson, 1991).

Post-Transcriptional Regulation: mRNA Processing and Decay

The unstable primary 7,300-nucleotides long transcript of the *his* operon has a half life of about 3 min in cells growing in minimal-glucose medium and is degraded with a 5' to 3' directionality. The decay process generates three major processed species, 6,300, 5,000 and 3,900 nucleotides in length, that encompass the last seven, six and five cistrons, respectively, and have increasing half-lives (4, 6 and 15 min, respectively) (Alifano et al., 1992; Alifano et al., 1994a; Carlomagno et al., 1988). The pattern of *his* mRNA decay is identical both in *S. typhimurium* and in *E. coli* irrespective of the presence of a 102 nucleotides long REP sequence in the intercistronic *hisG-hisD* region of the former microorganism which is absent in the other (see above and Carlomagno et al., 1988).

RNAase E controls the decay of the native transcript. Active translation of the 5'-end proximal cistrons of the processed species is required to stabilize temporarily these species. The overall process of decay may have functional relevance to balance the expression of the promoter proximal genes, which are the first to be transcribed, and the distal ones.

The most distal 3,900-nucleotides long processed species has a half life of about 15 min. The uncommon stability of this molecule suggests that the processing event that generates it has functional consequences. In fact, the processed species spans the distal cistrons which are involved, in addition to histidine biosynthesis, in a purine recycling pathway leading to production of the cellular "alarmone" ZTP (see above) (Alifano et al., 1994a). Notably, the same distal genes are also transcribed from the internal *hisp2* promoter which is subjected to promoter occlusion (see above).

The specific processing event leading to production of the 3,900 nucleotides species is mechanistically complex. It requires sequential cleavages by two endoribonucleases. RNAase E triggers the process by inactivating functionally the native transcript at the level of a major target site located in *hisC* cistron 620 nucleotides upstream of the 5'-end of the processed species and at the level of minor sites (Alifano et al., 1992; 1994b). RNAase P cleaves the processing products generated by RNAase E at a discrete site located 76 nucleotides upstream of the start codon of *hisB* cistron thus originating the mature 5' end (Alifano et al., 1994b). The RNAase P-dependent cleavage occurs at the 5'-end of a region that may fold into a short stem-loop structure followed by a 3'-distal NCCA sequence. RNA molecules with such features have been proposed to be minimal substrates for this endoribonuclease (Altman, 1993;

Forster and Altman, 1990; Schuster, 1995?). The considerable stability of the processed species may be conferred by the stem-loop structure sequestering the 5'-end of the mature RNA generated by RNAase P (Bouvet and Belasco, 1992).

Translational events modulate the mRNA processing efficiency. The RNAase P-catalyzed reaction requires binding of ribosome at the ribosomal binding site of *hisB* cistron (Alifano et al., 1992; 1994b). Ribosomes initiating translation of *hisB* might favor the formation of the RNAase P-targeted structure allowing RNAase P to cleave mRNA efficiently. Alternatively, they might contribute to stabilize the processed species by arresting temporarily the 5' to 3' wave of decay.

Metabolic perturbation of the translation process caused by limitation in the intracellular pool of initiator tRNA results in an increase in the amount of the processed species in vivo (Alifano et al., 1994a). This effect may be due to two mechanisms. Low levels of initiator tRNA might uncouple transcription and translation thereby improving the exposure of target sites to RNAase E which triggers the process. Moreover, reduction of the intracellular levels of initiator tRNA might affect processing by altering the kinetics of formation of the initiation complex at the intercistronic regions and causing stalling of ribosomes at the *hisB* ribosome binding site (Petersen et al., 1976a, 1976b). Ribosome stalling would in turn result into a stabilization of the RNAase P targeted structure. Even though translation is likely to control mRNA processing and decay in this as well as in other systems (reviewed in Alifano et al., 1994c; Belasco and Higgins, 1988; Petersen, 1992), the exact links of these distinct cellular processes should be clarified in more details.

REGULATION OF HISTIDINE BIOSYNTHESIS IN OTHER SPECIES

As discussed above, regulation of *his* operon expression in *E. coli* and *S. typhimurium* has been the subject of very intensive studies and the general

mechanisms and the molecular details of the process are fairly well established. On the contrary very few studies in this area have been performed in other prokaryotic cells. In general it seems that while the biochemical reactions leading to histidine biosynthesis are the same in all the organisms, the overall genomic organization, the structure of the *his* genes (see pertinent sections) and the regulatory mechanisms by which the pathway is regulated differ widely in tassonomic unrelated groups.

In the closely related organism *K. pneumoniae* the overall genomic organization, at least of the proximal and distal regions, appears to be conserved (Rieder et al., 1994; Rodriguez et al., 1981 and Table 2). DNA sequence analysis of the regulatory region and promoter-expression studies under different metabolic conditions and in regulatory mutants also indicate that the mechanisms controlling the histidine biosynthesis are well conserved (Rodriguez and West, 1984). In the other Gram- negative nitrogen-fixing bacterium *A. brasilense* a *his* operon comprising five genes has been cloned and characterized (see Organization of the Histidine Genes section) but no information on its regulation is available although a partial characterization of the transcripts has been performed (Fani et al., 1993).

Although several *his* biosynthetic genes have been cloned and sequenced from organisms belonging to the Gram-positive group of eubacteria (Hinshelwood and Stoker, 1992; Sonenshein, 1993 and see appropriate section), virtually nothing is known about regulation of histidine biosynthesis in these species. *S. coelicolor his* genes expression appears to be regulated by the intracellular histidine levels (Carere et al., 1973; Limauro et al., 1992) and the transcription initiation site of the cluster comprising five *his* genes (Limauro et al., 1990) has been determined (Limauro et al., 1992).

Finally, *his* biosynthetic genes from the other unrelated group of prokaryotes, the archaebacteria, have also been identified and characterized (see

Organization of the Histidine Genes section) but again studies of regulation of their expression have not been performed.

In all the above mentioned systems with the exception of *K. pneumoniae* the only conclusion that can be drawn is that regulation by attenuation can be excluded since the 5'-proximal regions of all the genes and gene clusters lack both the required palindromic structures and the histidine-rich leader peptide (Delorme et al., 1992). It is felt that studies in this area, making use of the different *his* systems, might be very important and rewarding to unravel other mechanisms by which simple unicellular organisms regulate expression of biochemical pathways.

The situation is completely different for the organisms belonging to the lower eukaryotes group and in particular for *S. cerevisiae*. In these cells the mechanisms that regulate histidine and other amino acids biosynthetic pathways are well understood and the *his* genes have been extensively used as a model to study these regulatory systems. Two different *trans*-acting systems are operational. The general amino acid control, which activates transcription of more than 30 genes in eleven biosynthetic pathways in response to amino acid starvation, and the basal level control, which maintains transcription in the absence of amino acid starvation. Both systems are dependent on transcriptional activators (GCN4, and BAS1 and BAS2) and a series of accessory factors including kinases and phosphatases. A detailed analysis of these mechanisms goes far beyond the scope of this article and the interested readers are referred to several excellent reviews that cover this subject (Arndt et al., 1987; Hinnnebush, 1988; Hinnebush, 1990).

EVOLUTION OF THE METABOLIC PATHWAY

Synthesis of Histidine in Possible Prebiotic Conditions

It is very likely that histidine was formed during the long period of chemical abiotic synthesis of organic compounds which, according to Oparin (1924, 1936), was a necessary precondition for the appearance of the first life form.

It is generally accepted that histidine is present in the active sites of enzymes because of the special properties of the imidazole group (Weber and Miller, 1981) which acts as a general acid-base catalyst in a number of biochemical reactions (Bender et al., 1984; Fersht, 1977). It is therefore reasonable to assume that His-containing small peptides could have been involved in the prebiotic formation of other peptides (White and Erickson, 1980) and nucleic acid molecules, once these monomers accumulated in primitive tidal lagoons or ponds. The synthesis of histidine must involve at some step the formation of an imidazole group. Nevertheless in the first successful synthesis of organic compounds under plausible prebiotic conditions, accomplished by the action of electric discharges for a week over a mixture of CH_4, H_2, H_2O and NH_3 (Miller, 1953), histidine and/or imidazole was not present among the reaction products. The formation of imidazole from glyoxal, ammonia, and formaldehyde under plausible prebiotic conditions was demonstrated much later on (Orò et al., 1984). Three years later Shen et al. (1987) showed that imidazole-4-glycol and imidazole-4-acetaldehyde could be synthesized from erythrose and formamidine. Later on it was demonstrated that imidazole-4- acetaldehyde, present among the crude reaction products from erythrose and formamidine, without isolation, can be directly converted to histidine by a Strecker cianohydrin synthesis (Shen et al., 1990a); moreover the formation in good yields of the dipeptide histidyl-histidine by the evaporation of an aqueous solution of histidine in the presence of condensing agents was also demonstrated (Shen et al., 1990b). This dipeptide was shown to have an enhancing effect in some prebiotic reactions involving nucleotide derivatives and oligonucleotides, such as in the dephosphorylation of deoxyribonucleoside monophosphate, in the hydrolysis of oligo (A)12, and in the

oligomerization of 2',3'-cAMP under cyclic wet-dry laboratory reaction conditions simulating a primitive evaporating pond (Shen et al., 1990c). This body of data supports the hypothesis that simple peptides of prebiotic origin containing at least two imidazole groups could have played a significant role in the chemical events preceding the evolutionary development of enzyme biosynthesis (Shen et al., 1990c).

Finally, since the biosynthesis of histidine requires a carbon and a nitrogen equivalent from the purine ring of ATP, it has also been suggested that it may be the molecular descendant of a catalytic ribonucleotide from an earlier biochemical stage in which RNA played a major role in catalysis (White, 1976).

Origin of the Histidine Biosynthetic Pathway

If histidine was required by primitive catalysts, then the eventual exhaustion of the prebiotic supply of histidine and histidine-containing peptides (Shen et al., 1990a, b, c) must have imposed an important pressure favoring those organisms capable of synthesizing imidazole-containing compounds. There are several independent indications for the antiquity of the histidine biosynthesis pathway. Even though how the biological synthesis of histidine actually originated can only be surmised, the apparently universal phylogenetic distribution of the *his* genes (Table 1) suggests that the histidine pathway was already part of the metabolic abilities of the last common ancestor of the three extant cell lineages (Lazcano et al., 1992; Fani et al., 1995). This conclusion is also supported by the fact that at least three different *his* genes (*hisA*, *hisC* and *hisI*) have been identified in each of the three cell lineages and by the robustness of the *his* genes phylogenetic trees depicting the evolutionary distances between different microorganisms (Fani et al., 1995). These data suggest that the evolution of the 16S-like rRNA and that of the *his* genes were roughly parallel. Hence this metabolic pathway might have been assembled long before the

divergence of the three cell lineages, most probably in the early stages of life evolution. How the *his* pathway actually originated is still an open question, but several different theories have been suggested accounting for the establishment of anabolic routes. These explanations include: i. the retrograde hypothesis (Horowitz, 1945; 1965) according to which the present biosynthetic pathways were organized stepwise and backwards from the final metabolites of the pathways. Based on this hypothesis, Horowitz also proposed that the group of genes involved in one biosynthetic pathway were generated by duplication followed by divergence of a common ancestral gene; ii. the possibility that at least some biosynthetic routes evolved forwards, i.e., from simple precursors to complex end products (Granick, 1965); iii. the idea that metabolic pathways appeared as a result of the gradual accumulation of mutant enzymes with minimal structural changes (Waley, 1969); and iv. the patchwork theory, according to which metabolic routes are the result of the serial recruitment of relatively small, inefficient enzymes endowed with broad-specificity that could react with a wide range of chemically related substrates (Ycas, 1974; Jensen, 1976).

The comparative studies of the known histidine genes that are involved in the same metabolic pathway in different organisms belonging to the three cell lineages may contribute to understand how a metabolic pathway has assembled.

The cladistic analysis of the available *his* sequences indicated that paralogous gene duplications played a major role in shaping the pathway. This is indicated by the evidence of two successive duplications involving an ancestral module which eventually led to the *hisA* and *hisF* genes and their homologues (Fani et al., 1994, 1995) (see below). On the basis of this analysis it has been postulated that the *hisA* and *hisF* genes and their homologues are the descendants of a gene encoding a less specific enzyme (Fani et al., 1995), an evidence supported also by the finding that the HisF protein is able to interact, even with an apparent reduced affinity, to 5'-ProFAR, the substrate of HisA

enzyme (Klem and Davidsson, 1993). Moreover Sheridan and Venkataraghavan (1992) have proposed a potential substrate recognition in HisA and HisF on the basis of a common signature in both proteins which was assumed to be a strand-helix- strand structure that could bind glycerol-phosphate moieties. It is then possible that 5'-PRFAR or 5'-ProFAR could have been the substrate of the ancestral HisA enzyme (Fani et al., 1995). These data appear to support the so-called patchwork hypothesis (Ycas, 1974; Jensen, 1976), and are consistent with the possibility that an ancestral histidine pathway may have implied a primitive enzyme catalyzing two or more similar reactions and whose substrate specificity was refined as a result of later duplication events. The possibility that histidine biosynthesis was originally mediated by less specific enzymes is strongly supported by the common origin of the imidazole glycerol-P synthase encoded by the enterobacterial *hisH* gene, with other *E. coli* G- type glutamine amidotransferases (GAT) which participate in the biosynthesis of purines, pyrimidines, arginine, tryptophan, and other ancient pathways (Fani et al., 1995). Although for the time being the lack of sequences of G-type GAT from the three cell lineages limits a complete evolutionary analysis, similarity data available suggest that they may be the descendants of an ancient, less-specific glutamine amidotransferase which mediated the transfer of the amide group of glutamine to a wide range of substrates (Fani et al 1995).

Evolution of *his* Genes

The evolutionary comparison of the *his* genes in the three cell lineages clearly indicates that, after the divergence from the last common ancestor, the structure, organization and order of these genes have undergone several major rearrangements in the different cellular lineages (Fig. 2). Although it is not possible from the analysis of the available data to infer the organization of the *his* genes in the last common ancestor (that is whether these genes were

clustered or scattered throughout its genome), nevertheless the same analysis may help to elucidate the primitive structure of some *his* genes and their evolution. It is worth noticing that some of them have undergone gene duplication and/or gene fusion events. In particular gene fusion, as well as gene duplication, appears to be one of the most important mechanisms of gene evolution in the histidine biosynthetic pathway. Several fusion events have occurred in both the genomes of bacteria and some eukaryotes, leading to longer genes encoding for bi- or multifunctional enzymes. Although gene fusions can be selected for substrate channeling, they also represent an effective mechanism ensuring the coordinate synthesis of two or more enzymatic activities. This may have special significance among eukaryotes, where the absence of operons does not allow coordinate regulation by polycistronic mRNAs (Davidson et al., 1993).

Evolution of *hisB*. As described above, in enterobacteria *hisB* codes for a bifunctional enzyme catalyzing the sixth and the eighth steps of histidine biosynthesis (Winkler, 1987). The most widely accepted model for the association of these two enzymatic activities of the *hisB* gene product predicts the existence of two independent domains in the gene, i.e., a proximal domain encoding the phosphatase moiety, and a distal one encoding the dehydratase activity (Brady and Houston, 1973; Chumley and Roth, 1981; Loper, 1961). The structural organization of the two enzymatic activities in other microorganisms where they are encoded by two separate genes confirms the two domain model discussed above (Fig. 2). Taken altogether, the available data support the idea that a bifunctional *hisB* gene is an enterobacterial peculiarity, i.e., it seems likely that the evolution of the *hisB* gene in *E.coli* and *S. typhimurium* could have involved the fusion of two independent cistrons, *hisB*px and *hisB*d, coding for a HOL-P phosphatase and an IGP dehydratase, respectively (Fani et al., 1989; 1995). The lack of significant sequence homology between the two moieties of the enterobacterial *hisB* gene suggests that domain-shuffling rather than gene duplication or gene elongation could be responsible for its present-day structure.

Moreover, it is likely that this fusion event took place after the evolutionary splitting between the α and the γ branches of the purple bacteria, as *A. brasilense* also showed separate genes.

Evolution of *hisI* and *hisE*. As reported above *hisI* and *hisE* genes are fused in *E. coli*, *S. typhimurium*, *L. lactis*, *S. cerevisiae*, *P. pastoris* and *N. crassa*, but separated in *A. brasilense* and *M. vannielii*. Two contrasting hypotheses could explain this difference. In one of them an ancestral *hisIE* bifunctional gene gave rise in some prokaryotes to monofunctional genes by at least two independent splitting events. Alternatively two ancestral genes, each of which encoding a monofunctional enzyme catalyzing sequential steps in histidine biosynthesis, might have undergone many independent fusion events in different cell lineages.(Fani et al., 1995).

In lower eukaryotes the *hisI* and *hisE* genes are fused to *hisD* giving a larger multifunctional gene. As shown in Fig. 2, these genes share the same internal organization, with the "*hisD* domain" located at the 3'-terminal region. It has been previously suggested that this eukaryotic multifunctional gene originated from the fusion of bacterial separated cistrons (Bruni et al., 1986), but just how many gene fusion and/or gene elongation events led to the extant fungal genes is an open question. Moreover, since *HIS4* and *his-3* genes share the same internal organization, the putative fusion event(s) took place before the evolutionary separation of these fungi, but after the separation of fungi from plants, in fact the *hisD* gene of *B. oleracea* is not fused to the *hisI* and *hisE* genes demonstrating that the evolution of *his* genes followed different paths among eukaryotes.

Evolution of paralogous *his* genes. An additional intriguing feature of histidine biosynthetic genes concerns the two genes *hisA* and *hisF*. The accurate observation of the sequence of *hisF* gene and of its product demonstrated a remarkable homology with *hisA*, the nearest preceding gene in eubacterial operons, and with its product 5'-ProFAR isomerase (Fani et al., 1994). This

homology, recognized in all *hisA* and *hisF* genes sequenced, has been interpreted as the consequence of a gene duplication event during the evolution of the operon (see below). The same analysis moreover demonstrated the presence of two homologous moieties inside each of the two genes *hisA* and *hisF* thus appearing constituted by four relatively homologous modules: A1, A2, F1 and F2 that are the result of two successive duplication events: the first one involving the *hisA*1 module and leading to the extant *hisA* gene, which in turn duplicated and gave rise to the *hisF* gene (Fani et al., 1994).

In *S. cerevisiae* the *hisF* counterpart is represented by *HIS7*, which is constituted by two moieties corresponding to the *E. coli hisH* and *hisF* genes, respectively (Kuenzler et al., 1993). Analysis of the deduced amino acid sequence of the HisF moiety coded by the 3' region of *HIS7* revealed a high degree of sequence similarity with the prokaryotic HisA proteins, especially with the archaebacterial ones; this moiety, like its bacterial counterpart, is formed by two homologous modules half the size of the entire moiety (Fani et al., 1995). More recently, the *HIS6* gene of *S. cerevisiae*, homologous to the prokaryotic *hisA*, has been cloned and analyzed (Fani et al., unpublished data b). Its product shows a significant degree of sequence homology with the prokaryotic HisA proteins, with the known eubacterial HisF proteins and also with the 3' domain of yeast *HIS7*, indicating that also the *S. cerevisiae HIS6* and *HIS7* genes are paralogous (Fani et al., unpublished data). The *HIS6* gene also shows the "two-modules" structure typical of all the known *hisA* and *hisF* genes. This set of data suggests that *hisA* (and probably *hisF*, although an archaebacterial *hisF* gene has not been identified yet), was part of the genome of the last common ancestor and that the two successive duplication events leading to the extant *hisA* and *hisF* took place long before the diversification of the three evolutionary domains, probably in the early stages of the molecular evolution of the histidine pathway.

The evolutionary history of *hisH* also probably involves a duplication event. In fact, the sequence similarity of the different imidazole glycerol P

synthases encoded by the enterobacterial *hisH* genes with the sequences encoding the so-called G-type glutamine amidotransferases (GAT) genes encoding the anthranilate synthase (Nichols et al., 1980), 4-amino-4-deoxychorismate synthase (Kaplan and Nichols, 1983), carbamoyl-P synthase (Piette et al., 1984), GMP synthase (Tiedeman et al., 1985), CTP synthetase (Weng et al., 1986), and formylglycinamide synthetase (Schendel et al., 1989; Sampei and Mizobuchi, 1989), implies that these different enzymes are also the products of ancient paralogous duplications (Fani et al., 1995). From an evolutionary point of view particularly interesting is the *S. cerevisiae HIS7* gene which appears to be constituted by the two bacterial-like cistrons *hisH* and *hisF* (Kuenzler et al., 1993), which in turn are the results of two series of independent gene duplication events. As previously reported the *E. coli hisF* and *hisH* genes code for proteins which associate to form a heterodimeric enzyme. The genes encoding them are not adjacent in any of the studied bacteria, but are fused in yeast as a result not of an elongation event, but of a domain shuffling phenomenon (Fani et al., 1995). However the possibility that the *HIS7* gene could have originated from the fusion of two genes via the deletion of the intervening region cannot be ruled out. As reported before the 3'-terminal region of *HIS7* is considerably longer than eubacterial *hisF* genes, due to the presence of six insertions in the first module (*hisF*1) of *HIS7*. Since these insertions are not present in *HIS6*, it is likely that they have been incorporated in *HIS7* after the divergence of prokaryotes and eukaryotes. Nevertheless we cannot a priori rule out the possibility that these insertions could have been incorporated in an ancestral *HIS7* gene before this divergence. The cloning and analysis of archaebacterial and additional eukaryotic *hisF* genes might solve this issue.

Evolution of *his* clusters. Differences in the relative *his* gene order may be observed in those prokaryotes in which at least some of the histidine biosynthetic genes are clustered as reported before. Nevertheless, four of the clustered genes (*hisB*d, *hisH* and *hisA*, and also *hisF*, except the latter for *S. coelicolor*) are always present and share the same relative order. These four genes encode enzymes involved in the central, sequential enzymatic steps of the pathway, connecting histidine biosynthesis with nitrogen metabolism and the de novo synthesis of purines (see above).

It is possible that the four genes *hisBdHAF* could represent the core of the histidine biosynthesis, and that the *his* operon is an open plastic operon, i.e. an operon in which except for the highly maintained core of the pathway, different ways of gene organizations are possible (Fani et al 1995). Indeed, the known *his* operons show different gene organization and most of them (*A. brasilense*, *S. coelicolor*, *L. lactis*), also contain non homologous ORFs with unknown function. Moreover it is possible that this set of enzymes (or at least HisF and HisH proteins) are a "metabolon" as defined by Srere (1987), promoting the preferential transfer of an intermediate metabolite from one enzyme to a physically adjacent enzyme, and limiting its diffusion into the surrounding milieu (Mathews, 1993). The question whether these four genes could have act as a unitary block in the evolution of eubacterial *his* genes will be solved when additional *his* operons from other different eubacteria will be described in detail (Fani et al 1995)

Molecular Phylogenies of *his* Genes

Because of the availability of so many sequence data in so many different organisms, the *his* biosynthetic genes constitute a precious opportunity to perform phylogenetic studies. The molecular phylogenetic analysis of the *his* genes performed in order to compare the evolutionary relationships among the organisms from which the different genes involved in the pathway have been sequenced, was consistent with the existence of three cellular domains (Fani et al., 1995). Some peculiarities of *his* genes evolution are however worth to mention. The detailed analysis of the phylogenetic trees showed that the nitrogen-fixing α-purple Gram negative bacterium *A. brasilense* was in most

cases nearer to Gram positive bacteria than to its close relatives, the γ purple enterobacteria *E. coli* and *S. typhimurium*. The same result was found for *K. pneumoniae*, when the first one hundred amino acid that are available for its *hisG* gene product were compared; but the latter result was not confirmed by the analysis of the HisF proteins, which placed *K. pneumoniae* closer to *E. coli* and *S. typhimurium*. Whether the peculiar position of this two nitrogen-fixing eubacteria reflects an ancient lateral gene transfer event is not clear yet. Another interesting feature of *his* genes evolution concerns the trees depicting the phylogenies of the *hisC* and *hisI* gene products that point towards the evolutionary proximity of the low GC Gram positive branch (*B. subtilis* and *L. lactis*) to the Archaea rather than to the other bacteria (Fani et al., 1995). This observation is consistent however with results obtained by the comparison of glutamate dehydrogenase (Benachenhou-Lafha et al., 1993), heat shock proteins (Gupta and Golding, 1993; Gupta and Singh, 1994), glutamine synthetase (Brown et al., 1994; Kumada et al., 1993; Tiboni et al., 1993), and carbamoyl synthetases (Lazcano, Puente and Gogarten, unpublished results), all of which have been interpreted as indicating an early massive lateral gene transfer event between the ancestors of both archaea and Gram positive bacteria (Gogarten, 1994). This hypothesis is therefore supported also by the phylogenetic trees constructed by using the histidine gene products.

It is also noteworthy that a systematic study of the paralogous duplications that took place prior to the diversification of the three cell domains, such as those involving *hisA* and *hisF* may provide promising set of data for the construction of deep phylogenies that may shed light on the proper rooting of the universal trees.

CONCLUSIONS

The histidine biosynthetic pathway has been largely used over the past fourty years as a powerful genetic tool to study general aspects of control of cellular metabolism. The pathway is somehow unique for the presence of several reactions quite unusual for a biosynthetic pathway and for structural features of several biosynthetic enzymes. In *S. typhimurium* and in *E. coli* seven enzymes, coded for by eight genes, control ten enzymatic steps: three of them are bifunctional and one ethero-dimeric being composed by the *hisH* and *hisF* gene products.

Regulation of the activity of the first enzyme of the pathway by the end-product histidine and by several other molecules signaling the energy charge of the cell represents a sophisticate example of enzymatic feed-back inhibition. The study of the correlation between the aggregation state of the enzyme and its activity will help clarify the mechanism of allosteric inhibition. The comparison of structural properties of several *his* gene products, each catalyzing more than one biosynthetic steps, provided the opportunity to study the evolution and the mechanisms of catalysis of bifunctional and multifunctional enzymes. HOL-dehydrogenase of *S. typhimurium* represents one of the first examples of bifunctional NAD-linked dehydrogenases. Even though the histidine biosynthetic pathway was almost completely known since a long time, the last blind spot concerning the conversion of 5-PRFAR (BBMIII) to IGP and AICAR (PRAIC or ZMP) has been elucitated only recently. The recent finding that in *Enterobacteria* the *hisH* and *hisF* gene products, in analogy to other glutamine amidotransferases, associate to form the dimeric active IGP synthase holoenzyme catalyzing this step will help to investigate the mechanisms of catalysis and the functional interactions of monomers in multimeric enzymes.

The purine and histidine biosynthetic pathways are connected *via* the "AICAR" cycle. The AICAR is converted either to IMP, in the presence of adequate folate levels, or to the unusual nucleotide ZTP, under metabolic conditions inducing a folate starvation. ZTP is a proposed alarmone signaling C-1-folate deficiency and mediating a physiological response to folate stress. Although inhibition of folate metabolism actually leads to alterations of a number of intracellular processes, by inducing metabolic transcriptional polarity and by affecting mRNA decay and processing, or by triggering other phoenomena, as sporulation in *B. subtilis*, the involvement of ZTP in these processes is only speculative and further studies will be required to demonstrate the existence of a folate stress regulon.

Overproduction of AICAR has been previously invoked to account for the multiple phenotypic changes resulting from the constitutive expression of *hisH* and *hisF* gene products. However, the evidence that the pleiotropic response still occurs in strains devoid of AICAR as a consequence of interruptiong the carbon flow through the histidine and the purine pathways has ruled out this hypothesis. Intergenic suppression analysis has allowed to speculate that several aspects of the pleiotropic response (e.g. filamentation) might be due to the interference of elevated expression of *hisH* and *hisF* with the synthesis of cell wall. This

genetic approach has revealed its importance in identifying novel loci on the *S. typhimurium* chromosome which control synthesis of cell wall and to establish a hierarchy between them by observing epistatic effects among suppressors.

-organization of *his* genes in other species.

The investigation of the mechanisms regulating the expression of the *his* operon in *S. typhimurium* and *E. coli* has help to clarify general aspects of control of gene expression in prokaryotes. Although transcription attenuation was demonstrated to regulate *his* expression and was formally theorized in its essence ~~of codon-specific translational control of transcription~~ over ~~twenty~~ fifteen years ago, several mechanistic aspects concerning the mechanism regulating the coupling of transcription and translation in the leader region have been elucitated only recently.

The availability of a large collection of polar mutations scattered along the different cistrons of the operon has provided the opportunity to study the mechanisms responsible for polarity. These studies have led to envisage the existence of a general mechanism regulating transcription in response to the rate of protein synthesis, operating at the level of intracistronic Rho-dependent terminators. Understanding the structure of the transcription elongation complex, the nature of the *cis* signals on RNA and DNA modulating its activity and the role of ribosomes in controlling the elongation step of transcription represent relevant subjects of this research field.

Post-transcriptional mechanisms have been more recently shown to regulate the expression of bacterial genes. The existence of a specific pattern of decay of the *his* native transcript has been evidenced. The decay process, which is dependent on RNase E, has a net 5' to 3' directionality and originates temporarily stabilized species whose 5' ends map immediately upstream to the intercistronic barriers. The role of ribosomes at the intercistronic barriers in this process will probabily help to formulate a general model of decay of polycistronic mRNAs.

In addition to this complex pattern of mRNA decay, whose physiological significance might be to balance the expression of the promoter proximal genes, which are the first to be transcribed, and the distal ones, a processing event generates a considerably stable transcript spanning the distal cistrons. The relevance of the processing event in the cellular economy is not clear. The presence of an internal promoter, evolutionarily conserved between *E. coli* and *S. typhimurium*, driving transcription of the same distal cistrons contained in the processed species, suggests the existence of a selective pressure favouring the differential expression of these cistrons, whose products are also involved in metabolic pathways other than histidine biosynthesis. The mechanism originating this species has been elucidated only in part. The processing event requires the sequential action of RNase E and RNase P and is dependent on the presence of ribosomes initiating translation at the intercistronic barriers located close downstream to the cleavage site. The role of ribosome in this reaction remains elusive as well as the elements responsible for stability of this species. Answering these questions will ~~help to clarify~~ understand several fundamental aspects of post-transcriptional control: the role of ribosomes; the nature of stability determinants; the role of RNase P in processing molecules other than precursors to tRNAs.

~~regulation of *his* in other species.~~

-evolution of genes and phylogeny.

CONCLUSIONS

The availability of information about over than 60 *his* genes from a wide range of (micro)organisms belonging to the three cell domains has permitted a deep analysis of the structure, organisation and evolution of the histidine biosynthetic genes. As reporetd above there are many clues indicating the antiquity of this anabolic pathway. The phylogenetic distribution of the *his* genes and the fact that at least three of them (*hisA*, *hisI* and *hisC*) have been identified in the three cell lineages strongly suggests that this pathway was one of the metabolic activity of the last common ancestor (LCA) and that the entire pathway might have been assembled well before the appearance of this cell. This possibility is strongly supported by the robustness of the trees depicting the evolutionary distances between different *his* genes, which have the same general topology and show similar interdomain relationships (Fani et al 1995).

The synthesis of histidine and of the dipeptide with catalytic activity hystidil-histidine under possible prebiotic conditions suggests that these molecules were present in the primordial soup (early stages of molecular evolution) and that the pathway might have assembled in consequence of the exhaustation of the prebiotic supply of histidine and histidine-containing peptides, which must have imposed an importance pressure favouring those organisms capable of synthesizing imidazole-containing compounds. The analysis and the comparison of the structure of some of the histidine genes (*hisH, hisA* and *hisF*) suggests that paralogous gene duplications have played an important role in shaping the pathway. This is indicated by the evidence of two successive duplications involving an ancestral module which eventually led to the *hisA* and *hisF* genes and their homologues (Fani et al. 1994) and by the possible common origin of the imidazole glycerol-P synthase encoded by the *E. coli hisH* with other G-type glutamine amidotransferases which participate in the biosynthesis of purines, pyrimidines, arginine, tryptophan, and other ancient pathways (Fani et al 1995).

It has also been postulated that the products of these three genes and their homologues are the descendants of genes encoding less specific enzymes supporting the so-called patchwork hypothesis (Ycas, 1974; Jensen, 1976). According to this idea, primitive metabolic routes were mediated by enzymes of low substrate specificity that were eventually recruited into different pathways (Jensen 1976). This hypothesis is also consistent, however, with the possibility that an ancestral pathway may have had a primitive enzyme catalyzing two or more similar reactions on related substrates of the same metabolic route and whose substrate specificity was refined as a result of later duplication events.

After the building of the entire pathway, and supposing that the population of last common ancestor's cells shared the same *his* genes structure and organization, these genes underwent several major rearrangements in the different cellular lineages. Concerning the organization, in those bacteria where the his genes have been identified, at least some of them are clustered in an operon (figure 2). This is particularly true for *hisBd, H, A* (and often *F*) which have always been found to be part of an operon (Figure 2), where they are contiguous and arranged in the same order and which might represent the core of the histidine biosynthetic pathway. This group of genes might have acted as a unitary block in the evolution of the eubacterial pathway. It is also possible that the set of enzymes they code for (or at least the HisF and HisH proteins) are a metabolon, promoting the preferential transfer of an intermediate from one enzyme to a physically adjacent enzyme, and limiting its diffusion into the surrounding melieu..

Although limited information about the organization of the *his* gene in archaea exists, the available data clearly suggest that they are scattered on the chromosome, resembling the situation of lower eukaryotes, where the histidine genes are not part of operons and are localized on different chromosomes (Figure 2). In the latter organisms the coordinate synthesis of two or more enzymatic activities is often achieved by gene fusion (see for eample the *S. cerevisiae HIS4* and *HIS7* genes). But gene fusions have also been occurred in eubacteria (*hisIE, hisB*)

The phylogenetic comparison of the available *his* genes sequences is consistent with the existence of three cellular domains, but this analysis also placed that the nitrogen-fixing α-purple Gram negative bacterium *Azospirillum* and some Archaea near to the low GC Gram positive bacteria including *B. subtilis* and *L. lactis*, which probably reflect massive lateral gene transfer events. Therefore the *his* genes and their products can be used to establish the phylogenetic relationshsips among microoranisms, and may also evidentiate events of lateral gene transfer. Moreover, since paralogous genes are often used for rooting the phylogenetic trees, a systematic study of these paralogous duplications that took place in the early stages of molecular evolution may provide promising sets of data for the construction of deep phylogenies that may shed light on the proper rooting of the universal trees.

ACKNOWLEDGMENTS

We thank the many Colleagues who provided unpublished data and suggestions used in this review. Work performed in the authors' laboratories was supported by grants from the Ministero dell'Università e della Ricerca Scientifica e Tecnologica, the targeted project "Ingegneria genetica" of the Consiglio Nazionale delle Ricerche and by the Human Capital Mobility Program founded by the Commission of the European Communities (grant n° CHRXCT930263)

REFERENCES

- Adams, E., 1954. The enzymatic synthesis of histidine from histidinol. J. Biol. Chem. 209:829-846.
- Adhya, S., and M. Gottesman. 1978. Control of transcription termination. Annu. Rev. Biochem. 47:967-996.
- Adhya, S., and M. Gottesman. 1982. Promoter occlusion: transcription through a promoter may inhibit its activity. Cell 29:939-944.
- Alifano, P., C. B. Bruni, and M. S. Carlomagno. 1994c. Control of mRNA processing and decay in prokaryotes. Genetica 94:157-172.
- Alifano, P., M. S. Ciampi, A. G. Nappo, C. B. Bruni, and M. S. Carlomagno. 1988. *In vivo* analysis of the mechanisms responsible for strong transcriptional polarity in a "sense" mutant within an intercistronic region. Cell 55:351-360.
- Alifano, P., C. Piscitelli, V. Blasi, F. Rivellini, A. G. Nappo, C. B. Bruni, and M. S. Carlomagno. 1992. Processing of a polycistronic mRNA requires a 5' *cis* element and active translation. Mol. Microbiol. 6:787-798.
- Alifano, P., F. Rivellini, D. Limauro, C. B. Bruni, and M. S. Carlomagno. 1991. A consensus motif common to all Rho-dependent prokaryotic transcription terminators. Cell 64:553-563.
- Alifano, P., F. Rivellini, A. G. Nappo, C. B. Bruni, and M. S. Carlomagno. 1994a. Alternative patterns of *his* operon transcription and mRNA processing generated by metabolic perturbation. Gene 146:15-21.
- Alifano, P., F. Rivellini, C. Piscitelli, C. M. Arraiano, C. B. Bruni, and M. S. Carlomagno. 1994b. Ribonuclease E provides substrates for ribonuclease P-dependent processing of a polycistronic mRNA. Genes Dev. 8:3021-3031.
- Altboum, Z., S. Gottlieb, G.A. Lebens, I. Polacheck, and E. Segal. 1990. Isolation of the *Candida albicans* histidinol dehydrogenase (*HIS4*) gene and characterization of a histidine auxotroph. J. Bacteriol. 172:3898-3904.
- Altman, S. 1990. Ribonuclease P. J. Biol. Chem. 265:20053-20056.
- Ames, B. N., and B. Garry. 1959. Coordinate repression of the synthesis of four histidine biosynthetic enzymes by histidine. Proc. Natl. Acad. Sci. USA 45:1453-1461.
- Ames, B. N., B. Garry, and L. A. Herzenberg. 1960. The genetic control of the enzymes of histidine biosynthesis in *Salmonella typhimurium*. J. Gen. Microbiol. 22:369-378.
- Ames, B. N., R. F. Goldberger, P. E. Hartman, R. G. Martin, and J. R. Roth. 1967. The histidine operon, p. 272-287. *In* V. V. Koningsberger and L. Bosch (ed.), Regulation of nucleic acid and protein biosynthesis. Elsevier Publishing Co., Amsterdam.
- Ames, B. N., and P. E. Hartman. 1963. The histidine operon. Cold Spring Harbor Symp. Quant. Biol. 28:349-356.
- Ames, B. N., P. E. Hartman, and F. Jacob. 1963. Chromosomal alterations affecting the regulation of histidine biosynthetic enzymes in *Salmonella*. J. Mol. Biol. 7:23-42.
- Ames, B. N., R. G. Martin, and B. Garry. 1961. The first step of histidine biosynthesis. J. Biol. Chem. 236:2019-2026.
- Anagnostopulos, C., P.J. Piggot, and J. A. Hoch. 1993. The genetic map of *Bacillus subtilis*, p. 425-462. *In* A. L. Sonenshein, J. A. Hoch and R. Losick (ed.), *Bacillus subtilis* and other gram-positive bacteria. American Society for Microbiology. Washington, D.C.
- Anton, D. N. 1968. Histidine regulatory mutants in *Salmonella typhimurium*. V. Two new classes of histidine regulatory mutants. J. Mol. Biol. 33:533-546.

Anton, D. N. 1978. Genetic control of defective cell shape and osmotic-sensitivity in a mutant of *Salmonella typhimurium*. Mol. Gen. Genet. 160:277-286.

Anton, D. N. 1979. Positive selection of mutants with cell envelope defects of a *Salmonella typhimurium* strain hypersensitive to the products of genes *hisF* and *hisH*. J. Bacteriol. 137:1271-1281.

Arndt, K. T., C. Styles, and G. R. Fink. 1987. Multiple global regulators control *HIS4* transcription in yeast. Science 237:874-880.

Artz, S. W., and J. R. Broach. 1975. Histidine regulation in *Salmonella typhimurium*: an activator-attenuator model of gene regulation. Proc. Natl. Acad. Sci. USA 72:3453-3457.

Artz, S. W., and D. Holzschu. 1983. Histidine biosynthesis and its regulation, p. 379-404. *In* K. M. Herrmann and R. L. Somerville (ed.), Amino acids: biosynthesis and genetic regulation. Addison-Wesley Publishing Co., Reading, Mass.

Atkins, J. F., and J. C. Loper. 1970. Transcription initiation in the histidine operon of *Salmonella typhimurium*. Proc. Natl. Acad. Sci. USA 65:925-932.

Barnes, W. 1978. DNA sequence from the histidine operon control region: seven histidine codons in a row. Proc. Natl. Acad. Sci. USA 75:4281-4285.

Bateman, E., and M. R. Paule. 1988. Promoter occlusion during ribosomal RNA transcription. Cell 54:985-992.

Bauerle, R. H., and P. Margolin. 1967. Evidence for two sites for initiation of gene expression in the tryptophan operon of *Salmonella typhimurium*. J. Mol. Biol. 26:423-436.

Bear, D. G., P. S. Hicks, K. W. Escudero, C. L. Andrews, J. A. McSwiggen, and P. H. von Hippel. 1988. Interaction of *Escherichia coli* transcription termination factor Rho with RNA. II. Electron microscopy and nuclease protection experiments. J. Mol. Biol. 199:623-635.

Beckler, G. S., and J.N. Reeve. 1986. Conservation of primary structure in the *hisI* gene of the archaebacterium *Methanococcus vannielii*, the eubacterium *Escherichia coli* and the eucaryote *Saccharomyces cerevisiae*. Mol. Gen. Genet. 204:133-140.

Belasco, J. G., and C. F. Higgins. 1988. Mechanisms of mRNA decay in bacteria: a perspective. Gene 72:15-23.

Bell, R. M., S. M. Parsons, S. A. Dubravac, A. G. Redfield, and D. E. Koshland, Jr. 1974. Characterization of slowly interconvertible states of phosphoribosyladenosine triphosphate synthetase dependent on temperature, substrates and histidine. J. Biol. Chem. 249:4110-4118.

Benachenhou-Lafha, N., P. Forterre, and B. Labedan. 1993. Evolution of glutamate dehydrogenase genes: evidence for two paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life. J. Mol. Evol. 36:335-346.

Bender, M. L., R. J. Ergeron, and M. Komiyama. 1984. The bio-organic chemistry of enzymatic catalysis, p. 116-129. Wiley, New York.

Berberich, M. A., P. Venetianer, and R. F. Goldberger. 1966. Alternative modes of derepression of the histidine operon observed in *Salmonella typhimurium*. J. Biol. Chem. 241:4426-4433.

Blasi, F., and C. B. Bruni. 1981. Regulation of the histidine operon: translation controlled transcription termination (a mechanism common to several biosynthetic operons). Curr. Top. Cell. Regul. 19:1-45.

Bochner, B. R., and B. N. Ames. 1982. ZTP (5-amino 4-imidazole carboxamide riboside 5'-triphosphate): a proposed alarmone for 10-formyl-tetrahydrofolate deficiency. Cell 29:929-937.

Bouvet, P., and J. G. Belasco. 1992 Control of RNase E-mediated RNA degradation by 5'-terminal base pairing in *E. coli*. Nature 360:488-491.

Brady, D. R., and L. L. Houston. 1973. Some properties of the catalytic sites of imidazoleglycerolephosphate dehydratase-histidinol posphate

phosphatase, a bifunctional enzyme from *Salmonella typhimurium*. J. Biol. Chem. 248:2588-2592.

Brenner, M., and B. N. Ames. 1971. The histidine operon and its regulation, p. 349-387. *In* H.J. Vogel (ed.), Metabolic pathways, vol. 5: Metabolic regulation. Academic Press,Inc., New York.

Brown, G. M., and J. M. Williamson. 1987. Biosynthesis of folic acid, riboflavin, thiamine, and pantothenic acid, p. 521-538. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium:* cellular and molecular biology, vol. 1. American Society for Microbiology, Washington, D.C.

Brown, J. R., Y. Masuchi, F. T. Robb, and W. F. Doolittle. 1994. Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. J. Mol. Evol. 38:566-576.

Bruni, C. B., M. S. Carlomagno, S. Formisano, and G. Paolella. 1986. Primary and secondary structural homologies between the *HIS4* gene product of *Saccharomyces cerevisiae* and the *hisIE* and *hisD* gene products of *Escherichia coli* and *Salmonella typhimurium*. Mol. Gen. Genet. 203:389-396.

Bürger, E., and H. Görisch. 1981a. Patterns of product inhibition of a bifunctional dehydrogenase; L-histidinol : NAD^+ oxidoreductase. Eur. J. Biochem. 116:137-142.

Bürger, E., and H. Görisch. 1981b. Evidence for an essential lysine at the active site of L-histidinol : NAD^+ oxidoreductase; a bifunctional dehydrogenase. Eur. J. Biochem. 118:125-130.

Bürger, E., H. Görisch, and F. Lingens. 1979. The catalytically active form of histidinol dehydrogenase from *Salmonella typhimurium*. Biochem. J. 181:771-774.

Carere, A., S. Russi, M. Bignami, and G. Sermonti. 1973. An operon for histidine biosynthesis in *Streptomyces coelicolor*. -I. Genetic evidence. Mol. Gen. Genet. 123:213-224.

Carlomagno, M. S., F. Blasi, and C. B. Bruni. 1983. Gene organization in the distal part of the *Salmonella typhimurium* histidine operon and determination and sequence of the operon transcription terminator. Mol. Gen. Genet. 191:413-420.

Carlomagno, M. S., L. Chiariotti, P. Alifano, A. G. Nappo, and C. B. Bruni. 1988. Structure and function of the *Salmonella typhimurium* and *Escherichia coli* K-12 histidine operon. J. Mol. Biol. 203:585-606.

Carlomagno, M. S., A. Riccio, and C. B. Bruni. 1985. Convergently functional, Rho-independent terminator in *Salmonella typhimurium*. J. Bacteriol. 163:362-368.

Casadesùs, J., and J. R. Roth. 1989. Absence of insertions among spontaneous mutants of *Salmonella typhimurium*. Mol. Gen. Genet. 216:210-216.

Cashel, M., and K. E. Rudd. 1987. The stringent response, p. 1410-1438. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium:* cellular and molecular biology, vol. 2. American Society for Microbiology, Washington, D.C.

Chan, C. L., and R. Landick. 1989. The *Salmonella typhimurium his* operon leader region contains an RNA hairpin-dependent transcription pause site. Mechanistic implications of the effect on pausing of altered RNA hairpins. J. Biol. Chem. 264:20796-20804.

Chan, C. L., and R. Landick. 1993. Dissection of the *his* leader pause site by base substitution reveals a multipartite signal that includes a pause RNA hairpin. J. Mol. Biol. 233:25-42.

Chen, C.-Y. A., and J. P. Richardson. 1987. Sequence elements essential for Rho-dependent transcription termination at λtR1. J. Biol. Chem. 262:11292-11299.

Chiariotti, L., P. Alifano, M. S. Carlomagno, and C. B. Bruni. 1986. Nucleotide sequence of the *Escherichia coli hisD* gene and of the *Escherichia coli* and *Salmonella typhimurium hisIE* region. Mol. Gen. Genet. 203:382-388.

Chopin, . 1993.

Chumley, F. G., and J. R. Roth. 1981. Genetic fusions that place the lactose genes under histidine operon control. J. Mol. Biol. 145:697-712.

Ciampi, M. S., P. Alifano, A. G. Nappo, C. B. Bruni, and M. S. Carlomagno. 1989. Features of the Rho-dependent transcription termination polar element within the *hisG* cistron of *Salmonella typhimurium*. J. Bacteriol. 171:4472-4478.

Ciampi, M. S., and J. R. Roth. 1988. Polarity effects in the *hisG* gene of Salmonella require a site within the coding sequence. Genetics 118:193-202.

Cole, S. T., and N. Honoré. 1989. Transcription of the *sulA-ompA* region of *Escherichia coli* during the SOS response and the role of an antisense RNA molecule. Mol. Microbiol. 3:715-722.

Conover, R. K., and W. F. Doolittle. 1990. Characterization of a gene involved in histidine biosynthesis in *Halobacterium (Haloferax) volcanii*: isolation and rapid mapping by transformation of an auxotroph with cosmid DNA. J. Bacteriol. 172:3244-3249.

Crawford, I. P., and R. Milkman. 1991. Orthologous and paralogous divergence, reticulate evolution, and lateral gene transfer in bacterial *trp* genes, p. 77-95. *In* R. K. Selander, A.G. Clark and T. S. Whittam (ed.), Evolution at the Molecular Level. Sinauer Press, Publishers, Sunderland.

Cue, D., G. S. Bekler, J. N. Reeve, and J. Konisky. 1985. Structure and sequence divergence of two archaebacterial genes. Proc. Natl. Acad. Sci. USA 82:4207-4211.

Dall-Larsen, T. 1988a. Regulation of the first step of histidine biosynthesis in *Escherichia coli*. Int. J. Biochem. 20:231-235.

Dall-Larsen, T. 1988b. Stopped flow kinetic studies of adenosine triphosphate phosphoribosyl transferase, the first enzyme in the histidine biosynthesis of *Escherichia coli*. Int. J. Biochem. 20:811-815.

Dall-Larsen, T., and L. Klungsøyr. 1976. The binding of specific ligands to adenosine-triphosphate phosphoribosyltransferase. Eur. J. Biochem. 69:195-201.

Davidson, J. N., K. C. Chen, R. S. Jamison, L. A. Musmanno, and C. B. Kern. 1993. The evolutionary history of the first three enzymes in pyrimidine biosynthesis. BioEssays 15:157-164.

Davis, L., and L. S. Williams. 1982. Characterization of a cold-sensitive *hisW* mutant of *Salmonella typhimurium*. J. Bacteriol. 151:867-878.

de Crombrugghe, B., S. Adhya, M. Gottesman, and I. Pastan. 1973. Effects of rho on transcription of bacterial operons. Nature (London) New Biol. 241:260-264.

Delorme, C., S. D. Ehrlich, and P. Renault. 1992. Histidine biosynthesis genes in *Lactococcus lactis* subsp. *lactis*. J. Bacteriol. 174:6571-6579.

Di Nocera, P. P., F. Blasi, R. Di Lauro, R. Frunzio, and C. B. Bruni. 1978. Nucleotide sequence of the attenuator region of the histidine operon of *Escherichia coli* K-12. Proc. Natl. Acad. Sci. USA 75:4276-4280.

Donachie, W. D., and A. C. Robinson. 1987. Cell division: parameter values and the process, p. 1578-1593. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.),

Escherichia coli and *Salmonella typhimurium:* cellular and molecular biology, vol. 2. American Society for Microbiology, Washington, D.C.

Donahue, T. F., P. J. Farabaugh, and G. R. Fink. 1982. The nucleotide sequence of the *HIS4* region of yeast. Gene 18:47-59.

Eccleston, E. D., M. L. Thayer, and S. Kirkwood. 1979. Mechanisms of action of histidinol dehydrogenase and UDP-Glc dehydrogenase. Evidence that the half-reactions proceed on separate subunits. J. Biol. Chem. 254:11399-11404.

Ely, B., and Z. Ciesla. 1974. Internal promoter P2 of the histidine operon of *Salmonella typhimurium*. J. Bacteriol. 120:984-986.

Fani, R., P. Alifano, G. Allotta, M. Bazzicalupo, M. S. Carlomagno, E. Gallori, F. Rivellini, and M. Polsinelli. 1993. The histidine operon of *Azospirillum brasilense*: organization, nucleotide sequence and functional analysis. Res. Microbiol. 144:187-200.

Fani, R., M. Bazzicalupo, G. Damiani, A. Bianchi, C. Schipani, V. Sgaramella, and M. Polsinelli. 1989. Cloning of the histidine genes of *Azospirillum brasilense*: organization of the *ABFH* gene cluster and nucleotide sequence of the *hisB* gene. Mol. Gen. Genet. 216:224-229.

Fani, R., P. Liò, I. Chiarelli, and M. Bazzicalupo. 1994. The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the *hisA* and *hisF* genes. J. Mol. Evol. 38:489-495.

Fani, R., P. Liò, and A. Lazcano. Submitted for publication.

Fani, R., et al. Unpublished data a

Fani, R., et al. Unpublished data b.

Faxén, M., and L. A. Isaksson. 1994. Functional interactions between translation, transcription and ppGpp in growing *Escherichia coli*. Biochim. Biophys. Acta 1219:425-434.

Fersht, A. 1977. Enzyme structure and mechanism, p.325-329. Freeman, San Francisco.

Figueroa, N., N. Wills, and L. Bossi. 1991. Common sequence determinants of the response of a prokaryotic promoter to DNA bending and supercoiling. EMBO J. 10:941-949.

Fink, G. R., T. Klopotowski, and B. N. Ames. 1967. Histidine regulatory mutants in *Salmonella typhimurium*. IV. A positive selection for polar histidine-requiring mutants from histidine operator constitutive mutants. J. Mol. Biol. 30:81-95.

Fink, G. R., and R. G. Martin. 1967. Translation and polarity in the histidine operon. II. Polarity in the histidine operon. J. Mol. Biol. 30:97-107.

Flores, A., M., and J. Casadesùs. Personal comunication.

Flores, A., M. Fox, and J. Casadesùs. 1993. The pleiotropic effects of *his* overexpression in *Salmonella typhimurium* do not involve AICAR-induced mutagenesis. Mol. Gen. Genet. 240:360-364.

Forster, A.C., and S. Altman. 1990. External guide sequences for an RNA enzyme. Science 249:783-786.

Fox, M., N. Frandsen, and R. D'Ari. 1993. AICAR is not an endogenous mutagen in *Escherichia coli*. Mol. Gen. Genet. 240:355-359.

Frandsen, N., and R. D'Ari. 1993. Excess histidine enzymes cause AICAR-independent filamentation in *Escherichia coli*. Mol. Gen. Genet. 240:348-354.

Franklin, N. C., and S. E. Luria. 1961. Transduction by bacteriophage P1 and the properties of the lac genetic region in *E. coli* and *S. depenteriae*. Virology 15:299-311.

Freedman, R., and P. Schimmel. 1981. *In vitro* transcription of the histidine operon. J. Biol. Chem. 256:10747-10750.

Freese, E., J. E. Heinze, and E. M. Galliers. 1979. Partial purine deprivation causes sporulation of *Bacillus subtilis* in the presence of excess ammonia, glucose and phosphate. J. Gen. Microbiol. 115:193-205.

Fridman et al.

Frunzio, R., C. B. Bruni, and F. Blasi. 1981. In vivo and in vitro detection of the leader RNA of the histidine operon of *Escherichia coli* K-12. Proc. Natl. Acad. Sci. USA 78:2767-2771.

Galloway, J. L., and T. Platt. 1988. Signals sufficient for Rho-dependent transcription termination at *trp* t' span a region centered 60 base pairs upstream of the earliest 3' endpoint. J. Biol. Chem. 263:1761-1767.

Geiger, J. R., and J. F. Speyer. 1977. A conditional antimutator in *E. coli*. Mol. Gen. Genet. 153:87-97.

Gibert, I., and J. Casadesùs. 1990. *sulA*-independent division inhibition in His-constitutive strains of *Salmonella typhimurium*. FEMS Microbiol. Lett. 69:205-210.

Gogarten, J. P. 1994. Which is the most conserved group of proteins? Homology, orthology, paralogy and the fusion of independent lineages. J. Mol. Evol. 39:541-543.

Görisch, H. 1979. Steady-state investigation of the mechanism of histidinol dehydrogenase. Biochem. J. 181:153-157.

Görisch, H., and W. Hölke. 1985. Binding of histidinal to histidinol dehydrogenase. Eur. J. Biochem. 150:305-308.

Goto, Y., H. Zalkin, P. S. Keim, and R. L. Heinrikson. 1976. J. Biol. Chem. 251:941-949.

Granick, S. 1965. Evolution of heme and chlorophyll, p. 67-88. *In* V. Bryson and H. J. Vogel (ed.), Evolving Genes and Proteins. Academic Press, New York.

Green, J. M., and B. P. Nichols. 1991. J. Biol. Chem. 266:12971-12975.

Grisolia, V., A. Riccio, and C. B. Bruni. 1983. Structure and function of the internal promoter (*hisBp*) of the *Escherichia coli* K-12 histidine operon. J. Bacteriol. 155:1288-1296.

Grubmeyer, C. T., and W. R. Gray. 1986. A cysteine residue (Cysteine-116) in the histidinol binding site of histidinol dehydrogenase. Biochemistry 25:4778-4784.

Grubmeyer, C. T., M. Skiadopoulos, and A. E. Senior. 1989. L-histidinol dehydrogenase, a Zn^{2+}-metalloenzyme. Arch. Biochem. Biophys. 272:311-317.

Gupta, R. S., and Golding. 1993.

Gupta, R. S., and B. Singh. 1994. Phylogenetic analysis of 70 kD heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. Curr. Biology 4:1104-1114.

Haas, F., M. B. Mitchell, B. N. Ames, and H. K. Mitchell. 1952. A series of histidineless mutants of *Neurospora crassa*. Genetics 37:217-.

Harley, C. B., and R. P. Reynolds. 1987. Analysis of *E. coli* promoter sequences. Nucleic Acids Res. 15:2343-2361.

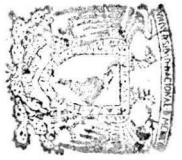
Harris, J. I., and M. Waters. 1976. , p. 1-49. *In* P. D. Boyer (ed.), The enzymes, vol. 13, 3rd Ed., Academic Press, New York.

Hartman, P. E. 1956. Linked loci in the control of consecutive steps in the primary pathway of histidine synthesis in *Salmonella typhimurium*. Carnegie Inst. Wash. Publ. 612:35-62.

Hartman, P. E., Z. Hartman, and D. Serman. 1960a. Complementation-mapping by abortive transduction of histidine-requiring *Salmonella* mutants. J. Gen. Microbiol. 22:354-368.

Hartman, P. E., Z. Hartman, R. C. Stahl, and B. N. Ames. 1971. Classification and mapping of spontaneous and induced mutations in the histidine operon of *Salmonella*. Adv. Genet. 16:1-34.

Hartman, P. E., J. C. Loper and D. Serman. 1960b. Fine structure mapping by complete transduction between histidine-requiring *Salmonella* mutants. J. Gen. Microbiol. 22:323-353.



Hartman, P. E., and S. R. Suskind. 1965. Gene action. Prentice-Hall Inc. Englewood Cliffs, New Jersey.

Heinze, J. E., T. Milani, K. E. Rich, and E. Freese. 1978. Induction of sporulation by inhibitory purines and related compounds. Biochim. Biophys. Acta 521:16-26.

Henner, D. J., L. Band, G. Flaggs, and E. Chen. 1986. The organization and nucleotide sequence of the *Bacillus subtilis hisH, tyrA* and *aroE* genes. Gene 49:147-152.

Hinnebusch, A. G. 1988. Mechanisms of gene regulation in the general control of amino acid biosynthesis in *Saccharomyces cerevisiae*. Microbiol. Rev. 52:248-273.

Hinnebusch, A. G. 1990. Transcriptional and translational regulation of gene expression in the general control of amino acid biosynthesis in *Saccharomyces cerevisiae*. Prog. Nucleic Acid Res. Mol. Biol. 38:195-240.

Hinshelwood, S., and N. G. Stoker. 1992. Cloning of mycobacterial histidine synthesis genes by complementation of a *Mycobacterium smegmatis* auxotroph. Mol. Microbiol. 6:2887-2895.

Hoppe, I., H. N. Johnston, D. Biek, and J. R. Roth. 1979. A refined map of the *hisG* gene of *Salmonella typhimurium*. Genetics 92:17-26.

Hopwood, D. A. 1986.

Hopwood, D. A., H. M. Kieser, and T. Kieser. 1993. The chromosome map of *Streptomyces coelicolor* A3(2), p. 497-506. *In* A. L. Sonenshein, J. A. Hoch and R. Losick (ed.), *Bacillus subtilis* and other gram-positive bacteria. American Society for Microbiology. Washington, D.C.

Horowitz, N. J. 1945. On the evolution of biochemical synthesis. Proc. Natl. Acad. Sci. USA 31:153-157.

Horowitz, N. J. 1965. The evolution of biochemical synthesis-retrospect and prospect, p. 15-23. *In* V. Bryson and H. J. Vogel (ed.), Evolving Genes and Proteins. Academic Press, Inc., New York.

Hsu, L. C., M. Okamoto, and E. E. Snell. 1989. L-Histidinol phosphate aminotransferase from *Salmonella typhimurium*. Kinetic behavior and sequence at the pyridoxal-P binding site. Biochimie 71:477-489.

Hubbard, J. S., and E. R. Stadtman. 1967. Regulation of glutamine synthetase. Interactions of inhibitors for *Bacillus licheniformis* glutamine synthetase. J. Bacteriol. 94:1016-1024.

Imamoto, F. 1970. Evidence for premature termination of transcription of the tryptophan operon in polarity mutants of *Escherichia coli*. Nature 228:232-235.

Imamoto, F., J. Ito, and C. Yanofsky. 1966. Polarity studies with the tryptophan operon. Cold Spring Harbor Symp. Quant. Biol. 31:235-.

Imamoto, F., and Y. Kano. 1971. Inhibition of transcription of the tryptophan operon in Escherichia coli by a block in initiation of translation. Nature 232:169-173.

Jackson, E. N., and C. Yanofsky. 1972. Internal promoter of the tryptophan operon of *Escherichia coli* is located in a structural gene. J. Mol. Biol. 69:307-315.

Jackson, E. N., and C. Yanofsky. 1973. The region between the operator and first structural gene of the tryptophan operon of *Escherichia coli* may have a regulatory function. J. Mol. Biol. 76:89-101.

Jacob, F., and J. Monod. 1961a. Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3:318-356.

Jacob, F., and J. Monod. 1961b. On the regulation of gene activity. Cold Spring Harbor Symp. Quant. Biol. 26:193-211.

Jensen, R. A. 1976. Enzyme recruitment in evolution of new function. Annu. Rev. Microbiol. 30:409-425.

Jink-Robertson, S., and M. Nomura. 1987. Ribosomes and tRNA, p. 1358-1385. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*

typhimurium: cellular and molecular biology, vol. 2. American Society for Microbiology, Washington, D.C.

Johnston, H. M., W. M. Barnes, F. G. Chumley, L. Bossi, J. R. Roth. 1980. Model for regulation of the histidine operon of *Salmonella*. Proc. Natl. Acad. Sci. USA 77:508-512.

Johnston, H. M., and J. R. Roth. 1979. Histidine mutants requiring adenine: selection of mutants with reduced *hisG* expression in *Salmonella typhimurium*. Genetics 92:1-15.

Joseph, E., A. Danchin, and A. Ullmann. 1978. Modulation of the lactose operon mRNA turnover by inhibitors of dihydrofolate reductase. Biochem. Biophys. Res. Commun. 84:769-776.

Jovanovich, et al. 1994.

Kaplan, J. B., and B. P. Nichols. 1983. Nucleotide sequence of *Escherichia coli pabA* and its evolutionary relationships to *trp(G)D*. J. Mol. Biol. 168:451-468.

Kasai, T. 1974. Regulation of the expression of the histidine operon in *Salmonella typhimurium*. Nature 249:523-527.

Kirschner, K., and H. Bisswanger. 1976. Annu. Rev. Biochem. 45:143-166.

Klem, T. J., and V. J. Davisson. 1993. Imidazole glycerol phosphate synthase: the glutamine amidotransferase in histidine biosynthesis. Biochemistry 32:5177-5186.

Klungsøyr, L. 1971. Conformational changes and aggregation in phosphoribosyladenosine triphosphate synthetase. Ligand effects on hydrogen exchange and hydrophobic probe uptake. Biochemistry 10:4875-4880.

Klungsøyr, L., J. H. Hageman, L. Fall, and D. E. Atkinson. 1968. Interaction between energy charge and product feedback in the regulation of biosynthetic enzymes. Aspartokinase, phosphoribosyladenosine triphosphate synthetase, and phosphoribosyl pyrophosphate synthetase. Biochemistry 7:4035-4040.

Klungsøyr, L., and H. Kryvi. 1971. Sedimentation behaviour of phosphoribosyladenosine triphosphate synthetase. Effect of substrate and modifiers. Biochim. Biophys. Acta 227:327-336.

Kolter, R., and C. Yanofsky. 1982. Attenuation in amino acid biosynthetic operons. Annu. Rev. Genet. 16:113-134.

Kuenzler, M., T. Balmelli, C. M. Egli, G. Paravicini, and G. H. Braus. 1993. Cloning, primary structure, and regulation of the *HIS7* gene encoding a bifunctional glutamine amidotransferase:cyclase from *Saccharomyces cerevisiae*. J. Bacteriol. 175:5548-5558.

Kumada, Y., D. R. Benson, D. Hillemann, T. J. Hosted, D. A. Rochford, C. J. Thompson, W. Wohlleben, and Y. Tateno. 1993. Evolution of the glutamine synthase gene, one of the oldest and functioning genes. Proc. Natl. Acad. Sci. USA 90:3009-3013.

Landick, R. 1987. The role of the paused transcription complex in *trp* operon attenuation, p. 441-444. *In* W. S. Reznikoff, R. R. Burgess, J. E. Dahlberg, C. A. Gross, M. T. Record, Jr., and M. P. Wickens (ed.), RNA polymerase and the regulation of transcription. Elsevier, New York.

Landick, R., J. Carey, and C. Yanofsky. 1987. Proc. Natl. Acad. Sci. USA 84:1507-1511.

Landick, R. and C. Yanofsky. 1984. Stability of an RNA secondary structure affects *in vitro* transcription pausing in the *trp* operon leader region. J. Biol. Chem. 259:11550-11555.

Landick, R. and C. Yanofsky. 1987. Isolation and structural analysis of the *Escherichia coli trp* leader paused transcription complex. J. Mol. Biol. 196:363-377.

Lazcano, A., G. E. Fox, and J. Orò. 1992. Life before DNA: the origin and evoution of early Archean cells, p. 237-339. *In* R. P. Mortlock (ed.), The Evolution of Metabolic Function. CRC Press, Boca Raton FL.

Lazcano, A., Puente, and J. P. Gogarten. Unpublished data.

Lee, D. N., and R. Landick. 1992. Structure of RNA and DNA chains in paused transcription complexes containing *Escherichia coli* RNA polymerase. J. Mol. Biol. 228:759-777.

Lee, D. N., L. Phung, J. Stewart, and R. Landick. 1990. Transcription pausing by *Escherichia coli* RNA polymerase is modulated by downstream DNA sequences. J. Biol. Chem. 265:15145-15153.

Legerton, T. L., and C. Yanofsky. 1985. Cloning and characterization of the multifunctional *his-3* gene of *Neurospora crassa*. Gene 39:129-140.

Leupold, U. 1958. Studies on recombination in *Schizosaccharomyces pombe*. Cold Spring Harbor Symp. Quant. Biol. 23:161-.

Limauro, D., A. Avitabile, C. Cappellano, A. M. Puglia, and C. B. Bruni. 1990. Cloning and characterization of the histidine biosynthetic gene cluster of *Streptomyces coelicolor* A3(2). Gene 90:31-41.

Limauro, D., A. Avitabile, A. M. Puglia, and C. B. Bruni. 1992. Further characterization of the histidine gene cluster of *Streptomyces coelicolor* A3(2): nucleotide sequence and transcriptional analysis of *hisD*. Res. Microbiol. 143:683-693.

Loper, J. C. 1961. Enzyme complementation in mixed extracts of mutants from the Salmonella histidine B locus. Proc. Natl. Acad. Sci. USA 47:1440-1450

Loper, J. C., and E. Adams. 1965. Purification and properties of histidinol dehydrogenase from *Salmonella typhimurium*. J. Biol. Chem. 240:788-795.

Malone et al. 1994.

Margolies, M. N., and R. F. Goldberger. 1966. Isolation of the fourth enzyme (isomerase) of histidine biosynthesis from *Salmonella typhimurium*. J. Biol. Chem. 241:3262-3269.

Margolies, M. N., and R. F. Goldberger. 1967. Physical and chemical characterization of the isomerase of histidine biosynthesis in *Salmonella typhimurium*. J. Biol. Chem. 242:256-264.

Martin, R. G. 1963a. The one operon-one messenger theory of transcription. Cold Spring Harbor Symp. Quant. Biol. 28:357-361.

Martin, R. G. 1963b. The first enzyme in histidine biosynthesis: the nature of feedback inhibition by histidine. J. Biol. Chem. 238:257-268.

Martin, R. G., M. A. Berberich, B. N. Ames, W. W. Davis, R. F. Goldberger, and J. D. Yourno. 1971. Enzymes and intermediates of histidine biosynthesis in *Salmonella typhimurium*. Methods Enzymol. 17B:3-44.

Martin, R. G., D. F. Silbert, D. W. E. Smith, and H. J. Whitfield Jr. 1966a. Polarity in the histidine operon. J. Mol. Biol. 21:357-369.

Martin, R. G., and N. Talal. 1968. Translation and polarity in the histidine operon. IV. Relation of polarity to map position in *hisC*. J. Mol. Biol. 36:219-229.

Martin, R. G., H. J. Whitfield Jr., D. B. Berkowitz, and M. J. Voll. 1966b. A molecular model of the phenomenon of polarity. Cold Spring Harbor Symp. Quant. Biol. 31:215-220.

Mathews, C. K. 1993. The cell- bag of enzymes or network of channels? J. Bacteriol. 175:6377-6381.

McSwiggen, J. A., D. G. Bear, and P. H. von Hippel. 1988. Interaction of *Escherichia coli* transcription termination factor Rho with RNA. I. Binding stoichiometries and free energies. J. Mol. Biol. 199:609-622.

Metha, K., T. I. Hale, and P. Christen. 1989. Evolutionary relationships among aminotransferases. Tyrosine aminotransferase, histidinol-

phosphate aminotransferase, and aspartate aminotransferase are homologous proteins. Eur. J. Biochem. 186:249-253.

Milani, T., J. E. Heinze, and E. Freese. 1977. Induction of sporulation in *Bacillus subtilis* by decoyinine or hadacidin. Biochem. Biophys. Res. Commun. 77:1118-1125.

Miller, S. L. 1953. A production of amino acids under possible primitive earth conditions. Science 117:528-529.

Morse, D. E., and M. Guertin. 1971. Regulation of mRNA utilization and degradation by amino acid starvation. Nature 232:165-169.

Morse, D. E., and P. Primakoff. 1970. Relief of polarity in *E. coli* by *"suA"*. Nature 226:28-31.

Morse, D. E., and C. Yanofsky. 1969. Polarity and degradation of mRNA. Nature 224: 329-331.

Mortimer, . 1994.

Mozier, N. M., M. P. Walsh, and J. D. Pearson. 1991. Characterization of a novel zinc-binding site of protein kinase C inhibitor-1. FEBS Lett. 279:14-18.

Mulligan, M. E., D. K. Hawley, R. E. Entriken, and W. R. McClure. 1984. *Escherichia coli* promoter sequences predict *in vitro* RNA polymerase selectivity. Nucleic Acids Res. 12:789-800.

Murray, M. L., and P. E. Hartman. 1972. Overproduction of *hisH* and *hisF* gene products leads to inhibition of cell division in *Salmonella*. Can. J. Microbiol. 18:671-681.

Murray, V. 1987. 5-Amino-4-imidazolecarboxamide is a mutagen in *E. coli*. Mutat. Res. 190:89-94.

Nagai, A., E. Ward, J. Beck, S. Tada, J. Y. Chang, A. Scheidegger, and J. Ryals. 1991. Structural and functional conservation of histidinol dehydrogenase between plants and microbes. Proc. Natl. Acad. Sci. USA 88:4133-4137.

Nègre, D., J.-C. Cortay, P. Donini, and A. J. Cozzone. 1989. Relationship between guanosine tetraphosphate and accuracy of translation in *Salmonella typhimurium*. Biochemistry 28:1814-1819.

Neuhard, J., and P. Nygaard. 1987. Purines and pyrimidines, p. 445-473. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium:* cellular and molecular biology, vol. 1. American Society for Microbiology, Washington, D.C.

Newton, W. A., J. R. Beckwith, D. Zipser, and S. Brenner. 1965. Nonsense mutants and polarity in the *Lac* operon of *Escherichia coli*. J. Mol. Biol. 14:290-296.

Nichols, B. P. et al. 1980.

O'Bryne, C. P., N. Ni Bhriain, and C. J. Dorman. 1992. The DNA supercoiling-sensitive expression of the *Salmonella typhimurium his* operon requires the *his* attenuator and is modulated by anaerobiosis and by osmolarity. Mol. Microbiol. 6:2467-2476.

Oparin, A. I. 1924. Proiskhodenie Zhizni. Moscow: Moscoksky Rabotichii. 7t pp. Transl., 1967, as appendix in Bernal, J.D. The origin of life. Cleveland: World 345 pp

Oparin, A. I. 1936. The origin of life. New York: Dover 270 pp

Oppenheim, and C. Yanofsky. 1980.

Orò, J., B. Basile, S. Cortes, C. Shen, and T. Yamron. 1984. The prebiotic synthesis and catalystic role of imidazoles and other condensing agents. Origins life 14:237-242.

Parsons, S. M., and D. E. Koshland Jr. 1974a. A rapid isolation of phosphoribosyladenosine triphosphate synthetase and comparison to native enzyme. J. Biol. Chem. 249:4104-4109.

Parsons, S. M., and D. E. Koshland Jr. 1974b. Multiple aggregation states of phosphoribosyladenosine triphosphate synthetase. J. Biol. Chem. 249:4119-4126.

Pattee, P. A. 1993. The genetic map of *Staphylococcus aureus*, p. 489-496. *In* A. L. Sonenshein, J. A. Hoch and R. Losick (ed.), *Bacillus subtilis* and other gram-positive bacteria. American Society for Microbiology. Washington, D.C.

Petersen, C. 1992. Control of functional mRNA stability in bacteria: multiple mechanisms of nucleolytic and non-nucleolytic inactivation. Mol. Microbiol. 6:277-282.

Petersen, H. U., A. Danchin, and M. Grunberg-Manago. 1976a. Toward an understanding of the formylation of initiator tRNA methionine in prokaryotic protein synthesis, I. In vitro studies of the 30S and 70S ribosomal-tRNA complex. Biochemistry 15:1357-1362.

Petersen, H. U., A. Danchin, and M. Grunberg-Manago. 1976b. Toward an understanding of the formylation of initiator tRNA methionine in prokaryotic protein synthesis, II. A two-state model for the 70S ribosome. Biochemistry 15:1362-1369.

Petersen, H. U., E. Joseph, A. Ullmann, and A. Danchin. 1978. Formylation of initiator tRNA methionine in procaryotic protein synthesis: in vivo polarity in lactose operon expression. J. Bacteriol. 135:453-459.

Piette, J., H. Nyunoya, C. J. Lusty, R. Cunin, G. Weyens, M. Crabeel, D. Charlier, N. Glansdorf, and A. Pierard. 1984. DNA sequences of the *carA* gene and the control region of *carAB*: tandem promoters, respectively controlled by arginine and the pyrimidines, regulate the synthesis of carbamoyl-phosphate synthetase in *Escherichia coli* K-12. Proc. Natl. Acad. Sci. USA 81:4134-4138.

Pittard, A. J. 1987. Biosynthesis of the aromatic amino acids, p. 368-394. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium:* cellular and molecular biology, vol. 1. American Society for Microbiology, Washington, D.C.

Rechler, M. M., C. B. Bruni, R. G. Martin and W. Terry. 1972. An intercistronic region in the histidine operon of *Salmonella typhimurium*. J. Mol. Biol. 69:427-452.

Riccio, A., C. B. Bruni, M. Rosenberg, M. Gottesman, K. McKenney, and F. Blasi. 1985. Regulation of single and multicopy *his* operons in *Escherichia coli*. J. Bacteriol. 163:1172-1179.

Richardson, J. P. 1991. Preventing the synthesis of unused transcripts by Rho factor. Cell 64:1047-1049.

Rieder, G., and D. Kleiner. 1993. Clarification of the last blind spot in the histidine biosynthesis - function of *hisF* and *hisH* gene products. Bioingeneering 9:30.

Rieder, G., M. J. Merrick, H. Castorph, and D. Kleiner. 1994. Function of *hisF* and *hisH* gene products in histidine biosynthesis. J. Biol. Chem., 269:14386-14390.

Riggs, D. L., and S. W. Artz. 1984. The *hisD-hisC* gene border of the *Salmonella typhimurium* histidine operon. Mol. Gen. Genet. 196:526-529.

Riggs, D. L., R. D. Mueller, H.-S. Kwan, and S. W. Artz. 1986. Promoter domain mediates guanosine tetraphosphate activation of the histidine operon. Proc. Natl. Acad. Sci. USA 83:9333-9337.

Rivellini, F., P. Alifano, C. Piscitelli, V. Blasi, C. B. Bruni, and M. S. Carlomagno. 1991. A cytosine- over guanosine-rich sequence in RNA activates rho-dependent transcription termination. Mol. Microbiol. 5:3049-3054.

Rodriguez, R. L., and R. W. West. 1984. Histidine operon control region of *Klebsiella pneumoniae*: analysis with an *Escherichia coli* promoter-probe plasmid vector. J. Bacteriol. 157:764-771.

Rodriguez, R. L., R. W. West, R. C. Tait, J. M. Jaynes, and K. T. Shanmugam. 1981. Isolation and characterization of the *hisG* and *hisD* genes of *Klebsiella pneumoniae*. Gene 16:317-320.

Rohlman, C. E., and R. G. Matthews. 1990. Role of purine biosynthetic intermediates in response to folate stress in *Escherichia coli*. J. Bacteriol. 172:7200-7210.

Roth, J. R., and B. N. Ames. 1966. Histidine regulatory mutants in *Salmonella typhimurium*. II. Histidine regulatory mutants having altered histidyl-tRNA synthetase. J. Mol. Biol. 22:325-334.

Roth, J. R., D. N. Anton and P. E. Hartman. 1966. Histidine regulatory mutants in *Salmonella typhimurium*. I. Isolation and general properties. J. Mol. Biol. 22:305-323.

Roth, J. R., and P. E. Hartman. 1965. Heterogeneity in P22 transducing particles. Virology 27:297-307.

Rudd, K. E., and R. Menzel. 1987. *his* operons of *Escherichia coli* and *Salmonella typhimurium* are regulated by DNA supercoiling. Proc. Natl. Acad. Sci. USA 84:517-521.

Sabina, R. L., E. W. Holmes, and M. A. Becker. 1984. The enzymatic synthesis of 5-amino-4-imidazolecarboxamide triphosphate. Science 223:1193-1195.

Sabina, R. L., D. Patterson, and E. W. Holmes. 1985. 5-Amino-4-imidazolecarboxamide riboside (Z riboside) metabolism in eukaryotic cells. J. Biol. Chem. 260:6107-6114.

Sampei, G., and K. Mizobuchi. 1989. The organization of *purL* gene encoding 5' phosphoribosyl-formyl-glycinamide amidotransferase of *Escherichia coli*. J. Biol. Chem. 264:21230-21238.

Schendel, F. J., E. Mueller, J. Stubbe,V Shiau, and J. M. Smith. 1989. Formyl-glycinamide ribonucleotide synthetase from *Escherichia coli*: cloning, sequencing, overproduction, isolation and characterization. Biochemistry 28:2459-2471.

Schmid, M. B., and J. R. Roth. 1983. Internal promoters of the *his* operon in *Salmonella typhimurium*. J. Bacteriol. 153:1114-1119.

Schuster, . 1995.

Shand, R. F., P. H. Blum, D. L. Holzschu, M. S. Urdea, and S. W. Artz. 1989a. Mutational analysis of the histidine operon promoter of *Salmonella typhimurium*. J. Bacteriol. 171:6330-6337.

Shand, R. F., P. H. Blum, R. D. Mueller, D. L. Riggs, and S. W. Artz. 1989b. Correlation between histidine operon expression and guanosine 5'-diphosphate-3'-diphosphate levels during amino acid downshift in stringent and relaxed strains of *Salmonella typhimurium*. J. Bacteriol. 171:737-743.

Shen, C., A. Lazcano, and J. Orò. 1990c. The enhancement activities of histidyl-histidine in some prebiotic reactions. J. Mol. Evol. 31:445-452.

Shen, C., T. Mills, and J. Orò. 1990a. Prebiotic synthesis of histidyl-histidine. J. Mol. Evol. 31:175-179.

Shen, C., L. Yang, S. L. Miller, and J. Orò. 1987. Prebiotic synthesis of imidazole-4-acetaldehyde and histidine. Origins life 17:295-305.

Shen, C., L. Yang, S. L. Miller, and J. Orò. 1990b. Prebiotic synthesis of histidine. J. Mol. Evol. 31:167-174.

Sheppard, D. E. 1964. Mutants of *Salmonella typhimurium* resistant to feedback inhibition by L-histidine. Genetics 50:611-623.

Sheridan, R. P., and R. Venkataraghavan. 1992. A systematic search for protein signature sequences. Proteins 14:16-28.

Shioi, J., R. J. Galloway, M. Niwano, R. E. Chinnock, and B. L. Taylor. 1982. J. Biol. Chem. 257:7969-7975.

Silbert, D. F., G. R. Fink, and B. N. Ames. 1966. Histidine regulatory mutants in *Salmonella typhimurium*. III. A class of regulatory mutants deficient in tRNA for histidine. J. Mol. Biol. 22:335-347.

Smith, D. W. E., and B. N. Ames. 1964. Intermediates in the early steps of histidine biosynthesis. J. Biol. Chem. 239:1848-1855.

Smith, D. W. E., and B. N. Ames. 1965. Phosphoribosyladenosine monophosphate, an intermediate in histidine biosynthesis. J. Biol. Chem. 240:3056-3063.

Sonenshein, A. L. 1993. Introduction to metabolic pathways, p. 127-132. *In* A. L. Sonenshein, J. A. Hoch and R. Losick (ed.), *Bacillus subtilis* and other gram-positive bacteria. American Society for Microbiology. Washington, D.C.

Srere, P. A. 1987. Complexes of sequential metabolic enzymes. Annu. Rev. Biochem. 56:21-56.

Stanssens, P., E. Remaut, and W. Fiers. 1986. Inefficient translation initiation causes premature transcription termination in the *lacZ* gene. Cell 44:711-718.

Stephens, J. C., S. W. Artz, and B. N. Ames. 1975. Guanosine 5'-diphosphate 3'-diphosphate (ppGpp): positive effector for histidine operon transcription and general signal for amino-acid deficiency. Proc. Natl. Acad. Sci. USA 72:4389-4393.

Stougaard, J., and C. Kennedy. 1988. Regulation of nitrogenase synthesis in histidine auxotrophs of *Klebsiella pneumoniae* with altered levels of adenylate nucleotides. J. Bacteriol. 170:250-257.

Taschner, P. E. M., P. G. Huls, E. Pas, and C. L. Woldringh. 1988. Division behavior and shape changes in isogenic *ftsZ*, *ftsQ*, *ftsA*, *pbpB*, and *ftsE* cell division mutants of *Escherichia coli* during temperature shift experiments. J. Bacteriol. 170:1533-1540.

Tebar, A. R., V. M. Fernandez, R. MartinDelRio, and A. O. Ballesteros. 1973. Studies on the quaternary structure of the first enzyme for histidine biosynthesis. Experientia 29:1477-1479.

Tebar, A. R., V. M. Fernandez, R. MartinDelRio, and A. O. Ballesteros. 1975. Fluorescences studies of phosphoribosyladenosine triphosphate synthetase of *Escherichia coli*. FEBS Letters 50:239-242.

Teng, H., E. Segura, and C. Grubmeyer. 1993. Conserved cysteine residues of histidinol dehydrogenase are not involved in catalysis. Novel chemistry required for enzymatic aldehyde oxidation. J. Biol. Chem. 268:14182-14188.

Threu-Cluot et al.

Tiboni, O., P. Cammarano, and A. M. Sanangelantoni. 1993. Cloning and sequencing of the gene encoding glutamine synthase I from the archaeum *Pyrococcus woesei*: anomalous phylogenies inferred from analysis of archaeal and bacterial glutamine synthase I sequences. J. Bacteriol. 175:2961-2969.

Tiedeman, A. A., J. M. Smith, and H. Zalkin. (1985) Nucleotide sequence of the *guaA* gene encoding GMP synthetase of *Escherichia coli* K12. J. Biol. Chem. 260:8676-8679.

Toone, W. M., K. E. Rudd, and J. D. Friesen. 1992. Mutations causing aminotriazole resistance and temperature sensitivity reside in *gyrB*, which encodes the B subunit of DNA gyrase. J. Bacteriol. 174:5479-5481.

Travers, A. A. 1984. Conserved features of coordinately regulated *E. coli* promoters. Nucleic Acids Res. 12:2605-2618.

Trotta, P. P., L. M. Pinkus, R. H. Haschemeyer, and A. Meister. 1974. J. Biol. Chem. 249:492-499.

Umbarger, H. E. 1956. Evidence for a negative-feedback mechanism in the biosynthesis of isoleucine. Science 123:848-.

Venetianer, P. 1968. Preferential synthesis of the messenger RNA of the histidine operon during histidine starvation. Biochem. Biophys. Res. Commun. 33:959-963.

Venetianer, P. 1969. Level of messenger RNA transcribed from the histidine operon in repressed, derepressed, and histidine-starved *Salmonella typhimurium*. J. Mol. Biol. 45:375-384.

Verde, P., R. Frunzio, P. P. Di Nocera, F. Blasi, and C. B. Bruni. 1981. Identification, nucleotide sequence and expression of the regulatory region of the histidine operon of *Escherichia coli* K-12. Nucleic Acids Res. 9:2075-2086.

Voll, M. J., E. Appella, and R. G. Martin. 1967. Purification and composition studies of phosphoribosyladenosine triphosphate: pyrophosphate phosphoribosyltransferase, the first enzyme of histidine biosynthesis. J. Biol. Chem. 242:1760-1767.

Wachi, M., M. Doi, S. Tamaki, W. Park, S. Nakajima-Iijima, and M. Matsuhashi. 1987. Mutant isolation and molecular cloning of *mre* genes, which determine cell shape, sensitivity to mecillinam, and amount of penicillin-binding proteins in *Escherichia coli*. J. Bacteriol. 169:4935-4940.

Waley, S. G. 1969. Some aspects of the evolution of metabolic pathways. Comp. Biochem. Physiol. 30:1- 7.

Weber, A. L., and S. L. Miller. 1981. Reasons for the occurrence of the twenty coded protein amino acids. J. Mol. Evol. 17:273-284.

Weil, C., G. Bekler, and J. Reeve. 1987. Structure and organization of the *hisA* gene of the thermophilic archaebacterium *Methanococcus thermolithotrophicus*. J. Bacteriol. 169:4857-4859.

Weng, M., C. A. Makaroff, and H. Zalkin. 1986. Nucleotide sequence of *Escherichia coli pyrG* encoding CTP synthetase. J. Biol. Chem. 261:5568-5574.

White, H. B. 1976. Coenzymes as fossils of an earlier metabolic state. J. Mol. Evol. 7:101-117.

White, D. H., and J. C. Erickson. 1980. Catalysis of peptide bond formation by histidyl-histidine in a fluctuating clay environment. J. Mol. Evol. 16:279-290.

Whitfield, H. J. Jr. 1971. Purification and properties of the wild type and a feed-back resistant phosphoribosyladenosine triphosphate pyrophosphate phosphoribosyltransferase, the first enzyme of histidine biosynthesis in *Salmonella typhimurium*. J. Biol. Chem. 246:899-908.

Winkler, M. E. 1987. Biosynthesis of histidine, p. 395-411. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium:* cellular and molecular biology, vol. 1. American Society for Microbiology, Washington, D.C.

Winkler, M. E. Personal comunication.

Winkler, M. E., D. J. Roth, and P. E. Hartman. 1978. Promoter- and attenuator-related metabolic regulation of the *Salmonella typhimurium* histidine operon. J. Bacteriol. 133:830-843.

Winkler, M. E., and C. Yanofsky. 1981. Pausing of RNA polymerase during *in vitro* transcription of the tryptophan operon leader region. Biochemistry 20:3738-3744.

Yager, T. D., and P. H. von Hippel. 1987. Transcript elongation and termination in *Escherichia coli*, p. 1241-1275. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium:* cellular and molecular biology, vol. 2. American Society for Microbiology, Washington, D.C.

Yanofsky, C. 1981. Attenuation in the control of expression of bacterial operons. Nature 289:751-758.

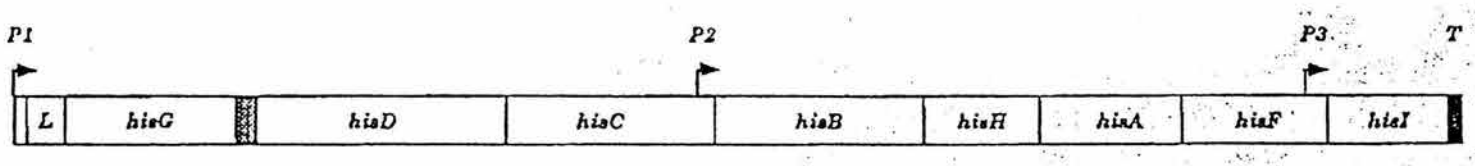
Ycas, M. 1974. On the earlier states of the biochemical system. J. Theoret. Biol. 44:145-160.

Ye, Q.-Z., J. Liu, and C. T. Walsh. 1990. *p*-Aminobenzoate synthesis in *Escherichia coli*: purification and characterization of PabB as aminodeoxychorismate synthase and enzyme X as aminodeoxychorismate lyase. Proc. Natl. Acad. Sci. USA 87:9391-9395.

Zipser, D. 1969. Polar mutations and operon function. Nature 221:21-25.

. *Pichia pastoris*. Unpublished data.

. *Phytophtora parasitica* . Unpublished data.


P1
P2
P3
T
L
hisG
hisD
hisC
hisB
hisH
hisA
hisF
hisI

1 kb

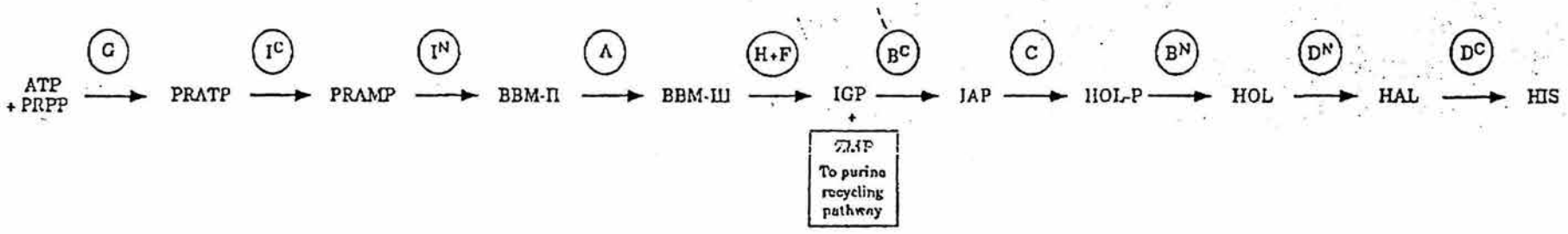
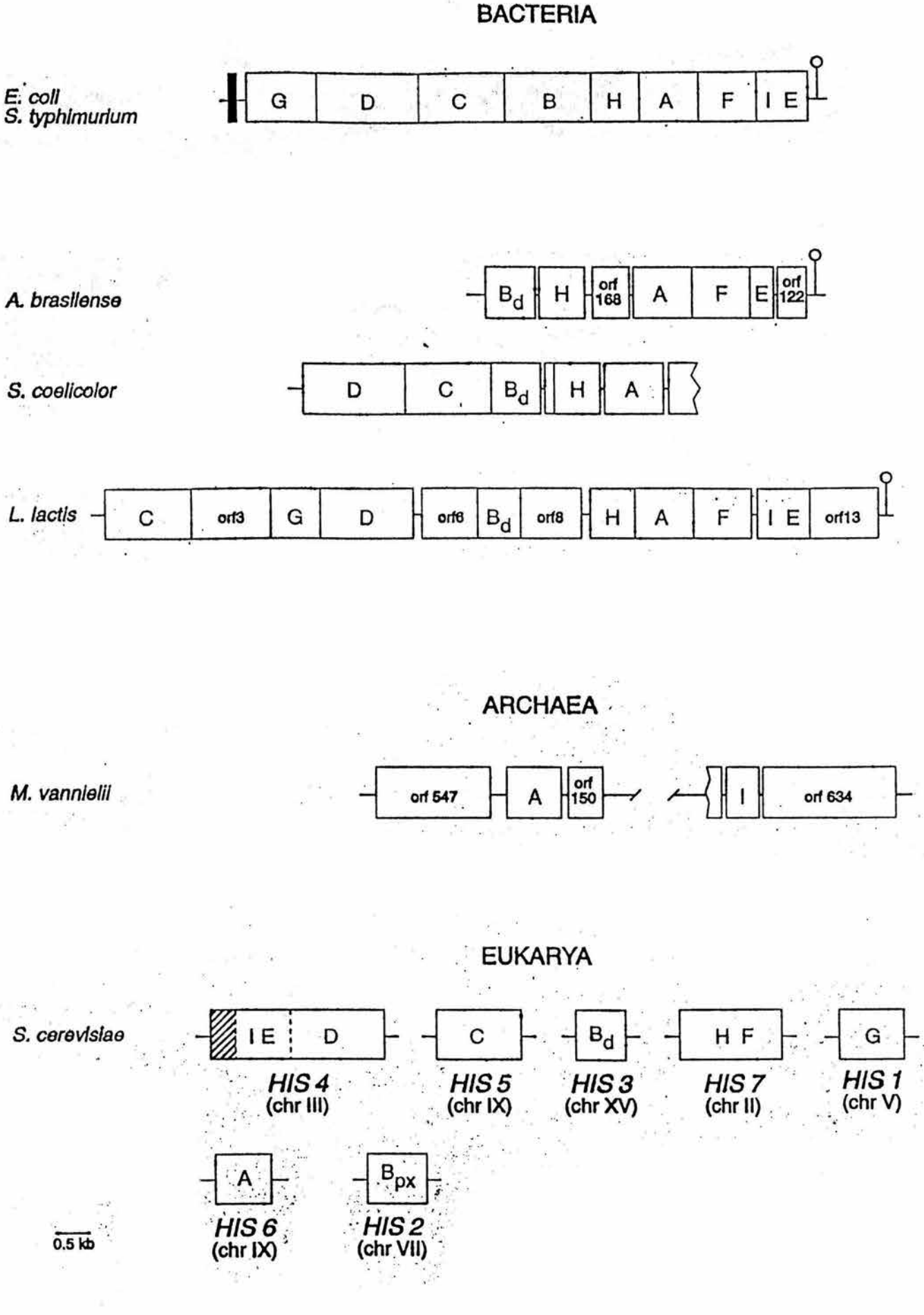
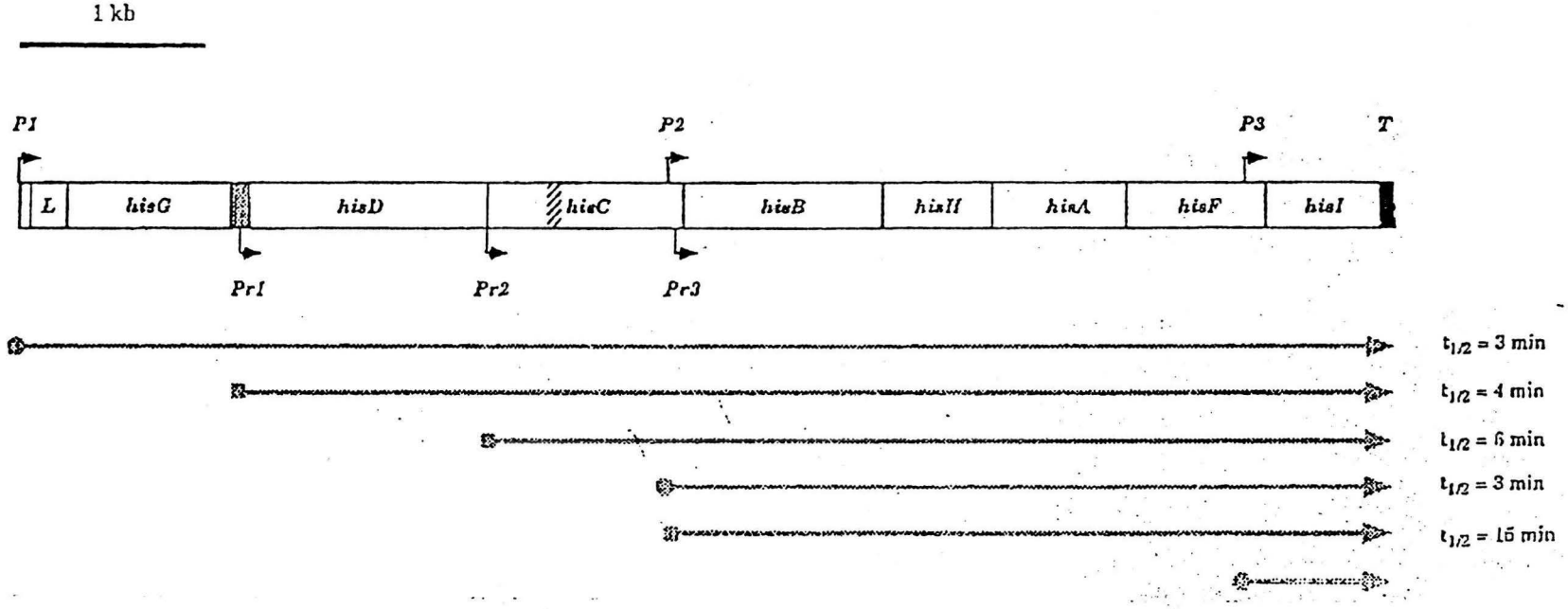

ATP
+ PRPP
G
PRATP
I^C
PRAMP
I^N
BBM-II
A
BBM-III
H+F
IGP
+
ZMP
To purine recycling pathway
B^C
IAP
C
HOL-P
B^N
HOL
D^N
HAL
D^C
HIS

Fig 1


BACTERIA
E. coli
S. typhimurium
G
D
C
B
H
A
F
I E
A. brasilense
Bd
H
orf 168
A
F
E
orf 122
S. coelicolor
D
C
Bd
H
A
L. lactis
C
orf3
G
D
orf6
Bd
orf8
H
A
F
I E
orf13
ARCHAEA
M. vannielii
orf 547
A
orf 150
I
orf 634
EUKARYA
S. cerevisiae
I E
D
HIS 4
(chr III)
C
HIS 5
(chr IX)
Bd
HIS 3
(chr XV)
H F
HIS 7
(chr II)
G
HIS 1
(chr V)
A
HIS 6
(chr IX)
Bpx
HIS 2
(chr VII)
0.5 kb



P1 ttgctttaaggcgtaaaagtggtttaggttaaaaggtAT

Pr1 Heterogeneous 5' ends at multiple RNase E cleavage sites

Pr2 Heterogeneous 5' ends at multiple RNase E cleavage sites

P2 tttaaatccttatgggatcagggcattatcttAC

Pr3 5'-gcggcugccugcggauuacggucggcacccgccaggaaaaccagccggucauugacgccuuacgugcggagccaguauga

P3 gcgtgccgctgatcgcctccggcggcgcgggcacgatggaaacacttttcTG

Fig. 2

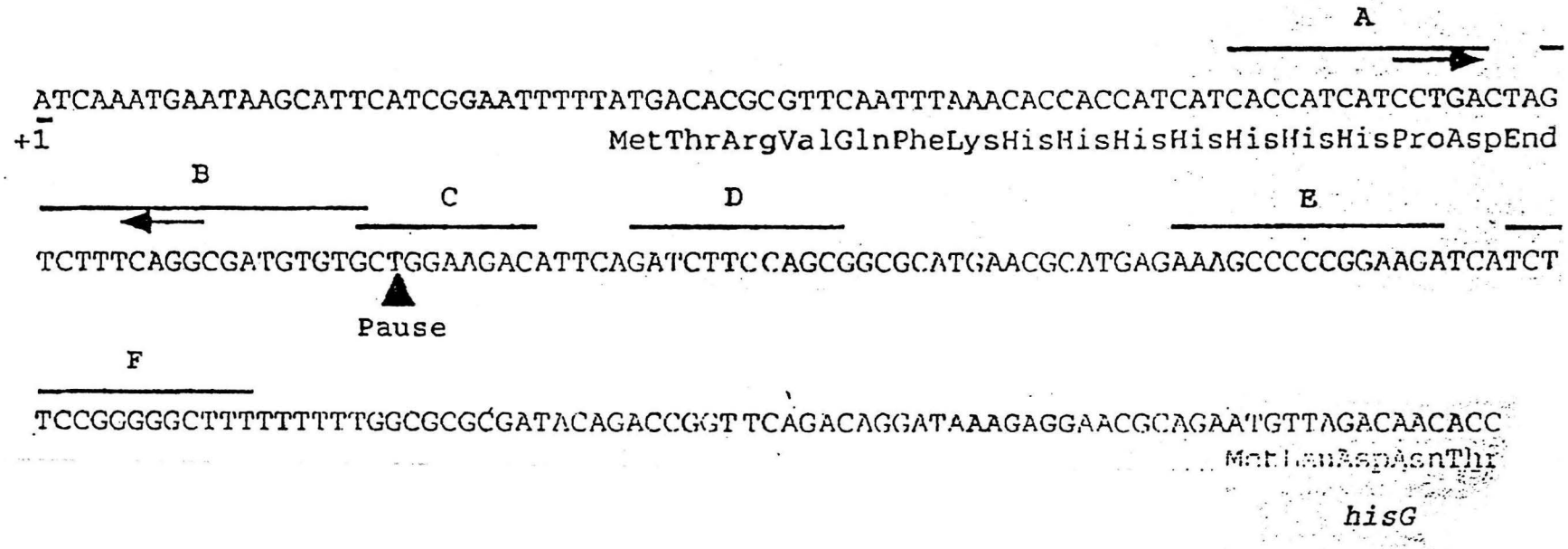

A
ATCAAATGAATAAGCATTCATCGGAATTTTTATGACACGCGTTCAATTTAAACACCACCATCATCACCATCATCCTGACTAG
+1
MetThrArgValGlnPheLysHisHisHisHisHisHisHisProAspEnd
B
C
D
E
TCTTTCAGGCGATGTGTGCTGGAAGACATTCAGATCTTCCAGCGGCGCATGAACGCATGAGAAAGCCCCCGGAAGATCATCT
Pause
F
TCCGGGGGCTTTTTTTTTGGCGCGCGATACAGACCGGTTCAGACAGGATAAAGAGGAACGCAGAATGTTAGACAACACC
MetLeuAspAsnThr
hisG

Fig 4

LEGEND TO FIGURES

FIG. 1. Gene structure of the *his* operon of *S. typhimurium* and metabolic pathway of histidine biosynthesis. Top. The leader region (*L*) and structural genes are drawn to scale. Arrows indicate the relative position of the *P1*, primary promoter; *P2* and *P3*, internal promoters. The solid box represents the *T*, Rho-independent bifunctional transcription terminator. The shaded box between *hisG* and *hisD* indicates the REP-like element which is absent in *E. coli* and substituted by a 5 bp intercistronic region. The single gene encoding a bifunctional enzyme, formerly known as *hisIE* (Winkler, 1987) has been renamed *hisI* in *E. coli* (Bachmann, 1990). In this review we have therefore used *hisI* for organisms with a single gene and *hisI* and *hisE* for organisms with two independent genes. Another gene encoding a bifunctional enzyme, *hisB*, is often split in two separate genes in different organisms. We refer to them as *hisB* proximal (*hisBpx*) encoding the HOL-P phosphatase and *hisB* distal (*hisBd*) encoding the IGP dehydratase.Bottom. The order of the reactions and the intermediates in the pathway are indicated. PRPP, 5-phosphoribosyl-α-1-pyrophosphate; PRATP, *N'*-5'-phosphoribosyl-ATP; PRAMP, *N'*-5'-phosphoribosyl-AMP; BBM-II (5'-ProFAR), N'-[(5'-phosphoribosyl)-formimino]-5 aminoimidazole-4 carboxamide-ribonucleotide; BBM-III (5'-PRFAR), N'-[(5'-phosphoribulosyl)-formimino]-5 aminoimidazole-4 carboxamide-ribonucleotide; IGP, imidazole glycerol-phosphate; ZMP (PRAIC, AICAR), 5'-phosphoribosyl- 4-carboxamide-5-aminoimidazole; IAP, imidazole acetol-phosphate; HOL-P, L-histidinol-phosphate; HOL, L-histidinol; HAL, L-histidinal; HIS, L-histidine. The enzymes encoded by the relative cistrons are indicated in the circles above the arrows. When a gene encodes a bifunctional enzyme the domains performing the single reaction: amino- or carboxyl-terminal are indicated by a superscript N or C, respectively.

FIG. 2. Clustering of *his* genes that have been sequenced and mapped in different organisms.

FIG. 3. Transcriptional map of the *S. typhimurium his* operon. Top. The operon is represented as in Fig. 1. The striped area within *hisC* defines the RNase E target sites. Arrows above the operon indicate the relative position of the *P1*, primary promoter; *P2* and *P3*, internal promoters; arrows below the operon indicate the relative position of the three processed (*Pr1*, *Pr2* and *Pr3*) species. Middle. Arrows indicate the limits and extent of the initiated (starting with a circle) and processed (starting with a square) transcripts. The half-lives of individual species are indicated on the right (Alifano, unpublished data; Alifano et al., 1994a). Bottom. DNA sequences of the promoter regions and RNA sequences and/or features of the 5'-regions of the processed transcripts (Alifano, unpublished data; Alifano et al., 1994b; Carlomagno et al., 1988). The -35 and -10 putative consensus sequences are in bold characters. Nucleotides corresponding to the 5' ends of the transcripts are in capital letters. The stem/loop structure at the 5' end of the processed *Pr3* species is indicated by convergent arrows. The CCA consensus, the ribosome binding site and overlapping stop and start codons of *hisC* and *hisB* are in bold characters.

FIG. 4.

Paralogous histidine biosynthetic genes: evolutionary analysis of the *Saccharomyces cerevisiae HIS6* and *HIS7* genes

Key words (histidine biosynthesis; yeast; evolution of metabolic pathways; phylogeny; *HIS6 HIS7* genes paralogous duplications)

Renato Fani, Claudia Barberio, Enrico Casalone, Duccio Cavalieri, Antonio Lazcano[a], Pietro Liò, Elena Mori, Brunella Perito and Mario Polsinelli

Dipartimento di Biologia Animale e Genetica, Università degli Studi di Firenze, Via Romana 17, I-50125 Firenze, Italy; and [a]Facultad de Ciencias, Universidad Nacional Autónoma de México, Apartado Postal 70-407, Cd Universitaria, México 04510, D.F., México

Correspondence to: Dr Renato Fani, Dipartimento di Biologia Animale e Genetica, Via Romana 17, I-50125 Firenze, Italy; tel +39-55-220498; Fax +39-55-222565; e-mail: r_fani@dbag.unifi.it

Abbreviations

LCA = last common ancestor

nt = nucleotide(s)

ORF = open reading frame

5'Pro-FAR isomerase = phosphoribosyl-5-amino-1-phosphoribosyl-4-imidazole carboxiamide isomerase (EC 5.3.1.16)

SUMMARY

The *HIS6* gene from *Saccharomyces cerevisiae* strain YNN282 is able to complement both the *S. cerevisiae HIS6* and the *Escherichia coli hisA* mutations. The cloning and the nucleotide sequence indicated that the gene encodes a putative phosphoribosyl-5-amino-1 - phosphoribosyl - 4 - imidazolecarboxiamide isomerase (5' Pro-FAR isomerase, EC 5.3.1.16) of 261 amino acids, with a molecular weight of 29554 d. The *HIS6* gene product shares a significant degree of sequence similarity with the prokaryotic HisA proteins and HisF proteins, and with the C-terminal domain of the *S. cerevisiae* HIS7 protein (homologous to HisF), indicating that the yeast *HIS6* and *HIS7* genes are paralogous. Moreover, the *HIS6* gene is organized into two homologous modules half the size of the entire gene, typical of all the known prokaryotic *hisA* and *hisF* genes. The structure of the yeast *HIS6* gene supports the two-step evolutionary model suggested by Fani et al. (1994) to explain the present day *hisA* and *hisF* genes; the model claims that *hisF* originated from the duplication of an ancestral *hisA* gene, which in turn was the result of an earlier gene elongation event involving an ancestral module half the size of the extant gene. Results reported in this paper also suggest that these two successive paralogous gene duplications took probably place in the early steps of molecular evolution of the histidine pathway, well-before the diversification of the three cell lineages, and that this pathway was one of the metabolic activities of the last common ancestor. The molecular evolution of the yeast *HIS6* and *HIS7* genes is also discussed.

INTRODUCTION

Histidine biosynthesis is one of the best characterized anabolic pathways. There is a large body of genetic and biochemical information, including operon structure, gene expression, and increasingly larger sequence databases. This pathway has been the subject of extensive studies, mainly in the enterobacteria *Escherichia coli* and *Salmonella typhimurium* (Alifano et al., 1996). The complete nucleotide sequence of the *his* operons of these two enterobacteria has been determined (Carlomagno et al., 1988), as well as that of *Haemophilus influenzae* (Fleischman et al., 1995). In *E. coli* and *S. typhimurium* the pathway is unbranched, and includes a number of complex and unusual biochemical reactions, consisting of nine intermediates and of eight distinct enzymes.

The histidine pathway has been also investigated in other prokaryotes, including archaebacteria and eubacteria (for recent reviews see Fani et al., 1995 and Alifano et al., 1996). In eukaryotes, where the pathway is similar to the enterobacterial one, many histidine genes from several different organisms have been identified, cloned and sequenced, but they have been extensively studied only in the yeast *Saccharomyces cerevisiae,* where the seven genes involved in histidine biosynthesis (*HIS1-HIS7*) are located on six different chromosomes (Broach, 1981). Six of these genes, *HIS1* (Hinnebusch and Fink, 1983), *HIS2* (Malone et al., 1994), *HIS3* (Sthrul, 1985), *HIS4* (Donahue et al., 1982), *HIS5* (Nishiwaki et al., 1987) and *HIS7* (Kuenzler et al., 1993) have been cloned and sequenced. The cloning of *HIS6,* which maps on chromosome IX very close to the *RPB3* gene (Kolodzei and Young, 1989) and encodes an enzyme homologous to the *E. coli* HisA protein that catalyzes the fourth step of histidine biosynthesis, is reported in this paper. Genes homologous to *HIS6* have been cloned and sequenced from a variety of archaea and bacteria, but up to now no eukaryotic gene homologous to the prokaryotic *hisA* has been cloned and/or sequenced. This gene is particularly interesting from an evolutionary point of view, since it has been shown that there is a paralogous gene, *hisF,* involved in the same metabolic route (Fani et al., 1994, 1995). All the available evidence suggests that the two genes, *hisA* and *hisF,* are the result of two successive duplications; the ancestral *hisA* gene might be the result of a gene elongation event involving an ancestral module *(hisA1)* half the size of the extant *hisA* gene (Fani et al., 1994, 1995). The *hisF* gene would have arisen as a result of a duplication of an ancestral *hisA* gene. Moreover, the 3' terminal region of *S. cerevisiae HIS7* gene (Kuenzler et al., 1993), which is homologous to

the eubacterial *hisF*, also has significant degree of sequence homology with the prokaryotic *hisA* genes and shares the same two-module structure of prokaryotic *hisA* and *hisF* genes (Fani et al., 1995). Identification and analysis of archaeal *hisF* genes and eukaryotic *hisA* genes could provide additional insights of the evolution of these putative paralogous genes, and could help not only in understanding how the histidine biosynthetic pathway was assembled, but also if it was indeed one of the metabolic abilities of the last common ancestor (Lazcano et al., 1992; Fani et al., 1995). Due to the apparent antiquity of this pathway, the available histidine genes have been successfully used to infer the evolutionary relationships between organisms belonging to the three domains, i.e. the archaea, bacteria and eucarya. Furthermore paralogous genes have been used to obtain rooted universal trees (Iwabe et al., 1989; Gogarten et al., 1989). Thus, a systematic study of the two paralogous duplications leading to *hisA* and *hisF*, that took probably place prior the diversification of the three cell lineages, may provide an important set of data for the construction of deep phylogenies that may shed some light on the proper rooting of the universal tree of life.

In this paper we report the cloning, the characterization and the structural analysis of the yeast *HIS6* gene homologous to the prokaryotic *hisA*, which encodes for 5'Pro-FAR isomerase, catalyzing the fourth step of histidine biosynthesis in *S. cerevisiae*. The evolutionary significance of this highly conserved gene is also discussed.

RESULTS AND DISCUSSION

(a) Cloning and sequencing of the *S. cerevisiae HIS6* gene

Since many yeast genes like *HIS2* (Ratzkin and Carbon, 1977), *HIS3* (Sthrul et al., 1976) and *HIS7* (Kuenzler et al., 1993) have been functionally expressed in *E. coli*, we attempted to isolate the yeast *HIS6* gene from a yeast genomic library of the strain YNN282 by complementation of the *hisA* mutation of *E. coli* strain FB184 (Goldschmidt et al., 1970). Three complementing clones, pRD2, pRD5 and pRD64, containing *Sau*3AI partial fragments of 7.7, 12.0 and 2.8 Kb, respectively, were isolated. Restriction endonuclease analysis of the three recombinant plasmids showed that their DNA inserts partially overlapped (not shown). A restriction map of plasmid pRD64, containing the 2.8 kb yeast DNA insert, was constructed (Fig. 1). Before further characterization, we verified that the cloned fragment had the same restriction map of the

corresponding chromosomal fragment and that no rearrangements occurred during cloning. For this purpose, the 2.8 Kb fragment was used as a probe in Southern blotting experiments with total DNA from the yeast strain YNN282 digested with *Eco*RI, *Hind*III or *Cla*I. The hybridization patterns obtained were in agreement with the restriction map of the cloned fragment (data not shown).
Functional analysis of the plasmid pRD64 in yeast showed that it is able to complement the *HIS6* mutation of *S. cerevisiae* strain 1437-8c, whereas the vector pRS316 does not. Results of plasmid-loss experiments (not shown) indicated that the His^+ phenotype of the transformants was associated to the presence of plasmid pRD64.

In order to locate the *S. cerevisiae* DNA region able to complement the *E. coli hisA* and the yeast *HIS6* mutations, we constructed the subclones shown in Fig. 1. Subcloning esperiments localized the complementing activity of pRD64 plasmid on a 1.5 Kb *Sau*3AI-*Cla*I fragment of plasmid pRD643 (Fig. 1). The nucleotide sequence of this fragment (1521 bp long) was then determined (Fig. 2), and its analysis evidentiated the presence of two divergent open reading frames (ORF), the first one of 783 bp (ORF 783) on one strand, and a truncated one (of 213 bp) on the other one. The truncated ORF showed 100% of sequence similarity with the yeast *RPB3* gene encoding the RNA polymerase II subunit (EMBLGenBank accession number M27496) (Kolodzei and Young, 1989). Moreover, the 284 nt of the 3' end of ORF783 gave 100% similarity with the sequence of a truncated ORF localized downstream from the above mentioned *RPB3* gene. Very tight linkage between the *RPB3* and *HIS6 S. cerevisiae* genes was shown by Kolodzei and Young (1989). The evidence that ORF 783 is able to complement the *S. cerevisiae HIS6* and the *E. coli hisA* mutation, together with its physical linkage to RPB3, strongly suggests the identity between ORF783 and *HIS6* gene. To confirm this possibility, we have compared the amino acid sequence of the putative protein encoded by ORF783 with the archaebacterial and eubacterial HisA protein amino acid sequences known to date. Results obtained are shown in Table I and Fig. 3; they reveal a good degree of similarity between the ORF783 encoded protein and the prokaryotic HisA polypepides (ranging from 20 to 28 % of identical amino acids, and from 32 to 42% of similar amino acids), the highest values being observed with the archaebacterial HisA proteins (Table I). The ORF783 encoded protein and prokaryotic HisA proteins also share very similar hydropathic profiles and predicted secondary structures (not shown). Taken altogether, these data indicate that

ORF783 actually corresponds to the yeast *HIS6* gene, encoding a 5' Pro-FAR isomerase which catalyzes the fourth step of histidine biosynthesis.

(b) The nucleotide sequence analysis of the yeast *HIS6* gene

The *HIS6* gene has a length of 783 bp, a 39.2 % GC content and a codon usage similar to that of other *S. cerevisiae* histidine genes sequenced. Analysis of the *HIS6* nucleotide (nt) sequence revealed the absence of introns.The *HIS6* gene is separated from *RPB3* gene by a short intergenic region (163 nt), in agreement with previous genetic analysis (Kolodzei and Young 1989) (Fig. 2). The *HIS6* flanking regions were analysed for the presence of regulatory sequences. The 346 nt region upstream of the putative ATG codon has a 60% AT content with three poly (dA-dT) stretches (positions 40-54, 107-119 and 160-170) and a putative TATA-box at positions 115-119 ($^{5'}$TATAA$^{3'}$). A putative mRNA initiation site of the TC(G/A)A class (Hahn et al., 1985) is also present at positions 303-306. The presence of these sequences suggests a constitutive transcription of *HIS6*. In addition, a GTCGCATGAGACG sequence which perfectly fits the ABF1 consensus sequence RTCYNNNNNACG (Verdier, 1990) was found at positions 183-195, suggesting that *HIS6* might be under the transcriptional control of the ubiquitous factor ABF1.

Finally, the analysis revealed the presence of four sequences matching the core sequence of the GCN4 binding site (TGACTC) (Arndt and Fink, 1986) , as for other *HIS* gene under the general control of amino acid biosynthesis, with the exception of *HIS2*. All the four GCN4 binding sites were in opposite orientation.

The 163 nt intergenic region between *HIS6* and *RPB3* genes showed an AT content of about 74%, which is characteristic of other similar regions in yeast. A TACATA sequence was also present, which has been reported to be able to promote mRNA termini production downstream from its position (Russo et al., 1993).

(c) Homology between the products of the *S. cerevesiae HIS6* and *HIS7* genes

It has been previously shown that the eubacterial *hisA* and *hisF* genes are paralogous, in that they are the descendants of a common ancestral gene (Fani et al., 1994). The only eukaryotic gene homologus to the eubacterial *hisF* cloned and characterized so far is the *S. cerevisiae HIS7* gene, which encodes a glutamine amidotransferase cyclase, a bifunctional enzyme whose C-terminal

moiety is homologous to the eubacterial HisF proteins (Kuenzler et al., 1993). Comparison of the amino acid sequence of the yeast HIS6 and HIS7 proteins revealed that they share a significant degree of similarity (32% of similar amino acids) (Table I and Fig. 4). This suggests that the two corresponding genes like the eubacterial *hisA* and *hisF* genes, are paralogous, i.e. that they are the descendants of a common ancestral gene that underwent duplication. The same analysis revealed that the yeast HIS6 protein shares a significative degree of similarity with the eubacterial HisF proteins (Table I).

(d) The *S. cerevisiae HIS6* gene has an internal duplication

Detailed analysis of the amino acid sequence of the *HIS6* gene product has also revealed an internal repetition of half the size of the molecule. As shown in Fig. 5, when the *HIS6* gene product was split at residue 127 and the two moieties aligned, a significant value of sequence similarity of 31% was found. Therefore, the *HIS6* gene shows the same internal organization of the other bacterial *hisA* and *hisF* genes already known (Fani et al., 1994, 1995). These data, on the whole, are in agreement with the two-step model for the evolutionary pathway leading to the *hisA* and *hisF* genes (Fani et al., 1994, 1995), and suggest that each of the corresponding proteins is formed by two similar halves with a pseudosimmetrical axis (McLachlan, 1987). According to this hypothesis, *hisA* was the early result of the duplication of an ancestral module, corresponding to the 5' half moiety (*hisA1*) of the present-day *hisA* gene, followed by the fusion of the two modules (Fig. 6) (Fani et al., 1994). Subsequently, *hisF* originated from the duplication of *hisA*. The finding that each half of the *HIS6* gene also has a higher degree of similarity with the corresponding portions of the *HIS7* gene than with the other parts (not shown), confirms the hypothesis that the *hisF* gene arose by the duplication of the entire *hisA* gene rather than from multiple successive duplications of the ancestral module (Fani et al., 1994). Although the *hisF* gene has not been described yet in archaea, the sisterhood relationship between them and the eucarya (Iwabe et al., 1989; Gogarten et al., 1989) lead us to predict a similar structure for the corresponding *hisF* genes. Thus, it is likely that the two successive duplication events leading to the extant *hisA* and *hisF* took probably place before the diversification of the three cell domains.

(e) Evolution of *S. cerevisiae HIS6* and *HIS7* genes

The fact that genes homologous to *hisA* have been identified in all of the three cell lineages, and show always the same two-module structure, strongly suggests that they are the orthologous descendants of a sequence that may have been part of the genome of the last common ancestor (LCA) (Fig. 6). The entire pathway was thus probably assembled prior of the appearance of the LCA, and paralogous duplications and gene fusion events appear to have played a major role in shaping the histidine biosynthetic pathway. Therefore, the ability to synthesize histidine is likely a very ancient property of biological systems.

It has been also postulated that during the early evolution of the histidine biosynthesis at least some of its reactions were mediated by low substrate specificity enzymes participating in different part of the same pathway (Fani et al., 1995). This interpretation is based on the so-called patchwork hypothesis, according to which primitive metabolic routes were mediated by enzymes of low substrate specificity that were eventually recruited into different pathways (Ycas, 1974; Jensen, 1976). This hypothesis is also consistent with the possibility that an ancestral pathway may have had a primitive enzyme catalyzing two or more similar reactions on related substrates of the same metabolic route. Its substrate specificity could have been subsequently refined as a result of duplication and divergence events of the gene encoding this ancestral enzyme.

The possibility that histidine biosynthesis was originally mediated by less specific enzymes is supported by the common origin of the imidazole glycerol-P synthase encoded by the *E. coli hisH* gene, with other *E. coli* G-type glutamine amidotransferases which participate in the biosynthesis of purines, pyrimidines, arginine, tryptophan, and other ancient pathways (Fani et al., 1995). The finding that the yeast *HIS6* gene has the same internal organization of its prokaryotic homologues and that it shares sequence similarity with *HIS7*, may also be interpreted to indicate that products of the *hisA* and *hisF* genes and their homologues are the descendants of a gene encoding a less specific enzyme. This possibility is supported not only by sequence analysis (Fani et al., 1994, 1995), but also by biochemical data. The purification and characterization of the enzymatic properties of the HisH and HisF proteins have shown that their non-covalent association is essential to form an active IGP synthase, composed of one subunit each of HisH and HisF (Klemm and Davisson, 1993). The IGP synthase is able to transform the PRFAR into AICAR and IGP *via* a glutamine molecule with no free intermediate (see below). As an isolated subunit, HisH has no detectable catalytic activities (Klemm and Davisson, 1993). An enhancement of the glutaminase activity (in the absence of

PRFAR) can be elicited by inclusion of IGP (which is one of the products of the reaction), or by 5' ProFAR, which is the biosynthetic precursor to the substrate of the HisA enzyme. Although 5'-ProFAR binds to the synthase with low affinity, the stimulation of the glutaminase activity has been interpreted as a specific interaction with the HisF domain in a manner that mimics the substrate PRFAR. This observation is in agreement with the model suggesting that *hisA* and *hisF* have a common ancestor. Sheridan and Venkataraghavan (1992) have proposed a potential substrate recognition in HisA and HisF on the basis of a common signature in both of these proteins which was assumed to be a strand-helix-strand structure that could bind glycerol-phosphate moieties. The identification of a highly conserved phosphate-binding site which is part of a β/α barrel in these proteins (Fig. 3) suggests that the HisA and HisF proteins may be related to a larger set of enzymes involved in the binding of heterocyclic compounds (Bork et al., 1995). It it thus possible that PRFAR or 5'ProFAR or both of them could have been the substrate(s) of the ancestral HisA enzyme or of a more ancestral enzyme half the size of the extant HisA protein.

After the divergence of the three cell lineages from the LCA, the structure and the organization of the above mentioned paralogous genes may have undergone rearrangements in the different cell lineages. In most bacteria, where at least some *his* genes are arranged in operons (*E. coli, S. typhimurium, Klebsiella pneumoniae, Lactococcus lactis, Streptomyces coelicolor, Haemophilus influenzae* and *Azospirillum brasilense*), the sequences are fused, and share the same relative order (Fani et al., 1995). Although the available information is still limited, among the archaea the *hisA* gene is not flanked by any *his* gene. In *S. cerevisiae HIS6* and *HIS7* are located on different chromosomes (IX and II, respectively).

The extant structure of the *S. cerevisiae HIS6* and *HIS7* also suggests that they underwent different genetic changes. The comparative analysis of all of the known *hisA* genes and of yeast *HIS6* gene, revealed that they have almost the same length and that the conserved sites they share are distributed throughout the sequence (Fig. 3); this suggests that no large rearrangements have been incorporated during evolution. However, the situation is different for the *S. cerevisiae HIS7* gene, which is composed by the two corresponding bacterial cistrons *hisH* and *hisF*, fused in the order *H-F* (Kuenzler et al., 1993). In fact, the HisH moiety and the second half of the HisF moiety (HisF2) of the polypeptide encoded by *HIS7* are conserved, whereas the HisF1 module shows a

strong rearrangement with six different insertions of different length. Since these insertions are not present in any other *hisA* and *hisF* genes, it is likely that they have been incorporated in *HIS7* after the fusion of *hisH* and *hisF*. Nonetheless on the basis of the available data it is not possible to conclude that the rearrangement leading to the extant *HIS7* took place only in eukaryotes, or that a similar fusion also took place in some archaebacteria or eubacteria. The cloning and analysis of archaebacterial and additional eukaryotic *hisF* genes might help to solve these issues. These rearrangements may be related to the function of the HisF protein itself. In fact, it has been postulated (Rieder et al., 1994) that PRFAR is transferred from HisA to HisF and subsequently cleaved ammonolyticaly by the glutamine amido group, which is received from a complex of glutamine-HisH. The latter protein do not take part in the cleavage of PRFAR, but it plays its role as donor of the glutamineamido group in a suitable and activated position to PRFAR on HisF. It has also been postulated that the two modules of the HisF protein (HisF1 and HisF2) have different activities, one (HisF1) responsible of the interaction with HisH, and the other (HisF2) responsible of catalysis (Fani et al., 1995). This possibility is supported by the recent finding of Rieder et al. (1994), suggesting that a mutation in the *K. pneumoniae hisF1* module of the *hisF* gene affects the interaction between HisH and HisF, but the catalytic capacity remains intact. It is possible that the fusion event involving *hisH* and *hisF* and leading to the ancestral yeast *HIS7* gene could have affected the correct interaction between the HisH moiety and the catalytic domain of HisF. It is also possible that the internal rearrangements, corresponding to the six insertions, restored the ability of the protein to perform its function. This could also explain why the fusion event between HisF and HisH occurred in such a way to place HisH upstream of HisF and not *vice versa*. In the later case, the fusion event would have probably destroyed the catalytic site, requiring a greater number of molecular rearrangements to restore the protein's functionality.

It is not known, at the time being, whether the two HisA modules perform different functions, or both of them are involved in the catalytic activity of the enzyme.

(f) Molecular phylogenies of the *hisA* and *hisF* genes

A molecular phylogenetic analysis of the proteins HisA and HisF and their eukaryotic homologoues (His6 and His7) was done in order to infer the possible evolutionary relationships among the organisms from which these genes have been sequenced. The phylogenetic analysis was

performed by using the maximum likelihood method, as described in the *Materials and Methods*. The unrooted maximum likelihood phylogenetic trees calculated for the *hisA* and *hisF* gene products are shown in Fig. 7. The phylogenetic analysis placed *S. cerevisiae* close to the archaea, but nearer to the α-purple nitrogen fixing bacterium *A. brasilense*, which in turn is closer to Gram positive bacteria than to enterobacteria (γ-purple bacteria). It is possible that the peculiar position of *A. brasilense* is due to a possible horizontal gene transfer between Gram positive and Gram negative bacteria (Fani et al., 1995). In fact, the proximity of *S. cerevisiae* with *A. brasilense* might be interpreted as an indication of a horizontal gene transfer between an ancestor of eukaryotes and an ancestor of the extant α-purple bacteria. This would be in agreement with recent proposal of a chimeric origin for the eukaryotic nucleocytoplasm (Gupta and Singh 1994).

(g) Conclusions

(1) The *HIS6* gene from *S. cerevisiae* can complement the *E. coli hisA* mutation.

(2) The analysis of the aa sequence of the *HIS6* gene product revealed that this gene is homologous to all the procaryotic *hisA* genes known so far. Moreover, the *HIS6* and *HIS7* genes are paralogous and the *HIS6* is composed by two homologus modules half the size of the entire gene.

(3) The *HIS6* and *HIS7* genes, as well as their prokaryotic counteparts, are the result of two successive paralogous duplication events which took probably place before the appearance of the LCA.

(4) A gene homologus to the yeast *HIS6* gene was part of the genome of the LCA.

(5) The two genes are probably the descendants of an ancestral gene encoding an enzyme with low substrate specificity.

REFERENCES

Alifano, P., Fani, R., Liò, P., Lazcano, A., Bazzicalupo, M., Carlomagno M.S. and Bruni, C.B.: Histidine biosynthetic pathway and genes: structure, regulation and evolution. Microbiol. Rev. (1996) (in press).

Arndt, K. and Fink G.R.: GCN4 protein, a positive transcription factor in yeast, binds general control promoters at all 5' TGACTC3' sequences. Proc. Natl. Acad. Sci. USA 83 (1986) 8516-8520.

Bork, P., Gellerich, J., Groth, H., Hooft, R. and Martin, F.: Divergent evolution of a β/α - barrel subclass: detection of numerous phosphate-binding sites by motif search. Protein Science 4 (1995) 268-274.

Broach , J.R.: Gene of *Saccharomyces cerevisiae*. In Strathern J.N., Jones E.W., Broach J. R. (Eds) The molecular Biology of the yeast Saccharomyces: life cycle and inheritance. Cold Sping Harbor Laboratory Press, Cold Spring Harbor (1981) 653-727.

Carlomagno, M.S., Chiarotti, L., Alifano, P., Nappo, A.G. and Bruni, C.B.: Structure of the *Salmonella typhimurium* and *Escherichia coli* K-12 histidine operons. J. Mol. Biol. 203 (1988) 585-606.

Cue, D., Beckler, G., Reeve, J. and Konisky, J.: Structure and sequence divergence of two archaebacterial genes. Proc Natl Acad Sci USA 81(1985) 8019-3023.

Delorme, C., Ehrlich, S.D. and Renault, P.: Histidine biosynthesis genes in *Lactococcus lactis* subsp.*lactis*. J Bacteriol 174 (1992) 6571-6579.

Donahue, T.F., Farabaugh, P.J. and Fink, G.R.: The nucleotide sequence of the *His4* region of yeast. Gene 18 (1982) 47-59.

Fani, R., Alifano, P., Allotta, G., Bazzicalupo, M., Carlomagno, M.S., Gallori, E., Rivellini, F. and Polsinelli, M.: The histidine operon in *Azospirillum brasilense*: organization, nucleotide sequence and functional analysis. Res. Microbiol. 144 (1993) 187-200.

Fani, R., Liò, P., Chiarelli, I. and Bazzicalupo M.: The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the *hisA* and *hisF* genes. J. Mol. Evol. 38 (1994) 489-495.

Fani, R., Liò, P. and Lazcano, A.: Molecular evolution of the histidine biosynthetic pathway. J Mol Evol 41 (1995) 760-774.

Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17 (1981) 368-376.

Fleischmann, R.D., Adams, M.A., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R.T., Bult, C.J., Tomb, J-F., Dougherty, B.A., Merrick, J.M., McKenney, K., Siutton, G., FitzHugh, W., Fields, C., Gocayne, J.D., Scott, J., Shirley, R., Liu, L-I., Glodek, A., Kelley, J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedlom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geohagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O. and Venter, J.C.: Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269 (1995) 496-512.

Fox, G.E., Pechmann, K.R. and Woese, C.R.: Comparative cataloging of 16S rRNA: a molecular approach to prokaryotic systematics. Int. J. Syst. Bacteriol. 27 (1977) 44-57.

Gietz, D., St Jean, A., Woods, R.A. and Schiestl, H.: Improved method for high efficiency transformation of intact yeast cells. Nucl. Acids Res. 20 (1992) 1425.

Gogarten, J.P., Kibak, H., Diitrich, P., Taitz, I., Bowman, E.J. and Bowman, B.J.: Evolution of the vacuolar H^+ ATPases: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA 86 (1989): 6661-6665.

Goldschmidt, E.P., Cater, M.S., Matney, T.S., Butler, M.A. and Greene, A.: Genetic analysis of the histidine operon of *Escherichia coli* K12. Genetics 66 (1970) 219-229.

Gupta, R.S. and Singh, B.: Phylogenetic analysis of 70 kD heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. Curr. Biology 4 (1994) 1104-1114.

Hahn, S., Hoar, E.T., and Guranete, L.: Each of three "TATA elements" specifics a subset of the transcription initiation sites at the CYC1 promoter of *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA 82 (1985) 8562-8566.

Hanahan , D.: Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. 166 (1983) 557-580.

Higgins, D.G. and Sharp, P.M.: CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. Gene 73 (1988) 237-244.

Hinnebusch, A.G. and Fink, G.R.: Repeated DNA sequences upstream from *HIS1* also occurr at several other co-regulated genes in *Saccharomyces cerevisiae*. J. Biol. Chem. 258 (1983) 5238-5247.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. and Miyata, T.: Evolutionary relationship of archabacteria, eubacteria and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA 86 (1989) 9355-9359.

Jensen, R.A. Enzyme recruitment in evolution of new function. Ann. Rev. Microbiol. 30 (1976) 409-425.

Klemm, T. and Davisson, V.J.: Imidazole glycerol phosphate synthase: the glutamine amidotransferase in histidine biosynthesis. Biochemistry 32 (1993) 5177-5186.

Kolodzei, P. and Young, R.A.: RNA polymerase II subunit RPB3 is an essential component of the mRNA transcription apparatus. Mol. Cell. Biol. 9 (1989) 5387-5394.

Kuenzler, M., Balmelli, T., Egli, C.M., Paravicini, G. and Braus, G.H.: Cloning, primary structure, and regulation of the *HIS7* gene encoding a bifunctional glutamine amidotransferase: cyclase from *Saccharomyces cerevisiae*. J. Bacteriol. 175 (1993) 5548-5558.

Lazcano, A., Fox, G.E. and Oro', J.:Life before DNA: the origin and evoution of early Archean cells. In Mortlock, R.P. (Ed.) The Evolution of Metabolic Function, CRC Press, Boca Raton FL, (1992) pp. 237-295.

Li, W.H. and Graur, D.: Fundamentals of Molecular Evolution. Sinauer Sunderland, 1991.

Limauro, D., Avitabile, A., Cappellano, C., Puglia, A.M. and Bruni, C.B.: Cloning and characterization of the histidine biosynthetic gene cluster of *Streptomyces coelicolor* A3(2). Gene 90 (1990) 31-41.

Malone, R.E., Sangkyu, K., Bullard, S.A., Lundquist, S., Hutchings-Crow, L., Crampton, S., Lutfiya, L. and Lee, J.: Analysis of a recombination hotspot for gene conversion occurring at the *HIS2* gene of *Saccharomyces cerevisiae*. Genetics 137 (1994) 5-18.

McLachlan, A.D.: Gene duplication and the origin of repetitive protein structures. Cold Spring Harbor Symp. Quant. Biol. 52 (1987) 411-420.

Nishiwaki, K., Hayashi, N., Irie, S., Chung, D.H., Harashima, S. and Oshima, Y.: Structure of the yeast *HIS5* gene responsive to general control of amino acid biosynthesis. Mol. Gen. Genet. 208 (1987) 159-167.

Ratzkin, B. and Carbon, J.: Functional expression of yeast DNA in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 74 (1977) 487-491.

Rieder, G., Merrick, M.J., Castorph, H. and Kleiner, D.: Function of *hisF* and *hisH* gene products in histidine biosynthesis. J Biol. Chem. 269 (1994) 14386-14390.

Rose, M.D., Winston, F. and Hieter, D.: In Methods in Yeast Genetics. A laboratory manual. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, NY, 1990.

Russo, P., Li, W., Guo, Z. and Sherman, F.: Signal that produce 3' termini in CYC1 mRNA of the yeast *Saccharomyces cerevisiae*. Mol. Cell. Biol. 13 (1993) 7836-7849.

Sambrook, J., Fritsch, E.F. and Maniatis, T.: Molecular Cloning. A laboratory manual. 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989.

Sanger, F.,Nicklen, S. and Coulson, A.R.: DNA sequencing with chain terminating inhibitors.Proc. Natl. Acad. Sci. USA 74 (1977) 5463-5466.

Sheridan, R.P. and Venkataraghavan, R.: A systematic search for protein signature sequences. Proteins 14 (1992) 16-28.

Sikorski, R.S. and Hieter, P.: A system of shuttle vectors and yeast host strain designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. Genetics 122 (1989) 19-27.

Sthrul, K.: Nucleotide sequence and transcriptional mapping of the yeast *pet56-his3-ded1* gene region. Nucl. Acids Res. 13 (1985) 8587-8601.

Sthrul, K., Cameron, J.R. and Davis, R.W.: Functional genetic expression of eukaryotic DNA in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 73 (1976) 1471-1475.

Verdier, J-M.: Regulatory DNA-binding proteins in yeast: an overview. Yeast 6 (1990) 271-297.

Weil, C., Bekler, G. and Reeve, J.: Structure and organization of the *hisA* gene of the thermophilic archaebacterium *Methanococcus thermolithotrophicus*. J. Bacteriol. 169 (1987) 4857-4859.

Weir, B. (1990) *Genetic Data Analysis*. (Sinauer Press, Sunderland), pp 377

Ycas, M.: On the earlier states of the biochemical system. J. Theoret. Biol. 44 (1974) 145-160.

Figure legends

Fig. 1. Restriction and genetic map of the *HIS6* locus in *S. cerevisiae* strain YNN282. Symbols: P, C, Sau3AI, recognition sites for *Pst*I, *Cla*I and *Sau*3AI restriction endonucleases, respectively. *RPB3* is the gene encoding the RNA polymerase II subunit. **Methods**: Plasmid pRD64 was obtained from a yeast genomic DNA library contained DNA of strain YNN282 α *TRP1Δ HIS3Δ200 URA3-52 LYS2-801$_a$ ADE2-1$_0$ GAL MAL CUP^R* partially digested with *Sau*3AI and cloned into the *Bam*HI site of the yeast-*E. coli* shuttle plasmid pRS316, harboring the *URA3*, *ARSH4-CEN6* yeast sequences (Sikorski and Hieter, 1989). Subclones pRD641 to pRD646, were obtained by cloning restriction fragments of different length derived from the insert contained in plasmid pRD64 into the plasmid vector pGEM7Zf(+) (Promega). Induction of competence and transformation of *E. coli* FB184 *hisA915* (Goldschmidt et al., 1970) cells with plasmid DNAs was carried out as described by Hanahan (1983). The *S. cerevisiae* strain 1437-8c *a HIS6 GAL7 SUC MAL TRP1 URA1 MET2 ADE6 LYS1* was transformed by the "colony" procedure according to Gietz et al. (1992). Plasmid loss experiments from yeast cells were carried out by growing His$^+$ trasformants for 15-20 generations on YPAD (yeast extract 10g/l, bacto-peptone 20 g/l, glucose 20 g/l, adenine 40 mg/ml); cells were then plated on YPAD, approximatively 100 cells/plate. Colonies were replicated on SD medium (Difco Yeast Nitrogen Base w/o aminoacids 6.7 g/l, glucose 20 g/l) plus the appropriate supplements (Rose et al., 1990) in the absence or in the presence of histidine (20 μg/ml). The presence of plasmids in His$^+$ and His$^-$ yeast cells was tested by colony hybridization (Rose et al., 1990) using as probe the bacterial moiety of plasmid pRS316.

All the other DNA manipulations were based on Sambrook et al. (1989).

Fig. 2. Nucleotide sequence of the 1521 bp *S. cerevisiae HIS6* gene and its flanking regions. The sequence spans from a *Sau*3AI site (position 1) to the *Cla*I site (position 1521). Amino acid are indicated by the ***single-letter code***; amino acids in bold or italic represent the translation of the *HIS6* and *RPB3* genes, respectively. Stop codons are indicated with an asterisck;. nt sequence underlined at positions 115-118: putative TATA box; bold lower case: hypothetical ABF1 binding site; nt in bold and italic (positions 303-306): putative mRNA initiation site; nt sequence in bold and underlined in the *HIS6-RPB3* intergenic region: site promoting mRNA terminus production **Methods:** DNA was sequenced by the method of Sanger et al. (1977), using the Sequenase^{Tm} 2.0 kit (US Biochemical, Cleveland, OH, USA); both dGTP and dITP wewre used to minimise band compression. Subclones for DNA sequencing were generated by cloning restriction fragments from plasmid pRD64 into the vector pGEM7Zf(+). Gaps in the sequence were filled by the use of *ad hoc* constructed oligodeoxyribonucleotides as primers. The sequence was determined on both strands, and all restriction sites used to subclone in the vector pGEM7Zf(+) were confirmed by overlapping sequencing. The sequence analysis was performed using the McVector program (IBI, New Haven, CT, USA). The nucleotide sequence presented in this paper has been assigned the GenBank/EMBL accession number X87341.

Fig. 3. Alignment of the amino acid sequences deduced from the *S. cerevisiae HIS6* gene and from the procaryotic *hisA* genes. The amino acids are indicated by the *single letter code*. Gaps were introduced for optimal alignment. Symbols under the Sce sequence indicate the position of at leats 80% of identical (stars) or similar (dots) amino acids (accepted substitutions: K-R, D-E, S-T, I-L-V-M, F-Y). The bottom line shows the consensus sequence of the phosphate-binding motif identified by Bork et al. (1995): h, mainly hydrophobic; p, mainly polar. Abbreviations: *Ab = Azospirillum brasilense, Ec = Escherichia coli, Hi= Haemophilus influenzae; Ll = Lactococcus lactis, Mt = Methanococcus thermolithotrophicus, Mva = Methanococcus vannielii, Mvo = Methanococcus voltae, Sce = Saccharomyces cerevisiae; Sco = Streptomyces coelicolor; Sty = Salmonella typhimurium.* **Methods:** Amino acid sequences were retrieved from the GenBank, EMBL and PIR databases. The *Clustal V* program was used for sequence alignments reported in Fig. 3-5 (Higgins and Sharp, 1988). For the structure of the histidine biosynthetic genes see legend of Table I.

Fig. 4. Alignment of the amino acid sequences deduced from the *S. cerevisiae HIS6* gene and from the *S. cerevisiae* 3'moiety of *HIS7* gene. The amino acids are indicated by the *single letter code*. Gaps were introduced for optimal alignment. Stars above the sequences indicate the position of identical or similar amino acids (accepted substitutions: K-R, D-E, S-T, I-L-V-M, F-Y). **Methods:** see legend of Fig. 3.

Fig. 5. Alignment of the amino acid sequences deduced from the 5'-terminal domain (His6/1) and the 3'-terminal domain (His6/2) of the gene *S. cerevisiae HIS6* gene. Gaps were introduced for optimal alignment. Stars above sequences indicate the position of identical or similar amino acids (accepted substitutions: K-R, D-E, S-T, I-L-V-M, F-Y). **Methods:** See legend of Fig. 3.

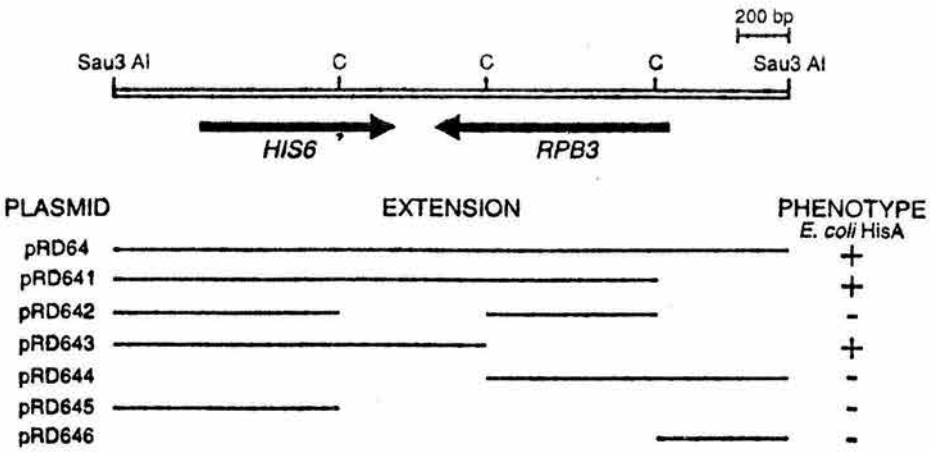
Fig. 6. Structure, organization and hypothetical evolutionary pathway of the extant *hisA*, *hisF* and *hisH* genes (see text, section e).

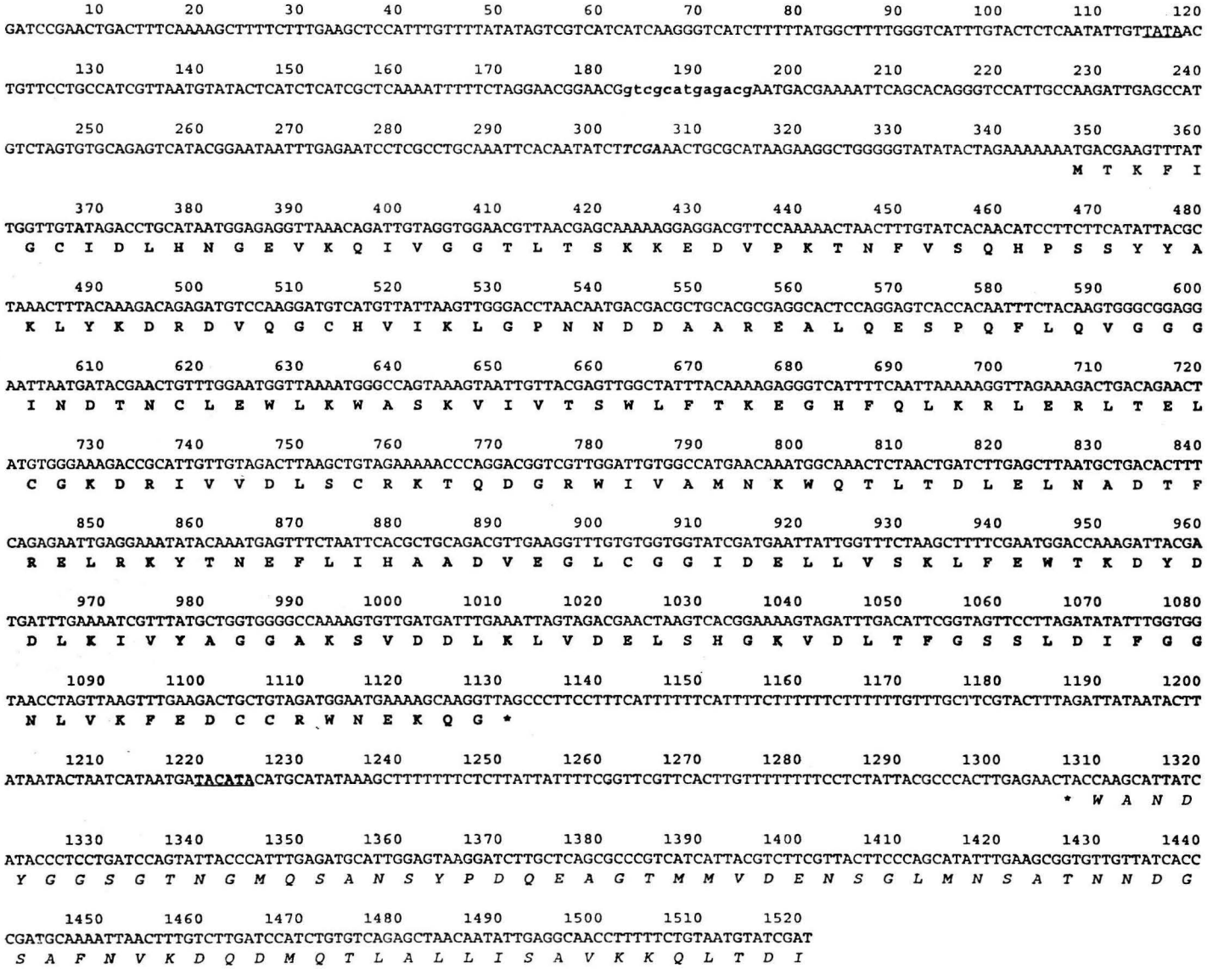
Fig. 7 . Unrooted phylogenetic tree constructed using the maximum likelihood method based on the HisA (upper) and HisF (lower) protein sequences from eucarya, archaea and bacteria. Abbreviations:*Kp = Klebsiella pneumoniae,* for the other ones see legend of Fig. 3. **Methods:** Phylogenetic analysis of the available sequences has been performed using the algorithms described by Weir (1990), Li and Graur (1991), and some additional ones prepared by the authors (available upon request). We have used *DNAML* by Felsenstein (1981) to compute maximum likelihood trees for the gene sequences, and the program *PROTML*, developed by Hasegawa and Adaki (personal communication), to obtain maximum likelihood trees for amino acid sequences. For the structure of the histidine biosynthetic genes see legend of Table I.

Table I. S_{AB} values (Fox et al., 1977) calculated for the deduced amino acid sequences of *hisA* and *hisF* gene products from different microorganisms. Upper right quadrant: identities, lower left: similarities.

	AbA	*AbF*	*EcA*	*EcF*	*HiA*	*HiF*	*LlA*	*LlF*	*MvoA*	*MtA*	*MvaA*	*Sce6*	*Sce7*	*ScoA*	
AbA		0.26	0.33	0.18	0.33	0.19	0.35	0.26	0.34	0.40	0.38	0.27	0.22	0.35	*AbA*
AbF	0.43		0.26	0.38	0.25	0.37	0.25	0.48	0.31	0.27	0.26	0.25	0.29	0.27	*AbF*
EcA	0.50	0.43		0.22	0.63	0.25	0.35	0.27	0.33	0.35	0.30	0.23	0.22	0.32	*EcA*
EcF	0.39	0.61	0.41		0.22	0.81	0.19	0.45	0.27	0.26	0.23	0.20	0.30	0.20	*EcF*
HiA	0.54	0.40	0.70	0.37		0.22	0.40	0.29	0.39	0.36	0.32	0.25	0.20	0.37	*HiA*
HiF	0.35	0.49	0.39	0.86	0.37		0.19	0.41	0.26	0.23	0.23	0.22	0.34	0.25	*HiF*
LlA	0.60	0.45	0.57	0.37	0.52	0.37		0.29	0.36	0.38	0.35	0.26	0.21	0.37	*LlA*
LlF	0.45	0.72	0.43	0.65	0.44	0.53	0.47		0.29	0.28	0.29	0.25	0.29	0.28	*LlF*
MvoA	0.57	0.54	0.50	0.50	0.53	0.40	0.60	0.61		0.60	0.67	0.26	0.25	0.26	*MvoA*
MtA	0.56	0.47	0.53	0.50	0.49	0.44	0.52	0.46	0.78		0.67	0.23	0.26	0.34	*MtA*
MvaA	0.55	0.45	0.54	0.47	0.45	0.41	0.56	0.62	0.81	0.79		0.28	0.25	0.30	*MvaA*
Sce6	0.35	0.34	0.36	0.34	0.32	0.32	0.35	0.36	0.40	0.40	0.42		0.20	0.25	*Sce6*
Sce7	0.35	0.54	0.37	0.37	0.31	0.48	0.38	0.57	0.40	0.41	0.38	0.32		0.21	*Sce7*
ScoA	0.50	0.47	0.47	0.32	0.49	0.39	0.52	0.44	0.51	0.52	0.48	0.36	0.36		*ScoA*
	AbA	*AbF*	*EcA*	*EcF*	*HiA*	*HiF*	*LlA*	*LlF*	*MvoA*	*MtA*	*MvaA*	*Sce6*	*Sce7*	*ScoA*	

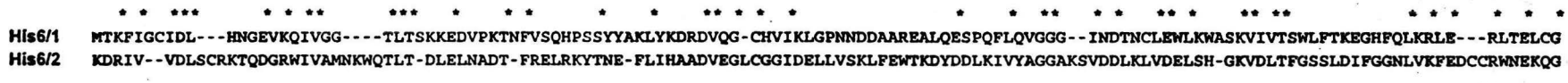
The amino acid sequences compared were deduced from: *A. brasilense hisA* (*AbA*) and *hisF* (*AbF*) (Fani et al., 1993); *E.coli hisA* (*EcA*) and *hisF* (*EcF*) (Carlomagno et al.,1988); *H. influenzae hisA* (*HiA*) and *hisF* (*HiF*) (Fleischmann et al., 1995); *L. lactis hisA* (*LlA*) and *hisF* (*LlF*) (Delorme et al., 1992).*M. vannielii hisA* (*MvaA*) and *M. voltae hisA* (*MvoA*) (Cue et al., 1985); *M. thermolithotrophicus hisA* (*MtA*) (Weil et al., 1987); *S. cerevisiae HIS6* (*Sce6*) (this article) and *HIS7* (*Sce7*) (Kluenzer et al., 1993); *S. coelicolor hisA (ScoA)* (Limauro et al., 1990).

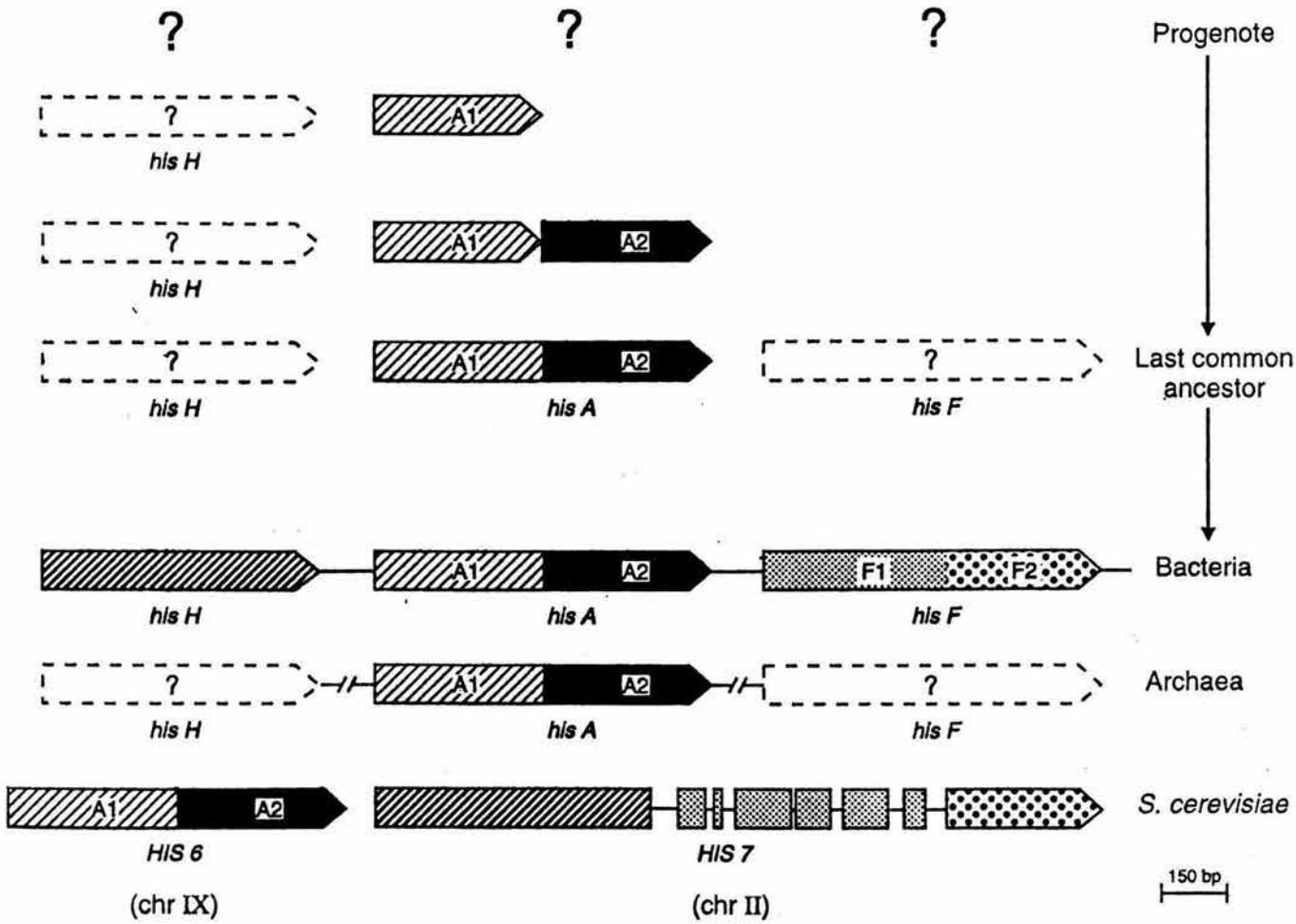

200 bp
Sau3 AI
C
C
C
Sau3 AI
HIS6
RPB3
PLASMID
EXTENSION
PHENOTYPE
E. coli HisA
pRD64
pRD641
pRD642
pRD643
pRD644
pRD645
pRD646
+
+
-
+
-
-
-


10 20 30 40 50 60 70 80 90 100 110 120
GATCCGAACTGACTTTCAAAAGCTTTTCTTTGAAGCTCCATTTGTTTTATATAGTCGTCATCATCAAGGGTCATCTTTTTATGGCTTTTGGGTCATTTGTACTCTCAATATTGTTATAAC
130 140 150 160 170 180 190 200 210 220 230 240
TGTTCCTGCCATCGTTAATGTATACTCATCTCATCGCTCAAAATTTTTCTAGGAACGGAACGgtcgcatgagacgAATGACGAAAATTCAGCACAGGGTCCATTGCCAAGATTGAGCCAT
250 260 270 280 290 300 310 320 330 340 350 360
GTCTAGTGTGCAGAGTCATACGGAATAATTTGAGAATCCTCGCCTGCAAATTCACAATATCTTCGAAACTGCGCATAAGAAGGCTGGGGGTATATACTAGAAAAAAATGACGAAGTTTAT
M T K F I
370 380 390 400 410 420 430 440 450 460 470 480
TGGTTGTATAGACCTGCATAATGGAGAGGTTAAACAGATTGTAGGTGGAACGTTAACGAGCAAAAAGGAGGACGTTCCAAAAACTAACTTTGTATCACAACATCCTTCTTCATATTACGC
G C I D L H N G E V K Q I V G G T L T S K K E D V P K T N F V S Q H P S S Y Y A
490 500 510 520 530 540 550 560 570 580 590 600
TAAACTTTACAAAGACAGAGATGTCCAAGGATGTCATGTTATTAAGTTGGGACCTAACAATGACGACGCTGCACGCGAGGCACTCCAGGAGTCACCACAATTTCTACAAGTGGGCGGAGG
K L Y K D R D V Q G C H V I K L G P N N D D A A R E A L Q E S P Q F L Q V G G G
610 620 630 640 650 660 670 680 690 700 710 720
AATTAATGATACGAACTGTTTGGAATGGTTAAAATGGGCCAGTAAAGTAATTGTTACGAGTTGGCTATTTACAAAAGAGGGTCATTTTCAATTAAAAAGGTTAGAAAGACTGACAGAACT
I N D T N C L E W L K W A S K V I V T S W L F T K E G H F Q L K R L E R L T E L
730 740 750 760 770 780 790 800 810 820 830 840
ATGTGGGAAAGACCGCATTGTTGTAGACTTAAGCTGTAGAAAAACCCAGGACGGTCGTTGGATTGTGGCCATGAACAAATGGCAAACTCTAACTGATCTTGAGCTTAATGCTGACACTTT
C G K D R I V V D L S C R K T Q D G R W I V A M N K W Q T L T D L E L N A D T F
850 860 870 880 890 900 910 920 930 940 950 960
CAGAGAATTGAGGAAATATACAAATGAGTTTCTAATTCACGCTGCAGACGTTGAAGGTTTGTGTGGTGGTATCGATGAATTATTGGTTTCTAAGCTTTTCGAATGGACCAAAGATTACGA
R E L R K Y T N E F L I H A A D V E G L C G G I D E L L V S K L F E W T K D Y D
970 980 990 1000 1010 1020 1030 1040 1050 1060 1070 1080
TGATTTGAAAATCGTTTATGCTGGTGGGGCCAAAAGTGTTGATGATTTGAAATTAGTAGACGAACTAAGTCACGGAAAAGTAGATTTGACATTCGGTAGTTCCTTAGATATATTTGGTGG
D L K I V Y A G G A K S V D D L K L V D E L S H G K V D L T F G S S L D I F G G
1090 1100 1110 1120 1130 1140 1150 1160 1170 1180 1190 1200
TAACCTAGTTAAGTTTGAAGACTGCTGTAGATGGAATGAAAAGCAAGGTTAGCCCTTCCTTTCATTTTTTCATTTTCTTTTTTCTTTTTTGTTTGCTTCGTACTTTAGATTATAATACTT
N L V K F E D C C R W N E K Q G *
1210 1220 1230 1240 1250 1260 1270 1280 1290 1300 1310 1320
ATAATACTAATCATAATGATACATACATGCATATAAAGCTTTTTTTCTCTTATTATTTTCGGTTCGTTCACTTGTTTTTTTTCCTCTATTACGCCCACTTGAGAACTACCAAGCATTATC
* W A N D
1330 1340 1350 1360 1370 1380 1390 1400 1410 1420 1430 1440
ATACCCTCCTGATCCAGTATTACCCATTTGAGATGCATTGGAGTAAGGATCTTGCTCAGCGCCCGTCATCATTACGTCTTCGTTACTTCCCAGCATATTTGAAGCGGTGTTGTTATCACC
Y G G S G T N G M Q S A N S Y P D Q E A G T M M V D E N S G L M N S A T N N D G
1450 1460 1470 1480 1490 1500 1510 1520
CGATGCAAAATTAACTTTGTCTTGATCCATCTGTGTCAGAGCTAACAATATTGAGGCAACCTTTTTCTGTAATGTATCGAT
S A F N V K D Q D M Q T L A L L I S A V K K Q L T D I

```
Hi        MKQSIIIPALDLINGQVVRLHQG-------DYAKQTTYSDNPIKQ-FDNY
Ll        M---KIIPAIDLQNGEAVRLYKG-------DYDKKTVYSKNPLEI-AQKF
Ab        M---IIYPAIDLKDGACVRLLRG-------EMSQATVFNTEPADQ-ARLF
Ec        M----IIPALDLIDGTVVRLHQG-------DYGKQRDYGNDPLPR-LQDY
Mt        M---IVIPAVDMKNGKCVQLIQG-------DPNKKHVELENPVEV-AEKW
Mva       M---LIIPAVDMKNKKCVQLIQG-------NPDKKHVELDNPPEI-AKKW
Mvo       M---YIIPAVDMKEGKCVQLIQG-------DPTKRHVEYDNPDEI-AKMW
Sco       MSKLELLPAVDVRDGQAVRLVHG-------ESGTETSY-GSPLEA-ALAW
Sty       M----IIPALDLIDGTVVRLHQG-------DYARQRDYGNDPLPR-LQDY
Sce       MTK--FIGCIDLHNGEVKQIVGGTLTSKKEDVPKTNFVSQHPSSYYAKLY
          *    .***.*.  *  * *  *        .          *

Hi        VRQGAKQLHLVDLTGAKNPQSRQTALIGKIVEATQCKVQVGGGIRTEQDV
Ll        ERMGATDLHLVDLDGAKIGQTRNLELVRKIKDETRLKIEIGGGIRDFDTV
Ab        ESQGFEWLHLVDLNGAFEGKPVNGKAVESILGAVTVPVQLGGGIRDLKTI
Ec        AAQGAEVLHLVDLTGAKDPAKRQIPLIKTLVAGVNVPVQVGGGVRTEEDV
Mt        VSEGAEMLHLVDLDGAIEGESVNRELIKEIIQTVNVPVQIGGGISVRCTE
Mva       VEQGAEMLHLVNLDGAINGKRVNDEFIEETIKNSGVPVQIGGGIRSVSDA
Mvo       IENGAEMLHLVDLDGAIDGERVNLPCVKKIIQESKVPVQMGGGIRTIQDV
Sco       QRSGAEWLHLVDLDAAF-GTGDNRALIAEVAQAMDIKVELSGGIRDDDTL
Sty       AAQGAGVLHLVDLTGAKDPAKRQIPLIKTLVAGVNVPVQVGGGVRTEEDV
Sce       KDRDVQGCHVIKLGPNNDDAAREA------LQESPQFLQVGGGINDTNCL
             **  ****** **       .  .  ..    . **.*****    .
                                                phPhhh-GGhp--pDh

Hi        A--DLLAV-GANRVVIGSTAVTHRSMVKNWFIKYGAEKFVLALDVNINAS
Ll        R--MYLEQIGVERVILGTAAVEKPDFLKELLIKYGPSRIVVGVDIR----
Ab        A--LWLE-KGVSRVILGTVALREPELVREARREF-PGKVAVGIDAR----
Ec        A--ALLEA-GVARVVVGSTAVKSQDMVKGWFERFGADALVVALDVRIDEQ
Mt        S--NLIE-IGAKRIILGTVAVENPDIVEEISKKVGKEKVMVALDAK----
Mva       L--YFIE-KGAEKVILGTVAIQNPKIVREISSIVGKEKVTVALDAK----
Mvo       E--ELVD-LGINKVIIGTVAVQNPDFVEQLAKKVGSDKIMVALDAK----
Sco       A--AALAT-GCTRVNLGTAALETPEWVAKVIAEHG-DKIAVGLDVR----
Sty       A--ALLKA-GVARVVIGSTAVKSPDVVKGWFERFGAQGVVLALDVRIDEQ
Sce       EWLKWASKVIVTSWLFTKEGHFQLKRLERLTELCGKDRIVVDLSCRKTQD
             .   *  .*..*. *.     *       *  .. * ** .
          h--ph----GA-hhhhGphhh------------G

Hi        GQKIVAISGWQEESGVLLETLIEDFQTVGLQQVLCTDISRDGTLTGSNIG
Ll        -EGFVSTSGWLEKTSLPYLSFLKKLERIGVKTTIITDISKDGTLTGPNFK
Ab        -EGYVAVAGWAETSTIKALDLALKFEDSGVAAIIYTDINRDGAMGGVNVE
Ec        GNKQVAVSGWQENSGVSLEQLVETYLPVGLKHVLCTDISRDGTLAGSNVS
Mt        -DGKVVIKGWKEKTKYTPVEMGKILEEKGAGSILFTNVNVEGLLTGMNVE
Mva       -DGKVLIKGWTEKTDYSPVQIGKILENMGAGSILFTNVDSEGLLEGINVL
Mvo       -DGKVVIKGWKEKTEYTPVQMGKILEEKGAGSILFTNVDSEGLLNGINIT
Sco       -GTTLRGRGWTRDGG-DLYETLDRLNKEGCARYVVTDIAKDGTLQGPNLE
Sty       GTKQVAVSGWQENSGVSLEQLVETYLPVGLKHVLCTDISRDATLAGSNVS
Sce       GRWIVAMNKWQTLTDLELNADTFRELRKYTNEFLIHAADVEGLCGGIDEL
             * . ** * .                *   .. * .  .* * * *.

Hi        LYQEICE---KYPPIQFQSSGGIGSLADIEALKGT---GVSGVIVGRAL-
Ll        LYDEISK---EN-SLNVIISGGVKDNSDI---QRATRSDFYGIIVGKAY-
Ab        ATSDLAF---HL-TTPVIASGAVSSIDDLIALKKEEDTGIQGVICGRAL-
Ec        LYEEVCA---RYPQVAFQSSGGIGDIDDVAALRGT---GVRGVIVGRAL-
Mt        PVETLVE---EL-EIPVIASGGVTTIEDLIKLK---AVGVEGVVVGSAI-
Mva       PTKELVD---NL-NIPIIASGGVTTVEDLLKFK---EIGVYAVVVGSAL-
Mvo       PTKELVD---NL-NIPIIASGGVTTIEDLIEFK---KIGVAGVVVGSAL-
Sco       LLKNVCA---A-TDRPVVASGGVSSLDDLRAIAGLVPAGVEGAIVGKAL-
Sty       LYEEVCA---RYPQIAFQSSGGIGDIDDDVAALRGT--GVRGMIVGRAL-
Sce       LVSKLFEWTKDYDDLKIVYAGGAKSVDDL-KLVDELSHGKVDLTFGSSLD
             .          .    ***.  ..*.          ** ...** **
                      phPhhh-GGhp--pDhhphhp----GA-hhhhGphh-

Hi        -LEGKFTLSEAIKCWQN-----------G
Ll        -YEGKINLEKEFR-------------NAN
Ab        -YDGRIDPKTALDLLSSVSVSGTPVTGKAG
Ec        -LEGKFTVKEAIACWQN-----------A
Mt        -YKNLIDLKEAIKACR-------------
Mva       -YKDMINLKDAILASK-------------
Mvo       -YKNNFKLQDAINTVNNFKI---------
Sco       -YAFAFTLEEALEA-------------TS
Sty       -LEGKFTVKEAIQCWQN-----------V
Sce       IPGGNLVKFEDCCRWNE----------KQG
             . . .*.
          h-G
```

```
     *** *  ****   *     * *                *    *       *        *     **  *     * *      ***  *   *  * ***** *
His6 MTKFIG-CIDLHNGE-----V-KQIV-----GGTLTSKKE-DVPKTN-FVSQHPSSYYAKL--YKDRD--VQGCHVIKLGPNNDDAAREALQESPQFLQVGGGINDTNC-----
His7 LTRRIIACLDVRTNDQGDLVVTKGDKYDVREKSDGKGVRNLGKPVQLAKYYQQGADEVTFLNITSFRDCPLKDTPMLEVLKQ---AAKT-VFV-P--LTVGGGIKDIVDVDGTK

        **   *   *   ** * *  **  *  * *  *      ** **   *   ***  *                  *   *       *     *    * ** *   *
His6 ---LE--WLKWAS---KVIV-TSWLFTKEGHFQL-KR------LERLTELCGKDRIVV--DLS-------------CRKTQD-GR------WIVAMNKWQTLT-DLELNADTFR
His7 IPALEVASLYFRSGADKVSIGTDAVYAAEKYYELGNRGDGTSPIETISKAYGAQAVVISVDPKRVYVNSQADTKNKVFETEYPGPNGEKYCWYQCTIKGGRESRDLGVWELTRA

           * **   * **   * *  **  * *  *    * **.  **  * *      *   * *  *  *  *      ** *       *
His6 ELRKYTNEFLIHAADVEGLCGGIDELLVSKLFEWTKDYDDLKIVYAGGAKSVDDLKLVDELSHGKVDLTFGSSLDIFGGN---LVKFEDCCRWNEKQG              261
His7 CEALGAGEILLNCIDKDGSNSGYDLELIEHVKDAVK----IPVIASSGAG-VPEHFEEAFLK-TRADACLGAGMFHRGEFTVNDVK-EYLLEHGLKVRMDEE          317
```


His6/1 MTKFIGCIDL---HNGEVKQIVGG----TLTSKKEDVPKTNFVSQHPSSYYAKLYKDRDVQG-CHVIKLGPNNDDAAREALQESPQFLQVGGG--INDTNCLEWLKWASKVIVTSWLFTKEGHFQLKRLE---RLTELCG
His6/2 KDRIV--VDLSCRKTQDGRWIVAMNKWQTLT-DLELNADT-FRELRKYTNE-FLIHAADVEGLCGGIDELLVSKLFEWTKDYDDLKIVYAGGAKSVDDLKLVDELSH-GKVDLTFGSSLDIFGGNLVKFEDCCRWNEKQG


?
?
?
Progenote
?
A1
his H
?
A1
A2
his H
?
A1
A2
?
his H
his A
his F
Last common ancestor
his H
A1
A2
F1
F2
his A
his F
Bacteria
?
A1
A2
?
his H
his A
his F
Archaea
A1
A2
HIS 6
HIS 7
(chr IX)
(chr II)
S. cerevisiae
150 bp

MOLECULAR EVOLUTION OF GLUTAMINE AMIDOTRANSFERASES: IMPLICATIONS FOR THE ORIGIN OF METABOLIC PATHWAYS

(título tentativo)

T.Mills[1], G. E. Fox[1], R. Fani2, J. I. Leguina[3], A. Lazcano[3], and J. Oró[1]

[1]Department of Biochemical and Biophysical Sciences
University of Houston
Houston, Texas, 77204-5934, USA

[2]Dipartamento di Biologia Animale e Genetica
Università degli Studi di Firenze
Via Romana, 17
I-50125 Firenze, ITALIA

[3]Facultad de Ciencias
Universidad Nacional Autónoma de México
Apdo. Postal 70-407
Cd. Universitaria, México D. F., 04510 MEXICO

Summary. In this paper we will suggest and demonstrate that the recent increase of published sequence data greatly facilitates the investigation of earlier suggestions concerning the evolution of metabolic pathways using well-known methodologies for phylogenetic investigation. Although it is unlikely that a single evolutionary mechanism will prove all-encompassing and be able to explain in its entirety the evolution of modern metabolism with all its complexity, we present here an original compilation and analysis of data on glutamine amidotransfersases which clearly demonstrate that the history of at least one small, yet prolific, catalyst, which has come to be present in numerous mainstream biosynthetic pathways, can indeed, be explained using the patchwork assembly hypothesis. Furthermore, the concordance between these, as well as other data, and the patchwork assembly hypothesis makes reasonable the suggestion that such a process may have played a major role in the diversification of early metabolic function.

Key Words: glutamine amidotransferase - patchwork assembly hypothesis

Introduction

A number of authors have undertaken a discussion of the evolutionary mechanisms behind the early development of metabolic pathways (Horowitz 1945; Waley 1969; Ycas 1974; Jensen 1976). Unfortunately, at the time that all of these thoughtful suggestions were put forth, very little relevant DNA and amino acid sequence data were available to aid in the determination of their legitimacy. Surprisingly, however, in spite of the ever increasing availability of pertinent DNA and amino acid sequence data, further discussion of this topic has been slow to develop and has surfaced only sparingly in the literature (e.g. Jensen 1985; Fothergill-Gilmore 1986). However, recent publications related to this topic have made it clear that the further development of this field is indeed possible, but relies to a great extent upon the elucidation of pairs or even familes of homologous enzymes serving roughly analogous catalytic functions, but in different metabolic pathways (Jensen 1992; Díaz-Villagómez et al. in press).

In light of these points, we have attmepted to review here, not only the present standing of the above hypotheses concerning the evolution of metabolic pathways, but also one specific data set, the glutmamine amidotransfersases, which we feel significantly enhances our perspective on the role of enzymes with common ancestry in the unfolding story of the evolution of metabolism. In the case of this family of enzymes, it has been demonstrated that a single ancestral catalytic activity has proliferated profusely, and is now found as a part of no less than seven distinct enzymes operating in at least five different biosynthetic pathways. Therefore, since a review of glutamine amidotranferases, especially one from an evolutionary perspective, is also long overdue, and may shed significant light upon the early evolution of biosynthetic metabolism, concurrently addressing both this family of enzymes, as well as their impact on our understanding of the evolution of metabolic pathways is indeed warranted.

Glutamine Amidotransferases

Glutamine amidotransferases (GATs) are a family of metabolically ubiquitous enzymes that participate in the catalytic transfer of the amide nitrogen of glutamine to a number of different substrates leading to the synthesis of various amino acids, purine and pyrimidine nucleotides, NAD, glucosamine, and antibiotics (Zalkin 1985). The generalized reaction displaying the metabolic importance of glutamine and glutamine amidotransferases is shown below in Figure 1. GATs typically function using two activities, a glutamine-dependent activity which binds glutamine and cleaves the amide nitrogen in the form of NH_3 (a glutaminase) and a second activity which utilizes the NH_3 to aminate a variety of substrates (a synthetase/synthase). Early on in the study of these enzymes it was suggested that these catalytic activities were brought together by the combination of a primitive ammonia-dependent enzyme and a glutaminase (Nagano et al. 1970; Li & Buchanan 1971; Trotta et al. 1971; Prusiner 1973). Structurally speaking, GATs characterized to date exhibit two distinct types of quaternary structure, an $\alpha_x\beta_x$ subunit composition, as well as an α_x form (Zalkin 1980; Zalkin et al. 1985; Hirai et al. 1987). As might be predicted from the description above, in the case of the $\alpha_x\beta_x$ configuration one subunit (β) binds glutamine and provides the glutamine amide cleavage activity, while the other subunit (α) is responsible for the NH_3-dependent reaction producing the aminated product. On the other hand, the identical subunits of the different oligomeric α_x GATs contain both catalytic activities on a single polypeptide. Interestingly, all of these enzymes have dual substrate specificity in that NH_3 can replace glutamine when asssayed *in vitro* (Buchanan 1973; Mantsala and Zalkin 1984a). Of further evolutionary importance is the fact that in some instances polypeptides possessing the amide cleavage activity are found fused not to their NH_3-dependent cohort, but to subsequent, yet unrelated enzymes of the pathway in which they serve (e.g. *E. coli trp* operon, see Yanofsky et al. 1981).

In terms of genetic organization the genes encoding GATs are found both as distinct genes and as constituents of operons and regulons, all of which have been shown to be under the control

of a collectively diverse array of regulatory mechanisms (e.g. Bae & Crawford 1990). Not surprisingly, the latter point is largely responsible for their being such widely studied genetic entities. GATs play a major role in the utilization of assimilated nitrogen and are found nearly ubiquitously in the biosynthetic pathways of all three extent cellular lineages. Furthermore, due to the somewhat notable fact that a number of the genes encoding glutamine amidotransferases have been shown to be homologous, not only in an orthologous sense (i.e. genes with a common origin functioning in the same pathway in other organisms), but also in a paralogous sense (i.e. genes with a common origin found as components of different pathways within the same organism), as a group, they may represent a useful repository of information relating to the evolution of biosynthetic pathways.

With the latter point in mind, the available glutamine amidotransferase DNA and derived amino acid sequence data were collected and analyzed in the hope that they might provide, not only a measure of the usefulness of the hypothetical models for the evolution of metabolic pathways mentioned above, but also, insight into the actual evolutionary history of the specific biosynthetic pathways containing this catalytic activity. Since the above mentioned orthologous/paralogous relationship seems to occur repeatedly across many mainstream biosynthetic pathways, but in this case, only for the subunits or domains of GATs which carry out the glutamine-dependent reaction, the majority of this paper will deal not with entire glutamine amidotransferases (i.e. both gln-dependent and NH_3-dependent activities), but primarily with the subunits or domains (henceforth GAT domains) able to free the amide nitrogen from glutamine.

The Two Sub-families of Glutamine Amidotransferases

Primary sequence analysis of the genes encoding various GAT domains has revealed that they are partitioned into two main sub-families, each possessing a distinct glutamine amide transfer domain (Walker et al. 1984; Weng & Zalkin 1987). Both of these sub-families contain

polypeptides which are approximately 200 amino acid residues in length, catalyze the nucleophilic cleavage of the amide nitrogen of glutamine, and possess regions or positions displaying striking sequence conservation. The two sub-families of GAT domains are the *purF*-types and the *trpG*-types (aka *purF*-related and *trpG*-related). GAT domains of the *purF*-type are homologs of amidophosphoribosylstransferase (glutamine PRPP amidotransferase) which catalyzes the first reaction in the *de novo* pathway for purine nucleotide biosynthesis (Weng & Zalkin 1987). In bacterial species, the genetic locus coding for this enzyme is known as *purF*, hence the name *purF*-type GAT domains. The *trpG*-type sub-family of GAT domains is exemplified by the small subunit of anthranilate synthase which, together with the large subunit, catalyzes the first step in the *de novo* synthesis of tryptophan (Weng & Zalkin 1987). The genetic locus encoding this enzyme is known as *trpG*, thus, its gene product, as well as its numerous homologs from other biosynthetic pathways are known as *trpG*-type GATs.

The genes within each sub-family have been shown to be homologous to the other genes within that sub-family, while comparisons between the two sub-families have revealed that the two groups have either evolved beyond recognition or are not likely to be of common ancestry (Tso et al. 1982a; Zalkin et al. 1985). Therefore, each sub-family may have arisen from a different ancestral enzyme, possibly a glutaminase (Nagano et al. 1970; Trotta et al. 1971) or an aminoacyl-tRNA synthetase (Di Guilio 1993). Of equal or greater importance is the fact that genes encoding homologous GAT domains are found repeatedly in a single genome, presumably as a result of numerous paralogous gene duplications. In many cases the gene encoding a distinct GAT domain which is used in a particular pathway, is found grouped together with many of the other genes encoding catalytic activities also used in that same pathway. Interestingly, in a few instances the product of a single gene encoding a GAT domain has been shown to function in two pathways simultaneously (e.g. Slock et al. 1990). Of the two sub-families of GAT domains, the *trpG*-types are the most well studied and will be discussed first below.

TrpG-type GAT domains

As mentioned above, the *trpG*-type GAT domains are all homologs of the small subunit of anthranilate synthase (*trpG*). These include the closely related small subunit of *p*-aminobenzoate synthase (Kaplan & Nichols 1983), as well as the glutamine amidotransferase domains or subunits found as parts of GMP synthetase (Zalkin et al. 1985), CTP synthetase (Weng et al. 1986), imidazole glycerol phosphate synthase (Carlomagno et al. 1988), carbamyl phosphate synthetase (Nyunoya & Lusty 1984), and the purine biosynthesis enzyme formylglycinamidine ribonucleotide synthetase (Ebbole & Zalkin 1987). Additionally, the existence of yet another pathway utilizing the amide nitrogen of glutamine using a second anthranilate synthase enzyme has been proposed to exist in the fluorescent pseudomonads (Essar et al. 1990a-c). While this pathway is itself quite hypothetical as this point, the genes for the second anthranilate synthase are very real and their discovery has given rise to some interesting suggestions concerning the evolutionary history of the anthranilate synthase genes in these organisms (Crawford & Milkman 1991). A summary of the reactions catalyzed by all of the enzymes containing a *trpG*-type subunit or domain can be found in Table 1. We will return shortly to a brief discussion of each of these GAT domains containing enzymes in turn, but first, we will discuss the general characteristics of *trpG*-type GAT domains.

In terms of overall structure, like GATs in general, the enzymes of the *trpG*-type subfamily of GAT domains exist as components of enzymes displaying a variety of subunit configurations (e.g. $\alpha\beta$, α_2 etc.). As shown schematically in Figure 2, *trpG*-type GAT domains are characterized at the amino acid sequence level by three highly conserved regions (designated I, II, and III) containing stretches of approximately five nearly invariant amino acids. Based on the numbering scheme corresponding to the amino acid sequence of *E. coli trpG* (Yanofsky et al. 1981), region I is centered at the glycine at position 57 and has the consensus sequence -S-P-G-P-. Region II is centered roughly at the cysteine residue in position 84 in the *E. coli* sequence and has

the consensus sequence -G-V-C-L-G-H-Q-, while region III has a consensus sequence of -Q-F-H-P-E- and is centered at the histidine residue in position 170.

As would be anticipated for highly conserved regions a number of the above residues have been demonstrated to play a role at the active site of *trpG*-type GAT domains. Zalkin and coworkers have investigated the mechanism of anthranilate synthase using a variety of techniques including affinity labeling via chemical modification (e.g. Tso et al. 1980; Tso & Zalkin 1981) and site-directed mutagenesis (Paluh et al. 1985; Amuro et al. 1985). As a result of these and numerous other works they have repeatedly concluded that all GAT domains utilize an active site cysteine in glutamine amide transfer, namely, Cys^{84} of region II. In a study of another critical residue, site-directed mutagenesis was used in combination with affinity labeling, to show that the invariant His^{170} of region III is also a likely participant at the GAT domain active site (Amuro et al. 1985). Amuro et al. (1985) have proposed a hypothetical reaction sequence for glutamine amide transfer in which His^{170} serves as a general base to increase the nucleophilicity of Cys^{84} through deprotonation of its sulfhydryl group. It is suggested, that this would subsequently facilitate the nucleophilic attack by the now electron-rich sulfide of cysteine on the carbonyl of glutamine, resulting in the formation of a covalent glutaminyl thioester intermediate leading to amide cleavage and transfer. This activity, when combined with the different NH_3-dependent synthetic capabilities, accounts for the observed phenotypes of all but one sub-group of the glutamine amidotransferase enzymes containing *trpG*-type GAT domains, the exception being mammalian carbamyl phosphate synthetase I which reacts directly with ammonia instead of glutamine to remove excess ammonia from the liver and small intestine (Nyunoya et al. 1985). As reviewed quite succinctly by Miran et al. (1991), the works of Zalkin and coworkers described above have been both preceded and corroborated by the works of numerous other researchers investigating anthranilate synthase, as well as other *trpG*-type GAT containing enzymes such as carbamyl phosphate synthetase (e.g. Miran et al. 1991) and FGAM synthetase (Schendel et al. 1989).

Presently there are more than 50 DNA and derived amino acid sequences available for *trpG*-type GATs. These sequences originate from all three of the extant cellular lineages and function as parts of as many as eight different enzymes in at least six, and possibly seven different pathways. Here we present the results of our compilation and analysis of the numerous *trpG*-type GAT domain data.

Anthranilate Synthase and PABA Synthase

The genes of the tryptophan biosynthetic pathway are collectively some of the most well-studied genetic entities known (for review see Crawford 1989). Of the more than 50 *trpG*-type GAT domain sequences now available, almost half are derived from the small subunit of anthranilate synthase (*trpG*), and its close *relative*, the small subunit of p-aminobenzoate synthetase (*pabA*), which together with the large subunit α, encoded by *pabB*, carries out the first step of folic acid biosynthesis in bacteria (Green et al. 1992). All of the well-studied enzymes encoded, in part, by these three particular genes, have been shown or inferred to be of the $\alpha_x\beta_x$ variety of subunit configuration, with the β being the subunit providing the GAT domain. Additionally, in some instances these GAT domain genes are found fused to other genes which encode enzymes catalyzing subsequent independent reactions in the same pathway , thus producing multifunctional proteins. A summary of these characteristics, shown together with the 30 corresponding GAT domain sequences from anthranilate synthase and PABA synthase, is presented in Table 2. Currently, there are 23 anthranilate synthase β or *trpG* DNA sequences available, 12 being eubacterial in origin, 6 being eukaryotic (fungal) in origin, and 3 originating from the archaebacteria. Of the 6 PABA synthase or *pabA* sequences published, all are eubacterial, as is the lone *phnB* sequence from the anthranilate synthase involved in the proposed phenazine pathway. It is noteworthy, however, that as a group, these sequences are derived from organisms from all three of the extant cellular lineages.

Alignment followed by metric analysis of a majority of the sequences in this rather large sub-group of the *trpG*-type GAT domains has been presented recently by Crawford (1989), and therefore, will not be discussed in detail here. However, this sub-set of data does provide some interesting insights into the evolution of biosynthetic pathways via gene duplication and recruitment. For example, the distance data clearly indicate that the *trpG* genes of the fluorescent pseudomonads are more closely related to the folic acid pathway *pabA* genes than to the *trpG* genes of other organisms. Crawford and Milkman (1991) have postulated that such a relationship could arise if an ancestral gene duplication occurred and was followed by sequence divergence, resulting in two separate enzymatic activities (e.g. PABA synthase and anthranilate synthase). They then suggest that, subsequently, along one line of descent, one of the duplicated genes became involved in one pathway, say tryptophan biosynthesis, and the other duplicated gene became specialized for another pathway, the folate pathway for example. At the same time, along another line of descent, the genes that in the first line of descent became specialized for the tryptophan pathway, might become specialized for folate synthesis and *vice versa.* This hypothesis seems to be a reasonable explanation for the clustering of the *Pseudomonas trpG* genes with the *pabA* genes of *E. coli* and other organisms. Furthermore, this phenomenon could in fact help to explain the diversification of a number of biosynthetic pathways.

If we add to this scenario the likely random occurrence of the deletion of one of the paralogous genes, we then arrive at a possible explanation for the existence of what are known as amphibolic enzymes. In the recent literature the term amphibolic has been used to describe enzymes that function in two pathways simultaneously. For example, as eluded to earlier, one of the *B. subtilis* genes encoding a GAT domain is clustered with genes for the first two steps of the folic acid pathway (Slock et al. 1990). Not surprisingly, the gene product functions in folic acid biosynthesis. However, this same product also has been shown to function in tryptophan biosynthesis in a manner analogous to the *trpG* gene product. (It should be noted, however, that amphibolic as originally coined was meant to describe an enzyme that could function both in

anabolic and catabolic metabolism. This is clearly not the case here and in other cases of similar use). As will be discussed below, this crossover of enzymatic activities between pathways has a direct bearing upon our models of the early evolution of biosynthetic pathways. Having now hinted at the importance of the paralogous and orthologous evolution of GAT domains, we will as briefly as possible describe other GAT domain containing enzymes and their previously uncompiled amino acid sequences.

Imidazole Glyerol Phosphate Synthase

The fifth step in the *de novo* biosynthesis of histidine involves the amidation of 5-amino-4-imidazolylcarboxamide to imidazole glycerol phosphate and aminoimidazole-4-carboxamide ribonucleotide (AICAR) (Table 1). Although it has not been completely characterized, the enzyme which is assumed to carry out this reaction is encoded by two genes, *hisF* and *hisH* (Pons et al. 1988). Therefore, as was the case above with anthranilate synthase and its closest relatives, the catalytic activities associated with this enzyme are believed to arise from two separate subunits (i.e. $\alpha_x\beta_x$ subunit configuration). *HisF* is thought to encode a cyclase which catalyzes the formation of the characteristic ring structure of histidine, while the *hisH* gene has been shown to be homologous to *trpG*-type GAT domains, and thus, is thought to provide the GAT activity. Hence, in line with other enzymes discussed above, we have tentatively assigned the eubacterial *hisF* product to be the "α" subunit and the *hisH* product to be the "β" subunit. At present, this enzyme is the only known case where a *trpG*-type GAT domain is used in conjunction with an activity other than a synthase or a synthetase. Accordingly, this enzyme could also be designated imidazole glycerol phosphate cyclase. Subsequent to this reaction, imidazole glycerol phosphate is further converted to histidine, while AICAR is an intermediate in the biosynthesis of purines. As shown in Table 3, all five of the available DNA sequences encoding the GAT domain of this enzyme come from eubacterial sources.

FGAM Synthetase

FGAM synthetase (aka FGAR amidotransferase) is an enzyme that catalyzes the amidation of FGAR (phosphoribosylformylglycinamide) to FGAM (phosphoribosyl-formylglycinamidine) in the fifth step in the *de novo* synthesis of purine nucleotides. Depending on the organism in which it is found, the GAT domain in this enzyme has been shown to exist in either the $\alpha\beta$ or the α subunit configuration, the latter having both catalytic activities on the same polypeptide. In the case of *B. subtilis*, the products of two genes, *purQ* and *purL*, are part of a twelve gene cluster from *Bacillus subtilis* encoding the nine enzymes of purine biosynthesis, and have been inferred to be responsible for encoding the catalytic activities of this enzyme using the $\alpha\beta$ subunit configuration (Ebbole & Zalkin 1987). However, in *E. coli* homologs of these two subunits are combined into a single polypeptide encoded by *pur(L)Q* with the GAT domain occupying the C-terminus (Schendel et al. 1989). Based on sequence comparison alone, it is clear that both forms of FGAM synthetase contain GAT domains. Additionally, Schendel et al. (1989) have presented compelling evidence supporting the proposed thioester glutamyl intermediate in the cleavage of the amide nitrogen from glutamine mentioned earlier.

GMP Synthetase

GMP synthetase catalyzes the synthesis of guanosine monophosphate from xanthosine monophosphate, ATP, and glutamine in a reaction that also implements a GAT domain (Table 1). In *E. coli*, the enzyme has been shown to exist as a dimer of identical subunits (α_2), both encoded by the *guaA* locus (Tiedeman et al. 1985). Each individual subunit provides on a single polypeptide both of the catalytic activities (glutaminase and synthetase) characteristic of glutamine amidotransferases. Nucleotide sequence data from *E. coli guaA* has indicated that the N-terminal 200 residues of the polypeptide are responsible for the glutaminase activity, and thus, comprise a GAT domain (Zalkin et al. 1985). As mentioned earlier, the glutaminase, and the synthetase

activities are believed to have come together as the result of an ancient gene fusion event involving a primordial GAT domain in both this case, and in the *E. coli* version of FGAM synthetase mentioned above (Zalkin & Truitt 1977; Zalkin et al. 1985). In addition to the *E. coli guaA* DNA sequence (Tiedeman et al. 1985), the *guaA* DNA sequences from *B. subtilis* (Mantsala & Zalkin 1992) and from the eukaryotic slime mold *Dictyostelium dicoideum* (Van Lookeren Campagne et al. 1991) have also been reported.

CTP Synthetase

CTP synthetase catalyzes the final reaction in the *de novo* synthesis of the pyrimidine nucleotide CTP. As shown in Table 1., this synthetase converts UTP, ATP, and glutamine to CTP, ADP, P_i, and glutamate. In *E. coli* it has been shown to exist as a tetramer of identical subunits (α_4) (Koshland & Levitski 1974). As is the case for GMP synthetase, the glutaminase and synthetase activities are provided together on a single peptide. However, in this case the GAT domain of the enzyme occupies the 300 C-terminal residues of the polypeptide (Weng & Zalkin 1987). As shown in Table 3, there are only two reported DNA sequences for the *pyrG* locus which encodes this glutamine amidotransferase, both of which are eubacterial in origin. They originate from *E. coli* (Weng et al. 1986) and the nitrogen fixing bacterium *Azospirillum brasilense* (Zimmer & Hundeshagen, unpublished), both of which are gram-negative purple bacteria of the γ and α subdivisions respectively.

Carbamyl Phosphate Synthetase

The last enzyme to be mentioned that contains a *trpG*-type GAT domain is carbamyl phosphate synthetase. This enzyme is interesting in that it displays various subunit and genetic configurations, as well as activities. Carbamyl phosphate is an essential intermediate in the biosynthetic pathways leading to both arginine and the pyrimidine nucleotides. It is synthesized by

carbamyl phosphate synthetase (CPS) from bicarbonate ion, two MgATPs, and either glutamine or NH_3 (Table 1). The organization and functional roles of CPSs are quite varied and the classification of CPSs is often based on substrate specificity and the pathways in which they are used. For example, particular CPSs can be either glutamine-dependent or NH_3-dependent, they can function solely in either the pyrimidine or arginine pathways, or they might function in both simultaneously. Additionally, they can exist as single enzymes, or as parts of large multifunctional enzymes. Below we will summarize the most salient characteristics of CPSs, especially those related to the evolution of their respective GAT domains.

In the eubacterial world, a single CPS often produces the carbamyl phosphate that is used in both arginine and pyrimidine biosynthesis. Organisms displaying the use of a single CPS in two different pathways include *E. coli*, *S. typhimurium*, and *P. aeruginosa* (Cunin et al. 1986). As shown in Table 4, the available DNA sequence data, in addition to protein purifications, have led to the conclusion that CPSs of these organisms are of the $\alpha\beta$ structural variety. Not surprisingly, it has been shown that the smaller subunit (β) provides the GAT domain, while the larger subunit (α) provides the synthetase activity. Hence, these enzymes are known to be glutamine-dependent, but, as is the case for all of the other GAT implementing enzymes, the synthetase subunit can also catalyze *in vitro* the synthesis of carbamyl phosphate with ammonia rather than glutamine (Trotta et al. 1974). Due to its function in the pathways leading to both arginine and pyrimidine biosynthesis, here we designate this form of carbamyl phosphate synthetase to be CPS-AP after the convention of Paulus and Switzer (1979).

The only known example of a eubacterial organism that carries two distinct CPSs, one for each pathway, is *B. subtilis* (Paulus & Switzer 1979). These two forms of CPS are referred to as CPS-A and CPS-P, indicating the arginine and pyrimidine pathways of use respectively. The genes encoding the glutamine-dependent CPS-P have recently been sequenced and have led the authors to conclude that CPS-P also conforms to the $\alpha\beta$ subunit configuration, as well as to the

catalytic division of labor described above (Quinn et al. 1991). The genes of *B. subtilis* CPS-A have been mapped, but await further study.

Turning to the eukaryotic domain, the picture becomes more complicated. Unlike the majority of the studied eubacteria, and similar to *B. subtilis*, eukaryotes posses at least two CPSs, one for the synthesis of arginine and one for pyrimidines. In yeast and other fungi, arginine specific CPS-A and pyrimidine specific CPS-P are found in different subcellular compartments and are separately regulated. Yeast CPS-A catalyzes the same overall reaction and has the same subunit structure ($\alpha\beta$) as the eubacterial enzyme (Nyunoya & Lusty 1984). The DNA sequences for the unlinked genes encoding each of the subunits of yeast CPS-A have been previously reported (Nyunoya & Lusty 1984; Werner et al. 1985), as has the sequence encoding the small GAT subunit of the *Neurospora crassa* CPS-A (Orbach et al. 1990).

The arginine specific CPS of ureogenic animals is known as CPS-I and is found in the mitochondria of the liver and small intestine where it functions in ammonia detoxification by means of the urea cycle. Unlike the $\alpha\beta$ subunit configuration of the eubacterial CPS and the eukaryotic CPS-A, this form of carbamyl phosphate synthetase maintains the α_x subunit configuration (e.g. Raijman & Jones 1976). Previously it has been demonstrated that significant sequence similarity exists between the rat CPS-I gene and the combination of the adjacent *carA* and *carB* genes encoding the two subunits of the *E. coli* CPS-AP enzyme, as well as with the combination of the unlinked *CPA1* and *CPA2* genes of the yeast CPS-A enzyme (Nyunoya et al. 1985a). Accordingly, it was proposed that the α_x form of the enzyme arose via a gene fusion involving the ancestors of the genes encoding the individual subunits of the $\alpha\beta$ form of CPS (Nyunoya et al. 1985b).

Interestingly, CPS-I, which is a mitochondrial enzyme, cannot use glutamine and requires NH_3 directly as a substrate. At first sight this might seem to cast doubt on the possibility that a

domain within CPS-I was in fact, homologous to other GAT domains. In both the rat and human cDNA sequences of CPS-I (Haraguchi et al. 1991), the previously discussed catalytically active cysteine at position 84 has been substituted with serine, rendering it inactive to glutamine. In spite of this difference at the GAT active site, the sequence homology mentioned above confirms that this enzyme was, at least at one time, a true glutamine amidotransferase. Thus, CPS-I contains a defective, and therefore NH_3-dependent, GAT domain that is indeed homologous to the other arginine specific glutamine-dependent CPSs. Incidentally, this naturally occurring point mutation in the active site of CPS-I provides, quite possibly the strongest *in vivo* evidence to date, of the importance of Cys^{84} to the catalytic functionality of GAT domains.

Thus, it seems clear that a mutation of the catalytically active residue in the GAT domain of CPS-I, combined most probably with the evolution of an increased affinity for NH_3, has facilitated the enzyme's role in the removal of excess ammonia from mammals. This point is also evidenced by the significantly lower K_m for ammonia of CPS-I. Still another mitochondrial CPS exists which is known as CPS-III and also functions in urea synthesis in both invertebrate and vertebrate species (Casey & Anderson 1983). Unlike CPS-I, this enzyme does indeed use glutamine as a substrate and one might hypothesize that when the gene(s) encoding one of these enzymes is sequenced, a close relative to and possibly the nearest ancestor of CPS-I will have been uncovered, but one with a cysteine at position 84 instead of serine.

In relation to eukaryotic pyrimidine biosynthesis, the pathway specific enzyme is known as CPS-II. This enzyme is known to exist as part of a multifunctional protein containing up to four catalytic activities. For example, in yeast, the reactions of pyrimidine biosynthesis are carried out by the products of five independent genes. One of those genes, *URA2*, encodes both the CPS and aspartate transcarbamylase activities, as well as a non-functional dihydroorotase, which together exist as a fused protein. As expected, part of the CPS enzyme, namely the N-terminus, encodes a GAT domain which is homologous to the many other GAT domains mentioned to this point

(Souciet et al. 1989). Mammals maintain CPS-II together with the pyrimidine biosynthetic enzymes aspartate transcarbamylase and dihydroorotase as a single multifunctional protein called CAD (see Simmer et al. 1990). *Drosophila melanogaster* and *Dictostelium discoideum* also possess the same catalytic activities in a single multifunctional protein in the same order and orientation. There are a total of four DNA or cDNA sequences available encoding CPS-II and the GAT domain therein. As shown in Table 5, they originate from yeast (Souciet et al. 1989), Syrian hamster (Simmer et al. 1990), *Drosophila melanogaster* (Freund & Jarry 1987), and *Dictostelium discoideum* (Faure et al. 1989).

Analysis of *TrpG*-type GAT Domain Data Set

In total, we have described in brief detail, seven enzymes involved in six different biosynthetic pathways. There are total of 54 DNA sequences available which have been shown to encode an entire *trpG*-type GAT domain. As described below, these sequences have been compiled and analyzed together as a group here for the first time. Only 53 of the above sequences were used as part of the data set however. The *hisH* sequence from *Azospirillum brasilense* was omitted since the sequence corresponding to its catalytic site was questionable. Also, the sequence for CPS-II from *Drosophila melanogaster* (Freund & Jarry 1987) was modified after Simmer et al. (1990) to correct for possible experimental artifacts which caused the conserved sequence about the catalytic site to be published in an alternate reading frame.

The simple investigative approach used involved gathering and "hand-aligning" the GAT domain amino acid sequences based first on classification by enzyme type (e.g. all of the *hisH* sequences), and then by sub-family type (or all of the *trpG*-types together). Due to the highly conserved nature of these genes "hand" or "eye" alignments in all cases resulted in nearly identical, if not more biologically relevant, sequence alignments compared to those produced via the computer driven techniques discussed below (data not shown).

Qualitatively speaking, alignment of the three highly conserved GAT regions (I, II, and III) was quite straightforward, while the sequences between the regions led to less impressive, albeit easily discernible, sequence alignment and conservation. The alignments of the terminal portions of the sequences found between the N-terminus and the 5'-end of region I, as well as between the 3'-end of region III and the C-terminus, were less impressive. Thus, in order to include sequence alignments of only the highest confidence in the metric analysis, the sequences were truncated somewhat arbitrarily at the 5'-end of region I and the 3' end of region III, and the alignments of the terminal sequences were not included in further analysis.

The remaining sequences, (i.e. the sequences bounded by the highly conserved regions I and III) were then "unaligned" and fed into progressive sequence alignment programs of Feng and Doolittle (1987). This set of programs collectively performs multiple sequence alignment through the iterative use of the algorithm of Needleman and Wunsch (1970) and the Mutation Matrix of Dayhoff et al. (1978). The result is a multiple sequence aligment with corresponding pairwise distance scores which have been converted and normalized from similarty scores based on the equation

$$\text{Distance} = -\ln \frac{S_{real} - S_{rand}}{S_{ident} - S_{rand}} \times 100, \qquad \text{(Equation 1)}$$

where S_{real} is the alignment score itself; S_{rand} is the score obtained with random sequences of the same lengths and compositions as the analyzed sequences, and S_{ident} is the average score of the two sequences being compared when each is aligned to itself (Feng & Doolittle 1987). Next, if phylogenetic analysis is to be undertaken, the distance scores are used with the aid of a least-squares approach to determine the branching order and branch lengths for a tree based on the aligned sequences (Fitch & Margoliash 1967).

The progressively aligned *trpG*-type GAT domain sequences are shown in Figure 3. There are a total six invariant residues in the aligned portions of these 53 genes. Half of these residues are glycines, one in region I, and two in region II. It might be surmised that the invariant nature of glycine in the protein would most likely be that due to a structural role. Generally, the small size of glycine allows it to play a pivotal role in polypeptide turns or regions where packing might be important. Thus, it is usually considered reasonable to presume that the glycine plays a role in bringing another important residue (or type of residue) into "register" so that it may carry out its highly conserved function. For example, each of the two of the invariant glycines in region II are found two residues removed on either side of the active site cysteine. Interestingly, this cysteine, which is well known to be the catalytic center of the GAT domain, is conspicuous in its absence from the short list of invariant residues. However, as mentioned above, the only two sequences which do not contain a cysteine at this position are those from carbamyl phosphate synthetase I, an enzyme well known to be NH_3-dependent. The correlation, in this case, between the lack of the catalytically active cysteine, and the inability to utilize glutamine as a nitrogen donor, is indeed believed to be the cause of the NH_3 dependence of this enzyme. This fact serves the molecule well in its function in the removal of NH_3 from the ureotelic mammals. Obviously, the most parsimonious evolutionary explanation for the emergence of CPS-I is one in which the enzyme is a more recent evolutionary modification of a once functionally active GAT domain. Its occurrence only in mammals supports this conclusion. Lastly, in region III, the histidine believed to act as the general base in catalysis is totally conserved along with the proline and glutamic acid adjacent to it on its C-terminal side.

Careful inspection of the alignment reveals also that a number of residues are strongly or nearly conserved, especially with respect to amino acid character. Therefore, this figure would appear to be a reasonable approximation of the alignment of homologous positions within these domains. Accordingly, this alignment is biologically reasonable and supports the numerous suggestions that GAT domains share a common ancestral gene.

Although the mixture in the data set of paralogous and orthologous genes ~~makes~~ may make phylogenetic analysis inappropriate for the data set as a whole, phylogenetic analysis based on a sub-set of the data, say the *hisH* genes for example, should be possible. Unfortunately, the sparcity of data in most of the sub-sets of data limits the usefulness of such analysis. In the cases of the anthranilate synthase/PABA synthase and the carbamyl phosphate synthetase data sets however, the number of sequences available is more inviting to phylogenetic analysis. Although phylogenetic analysis was not the goal of this work, the elucidation of a phylogeny of organisms based on these genes that is consistent with accepted phylogenies would cast serious doubt upon the inevitable suggestion that this particular domain has proliferated due to lateral gene transfer. As mentioned above, the anthranilate synthase/PABA synthase data sub-set has eluded phylogenetic analysis to this point. However, this is probably due to sequence conservation above the window of usefulness for phylogenetic analysis, the duplicity of function of some of these genes (amphibolic enzymes), the different evolutionary fates for duplicated genes in different organisms as described for the pseudomonads, and lastly, one presumably isolated case of putative lateral gene transfer. On the other hand, although the data await further detailed analysis, preliminary phylogenetic studies of the carbamyl phosphate synthetase genes seems to point to an approximation of an expected phylogenetic pattern (data not shown), and thus, support the ancient nature of this catalyst.

Therefore, although the alignment does not lead directly to a phylogeny of organisms based on the 53 sequences studied, it does allow the determination of the approximate relative relationship between each of the paralogous genes. As shown in Table 6, the different roles of GAT domains clearly show themselves when one computes the average distance (based on the distances calculated as described above) between the sequences of each sub-set of the data. Each sub-set of data or more correctly, each set of orthologous genes, clearly clusters closely together with the other genes of the same type.

Seven Paralogous TrpG-type GAT Domain Genes of E. coli

This entire study is significant in that it attempts to focus not on orthologous genes and phylogenetic trees, but rather on paralogous genes and what knowledge they might impart concerning the evolution of biosynthetic pathways. With the latter point in mind, and in an attempt to overcome the difficulties involved in comparing orthologous and paralogous genes simultaneously, we have compared the seven paralogous GAT domain genes from *E. coli*. Ironically, this approach will take on the familiar flavor of a phylogenetic study based on orthologous genes.

The *trpG*-type *E. coli* data set is merely a sub-set of the data set whose alignment is shown in Figure 3. Using the same methods described above for the larger data set, these genes were aligned and compared. In this case however, effort was made to "force" the resulting alignments to conform to the alignment of the larger data set. This seems appropriate if one considers the amount of information relative to the alignment of the sequences which is lost by using the smaller data set. For example, regions of local homology and conservation in one sub-set, say the *hisH* sub-set, might align nicely with a similar region in another sub-set, *trpG* for example. However, when the cumulative effect of the larger sample size upon the alignment is lost, such as when only the *E. coli* genes are used by themselves, the relationship between some regions which aligned rather well with the large data set are blurred when the smaller number of sequences were used. Therefore, in order to use what we believed to be the most biologically relevant, as opposed to mathematically optimized, alignment, the *E. coli* genes were forced, as much as possible, to conform to the alignment based on the larger data set. In retrospect it is clear that no matter which of the two types of alignments were used, the resulting relative relationships between the *E. coli* genes were unchanged and therefore did not depend upon the forcing of the alignments.

The distance scores calculated for the best alignment of *E. coli* GAT genes are shown in Table 7. We have imposed upon a phylogenetic method to derive from these distances what might tentatively be called a *paralogous gene tree* based on these genes. The dendogram based on the distance scores from the most acceptable alignment is shown in Figure 4. This is definitely not meant to be a phylogenetic tree. It merely demonstrates graphically, not only the relative distances between these paralogous genes, but the possibility that through judicious use of current methods or modifications thereof, one might be able to deduce the evolutionary history of the biosynthetic pathways. The first and foremost problem is that of determining the root or an outgroup of such a tree. More to the point, it is well worth considering the question of weather or not one can indeed "root a tree of pathways" and thus justify their existence. Regrettably, these questions are beyond the scope of this review and await futher attention. What can be concluded here is that comparison of paralogous genes does indeed provide insight useful for the consideration of previously proposed models for the evolution of biosynthetic pathways. Specifically, as will be discussed below, the data compiled and presented together here for the first time supports the patchwork theory for the development of catalytic biosynthetic diversity which maintains that metabolic pathways were formed through the duplication and recruitment of genes encoding globally useful catalytic domains. Another data set which appears also to support this theory is that of the *purF*-type GAT domains.

PurF-type GAT domains

The *purF*-type GAT domains represent a second group of homologous genes that, like the *trpG*-types, transfer the amide nitrogen of glutamine to variety of substrates. As a group, these enzymes also contain a combination of both paralogous and orthologous genes. As presented earlier, *purF*-type GAT domains are homologs of amidophosphoribosylstransferase (glutamine PRPP amidotransferase) which catalyzes the first reaction in the *de novo* synthesis of purine

nucleotides. In bacterial species, the genetic locus coding for this enzyme is known as *purF*. Therefore, this sub-family is known as the *purF*-type GAT domains.

Although the *purF*-type and *trpG*-type GAT domains do not appear to share common ancestry, their proposed catalytic mechanisms are very similar. As demonstrated by affinity labeling (Nagano et al. 1970; Tso et al. 1982a; Vollmer et al 1983; Badet et al. 1987) and site-directed mutagenesis (Mantsala & Zalkin 1984a), the *purF*-type sub-family of GAT domains, utilize a N-terminal active site cysteine residue which forms a covalent glutaminyl thioester intermediate leading to a nucleophilic amide cleavage. Although the glutaminyl intermediate should be familiar from the discussion of the *trpG*-types, the location of the active site cysteine, being N-terminal in the case of *purF*-types is however one notable difference. More recently, on the basis of additional site-directed mutagenesis, as well as deletion experiments (Mei & Zalkin 1989; Mei & Zalkin 1990), Zalkin and coworkers have put forth a model of *purF*-type GAT domains in which they are proposed to contain between 194 and 200 amino acid residues (as in *trpG*-types), and implement a Cysteine^{1}-Histidine^{101}-Aspartate^{29} catalytic triad in carrying out the function characteristic of these enzymatic domains. Mei and Zalkin (1989) have suggested that His^{101} functions to increase the nucleophilicity of Cys^{1}, which is used to form the covalent cysteinyl-glutamine tetrahedral intermediate. This is strikingly similar to the proposed roles of the conserved cysteine and histidine of *trpG*-type domains. Asp^{29} is proposed to act either to increase the nucleophilicity of a water molecule or might itself act as a nucleophile. In either case Asp^{29} is thought to assist in the removal of the glutamic acid from Cys^{1} remaining after the removal of the amide nitrogen from glutamine. For a more detailed discussion of the proposed mechanism of amide transfer, as well as the other conserved residues presumed to play structural roles such as Arg^{26} and Gly^{27}, consult Mei and Zalkin (1989).

Known homologs of amidophosphoribosyltransferase (*purF*) include both asparagine synthetase (Scofield et al. 1990) and glucosamine synthase (Walker et al. 1984). There are a total

total of fourteen sequences available for *purF*-type GAT domains, four function as amidophosphoribosyltransferases, seven as asparagine synthetases, and three as glucosamine synthases. The specific reactions catalyzed by each of these enzymes can be seen in Table 8. As mentioned above, amidophosphoribosyltransferase catalyzes the first step in the *de novo* synthesis of purines in which glutamine and 5-phosphoribosyl-1-pyrophosphate are converted to glutamate and phosphoribosylamine. Of the four sequences available for amidophosphoribosyltransferases, two are of eubacterial origin, *Escherichia coli* (Tso et al. 1982b) and *Bacillus subtilis* (Makaroff et al. 1983), while two are of eukaryotic origin, chicken (Zhou et al. 1990) and yeast (Mantsala & Zalkin, 1984b).

Asparagine synthetase, which is encoded by the *asnB* locus in *E. coli*, is an ATP dependent GAT which catalyzes the conversion of aspartate and glutamine to asparagine and glutamate. As in the case for amidophosphoribosyltransferases, the distinct GAT domain of this enzyme occupies the N-terminal of the polypeptide with a terminal cysteine presumably being the most important catalytic residue. Eukaryotic sequences available include those from human (Andrulis et al. 1987; Greco et al. 1989), Chinese hamster (Andrulis et al. 1989), Syrian hamster (Gong & Basilico 1990), as well as those from asparagus (Davies et al., unpublished) and pea plants (Tsai & Coruzzi 1990). Interestingly, in the case of the pea plant, two copies of the gene encoding asparagine synthetase have been discovered. These genes show roughly 86% identity at the amino acid level. *E. coli* is the only published eubacterial *asnB* sequence currently available (Scofield et al. 1990).

The last of the known *purF*-type homologs is glucosamine synthase which catalyzes the first reaction in hexosamine biosynthesis in which fructose-6-phosphate and glutamine are converted to glucosamine-6-phosphate and glutamate. In *E. coli* this N-terminal active GAT domain is part of the genetic locus named *glmS*. Two eubacterial glucosamine GAT domains have been sequenced in addition to *E. coli* (Walker et al. 1984). Both exist as one of a number of nodulation genes in the symbiotic nitrogen fixing bacteria *Rhizobium meliloti* (Baev et al. 1991)

and *Rhizobium leguminosarum* (Surin & Downie 1988). Although the latter two genes are known to encode glucosamine synthetase activity, their role in the nodulation process results in their loci being known as *nodM*. To date, no DNA sequences for eukaryotic glucosamine synthetase have been reported, and unlike the case for *trpG*-type GAT domains shown below, no archaebacterial sequences have been reported for any of the above variations of *purF*-type GAT domains. The available *purF*-type DNA sequences are summarized below in Table 9.

Analysis of *PurF*-type GAT Domain Data Set

The *purF*-type GAT domains were compiled and analyzed using the same methodology which was used for the *trpG*-types discussed above. The results of the progressive alignment of the fourteen sequences available are shown in Figure 5. Notable is the absolutely conserved N-terminal cysteine. However, in the case of His^{101} and Asp^{29} which, as discussed above, have both been implicated to play a significant role at the active site, these residues are not 100% conserved. In the case of Asp^{29}, all of the substitutions involved replacement by glutamic acid, which of course, can act in a manner similar to that of aspartic acid. The alanine replacement of histidine in the *E. coli* asparagine synthetase is more difficult to explain. Nevertheless, the overall sequence similarity, although less impressive when compared to that of the *trpG*-types, has led numerous authors to conclude that these genes are indeed derived from a common ancestor. Thus, this data set provides a second, albeit less striking, example of paralogous genes acting in different pathways. Accordingly, one can then conclude that this data set also supports the patchwork assembly theory for the evolution of biosynthetic pathways.

Although a "full-blown" comparison of these genes was performed, it can be summarized with little fanfare by stating that the *purF* genes appear to be most closely related to the genes encoding glucosamine synthetase (*glmS* and *nodM*), while the *asnB* genes of asparagine synthetase are the most distantly related to each of the other two genes (data not shown).

Unfortunately, the nature of this data set (i.e. the pathways from which the genes arise) is of little help in corroborating the branching order determined from the *trpG*-type data. Wishful thinking would have us believe that analysis of the genes of two sets of paralogous genes might provide some overlap between the data sets and thus the pathways. This might prove useful for rooting a tree of the pathways, and might at least corroborate the branching order, provided the genes were recruited more or less simultaneously and evolved at nearly the same rate. As will be brought out in the discussion section, there are other examples of paralogous genes acting in different pathways. However, such examples are supported by significantly less data and involve in most cases only two biosynthetic pathways. It is anticipated that accumulation of DNA and amino acid sequence data will provide more insight into the evolution of biosynthetic pathways. For example, NAD synthetase, encoded by *nadE* in *E. coli*, is a glutamine amidotransferase that awaits sequencing and classification. Hopefully, numerous other similar examples will come to light so that a broad picture of how metabolic function diversified will materialize.

Discussion

The analysis of both *trpG*-type and *purF*-type GAT domains clearly demonstrates that effective, globally useful catalysts, as well as the genes which encode them, have been subject to recruitment by many mainstream biosynthetic pathways leading to their broad dispercement in the genomes of extant organisms. As mentioned already, examples of this phenomenon are not limited to glutamine amidotransferase. Collectively, however, the GAT domain data sets are unique among these other examples due to their number and especially the wide variety of functional roles displayed by their members. It is also significant that glutamine amidotransferases serve pathways of fundamental importance to all living organisms. Although nearly all of these enzymes catalyze a similar reaction that removes the amide nitrogen from glutamine, they each do so in conjunction with a different "partner" of sorts which allows this catalytic activity to be broadly effective in

nitrogen metabolism. In essence, we believe this ability to serve the needs of multiple pathways is the heart of the patchwork theory of the evolution of biosynthetic pathways.

insert discussion

References

Abdelal AT, Bussey L, Vickers L (1983) Carbamylphosphate synthetase from Pseudomonas aeruginosa: subunit composition, kinetic analysis, and regulation. Eur J Biochem 129:697-702

Abdelal ATH, Ingraham JL (1975) Carbamylphosphate synthetase from Salmonella typhimurium. J Biol Chem 250:4410-4417

Adams RR, Royer T (1990) Complete genomic sequence encoding *trpC* form *Aspergillus niger* var. *awamori*. Nuc Acids Res 18:4931-4931

Amuro N, Paluh JL, Zalkin H (1985) Replacement by site-directed mutagenesis indicates a role for histidine 170 in the glutamine amide transfer function of anthranilate synthase. J Biol Chem 260:14844-14849

Andrulis IL, Chen J, Ray PN (1987) Isolation of human cDNA for asparagine synthetase and expression in Jenson rat sarcoma cells. Mol Cell Biol 7:2435-2443

Andrulis IL, Shotwell M, Evans-Blacker S, Zalkin H, Siminovitch L, Ray PN (1989) Fine structure analysis of the Chinese hamster *AS* gene encoding asparagine synthetase. Gene 80:75-85

Arhin FF, Vining LC (unpublished) *Streptomyces lividins p*-aminobenzoic acid synthase *pabB* and *pabA* genes. Genbank/EMBL DNA Sequence Database 1991, acc. no. m33811

Badet B, Vermoote P, Haumont PY, Lederer F, Le Gofic F (1987) Glucoseamine synthetase from *Escherichia coli*: purification, properties, and glutamine-utilizing site location. Biochemistry 26:1940-1948

Bae YM, Crawford IP (1990) The *Rhizobium meliloti trpE(G)* gene is regulated by attenuation, and its product, anathranilate synthase, is regulated by feedback inhibition. J Bacteriol 172:3318-3327

Bae YM, Holmgren E, Crawford IP (1989) *Rhizobium meliloti* anthranilate synthase gene: cloning, sequence, and expression in *Escherichia coli*. J Bacteriol 171:3471-3478

Baev N, Endre G, Petrovics G, Banfalvi Z, Kondorosi A (1991) Six nodulation genes of *nod* box locus 4 in *Rhizobium meliloti* are involved in nodulation signal production: *nodM* codes for D-glucosamine synthetase. Mol Gen Genet 228:113-124

Bardowski J, Ehrlich SD, Chopin A (1992) Tryptophan biosynthesis genes in *Lactococcus lactis* subsp. *lactis*. J Bacteriol 174:6563-6570

Buchanan JM (1973) The amidotransferases. Adv. Enzymol Relat Areas Mol Biol 38:1-39

Buchanan JM (1982) Covalent reaction of substrates and antimetabolites with formylglycinamide ribonucleotide amidotransferase. Methods Enzymol 87:76-84

Carlomagno MS, Chiariotti L, Alifano P, Nappo AG, Bruni CB (1988) Structure and function of the *Salmonella typhimurium* and *Escherchia coli* K-12 histidine operons. J Mol Biol 203:585-606

Casey CA, Anderson PM (1983) Glutamine- and *N*-acetyl-L-glutamate-dependent carbamoyl phosphate synthetase from *Micropterus salmoides*. J Biol Chem 258:8723-8732

Choi HT, Revuelta JL, Sadhu C, Jayaram M (1988) Structural organization of the *TRP1* gene of *Phycomyces blakesleeanus*: implications for evolutionary gene fusion in fungi. Gene 71:85-95

Crawford IP (1989) Evolution of a biosynthetic pathway: the tryptophan paradigm. Annu Rev Microbiol 43:567-600

Crawford IP, Han CY, Silverman M (1991) Sequence and features of the tryptophan operon of *Vibrio parahaemolyticus*. DNA Seqeunce 1:189-196

Crawford IP, Milkman R (1991) Orthologous and paralogous divergence, reticulate evolution, and lateral gene transfer in bacterial *trp* genes. In: Selander RK, Clark AG, Whittam TS (eds.) Evolution at the molecular level, Sinauer, Chicago, pp. 77-95

Cunin R, Glansdorff N, Pierard A, Stalon V (1986) Biosynthesis and metabolism of arginine in bacteria. Microbiol Rev 50:314-352

Dahoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change. In: Atlas of protein Ssequence and structure. vol. 5. suppl. 3. National Biomedical Research Foundation. Washington, D.C. pp 345-358

Davies KM, Hurst PL, Wollard DC, King GA (unpublished) Isolation and characterization of a cDNA clone for a harvest-induced asparagine synthetase from *Asparagus officianalis*. Genbank/EMBL DNA Sequence Database 1992, acc no x67958

Davis RH, Ristow JL, Hanson BA (1980) Carbamyl phosphate sythetase A of *Neurospora crassa*. J Bacteriol 141:144-155

Delorme C, Ehrlich SD, Renault P (1992) Histidine biosynthesis genes in *Lactococcus lactis* subsp. *lactis*. J Bacteriol 174:6571-6579

Di Guilio,M (1993) Origin of glutminyl-tRNA synthetase: an example of palimpsest? J Mol Evol 37 5-10

Díaz-Villagómez E, Lazcano A, Mills TM, Fox GE, Oró J (in press) The early evolution of metabolic pathways. Origins Life xx:xxx-xx

Ebbole DJ, Zalkin H (1987) Cloning and characterization of a 12-gene cluster from *Bacillus subtilis* encoding nine enzymes for de novo purine nucleotide synthesis. J Biol Chem 262:8274-8287

Essar DW, Eberly L, Crawford IP (1990b) Evolutionary differences in chromosomal locations of four early genes of tryptophan pathway in fluorescent pseudomonads: DNA sequences and characterization of *Pseudomonas putida trpE* and *trpGDC*. J Bacteriol 172:867-883

Essar DW, Eberly L, Hadero A, Crawford IP (1990c) Identification and characterization of genes for a second anthranilate synthase in *Pseudomonas aeruginosa*: Interchangeability of two anthranilate synthases and evolutionary implications. J Bacteriol 172:884-900

Essar DW, Eberly L, Han CY, Crawford IP (1990a) DNA sequences and characterization of four early genes of the tryptophan pathway in *Pseudomonas aeruginosa*. J Bacteriol 172:853-866

Fani R, Bazzicalupo M, Damiani G, Bianchi A, Schipani C, Sgaramella V, Polsinelli M (1989) Cloning of histidine genes of *Azosprillum brasilense*: organization of the *ABFH* gene cluster and nucleotide sequence of the *hisB* gene. Mol Gen Genet 216:224-229

Faure M, Camonis JH, Jacquet M (1989) Molecular characterization of a *Dictyostelium discoideum* gene encoding a multifunctional enzyme of the pyrimidine pathway. Eur J Biochem 179:345-358

Feng D-F, Doolittle RF (1987) Progressive sequence aligment as a prerequisite to correct phylogenetic trees. J Mol Evol 25:351-360

Fitch WM, Margoliash E (1967) Costruction of phylogenetic trees. Science 155:279-284

Freund, JN, Jarry BP (1987) The *rudimentary* gene of *Drosophila melanogaster* encodes four enzymic functions. J Mol Biol 193:1-13

Gong SS, Basilico C (1990) A mammalian temperature-sensitive mutation affecting G1 progression results from a single amino acid substitution in asparagine synthetase. Nuc Acids Res 18:3509-3513

Greco A, Gong SS, Ittmann M, Basilico C (1989) Organization and expression of the cell cycle gene, *ts11*, that encodes asparagine synthetase. Mol Cell Biol 9:2350-2359

Green J, Merkel W, Nichols B (1992) Characterization and sequence of *Escherichia coli pabC*, the gene encoding aminodeoxychorismate lyase, a pyridoxal phosphate-containing enzyme. J Bacteriol 174:5317-5323

Haraguchi Y, Uchino T, Takiguchi M, Endo F, Mori M, Matsuda I (1991) Cloning and sequence of a cDNA encoding human carbamyl phosphate synthetase I: molecular analysis of hyperammonemia. Gene 107:335-340

Hirai K, Matsuda Y, Nakagawa H (1987) Purification and characterization of GMP synthetase from Yoshida sarcoma ascites cells. J Biochem 102:893-902

Horowitz NH (1945) On the evolution of biochemical pathways. Proc Natl Acad Sci USA 31:153-157

Horowitz, NH (1965) The evolution of biochemical syntheses-restrospect and prospect. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 15-23

Horrowitz H, Van Arsdell J, Platt T (1983) Nucleotide sequence of the *trpD* and *trpC* genes of *Salmonella typhimurium*. J Mol Biol 169:775-797

Jensen RA (1976) Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409-425

Jensen RA, Byng GS (1982) The partitioning of bochemical pathways with isozyme systems. In: Rattazi MC, Scandios JG, Whitt GS (eds) Isosymes: current topics in biological and medical research. Alan R. Liss Co., New York, pp 143-174

Jensen, R.A. (1985) Biochemical pathways in prokaryotes can be traced backward through evolutionary time. Mol Biol Evol 2:92-108

Jensen, R.A. (1992) An emerging outline of the evolutionary history of aromatic amino acid biosynthesis. In: Mortlock RP (ed) The evolution of metabolic function. CRC Press, Boca Raton, p. 205-236

Kaplan JB, Goncharoff P, Seibold AM, Nichols BP (1984) Nucleotide sequence of the *Acinetobacter calcoaceticus trpGDC* gene cluster. Mol Biol Evol 1:456-472

Kaplan JB, Merkel WK, Nichols BP (1985) Evolution of glutamine amidotransferase genes: Nucleotide sequences of the *pabA* genes from *Salmonella typhimurium, Klebsiella aerogenes* and *Serratia marcescens.* J Mol Biol 183:327-340

Kaplan JB, Nichols BP (1983) Nucleotide sequence of *Escherichia coli pabA* and its evolutionary relationship to *trp(G)D.* J Mol Biol 168:451-468

Kilstrup M, Lu C-D, Abdelal A, Neuhard J (1988) Nucleotide sequence of the *carA* gene and regulation of the *carAB* operon in *Salmonella typhimurium.* Eur J Biochem 176:421-429

Koshland DE, Levitski A (1974) CTP synthetase and related enzymes. In: Boyer PD (ed) The enzymes. vol. X. Academic Press, New York, pp 539-559

Lam WL, Logan SM, Doolittle WF (1992) Genes for tryptophan biosynthesis in the halophilic archaebacterium *Haloferax volcanii*: the trpDFEG cluster. J Bacteriol 174:1694-1697

Levitzki A, Stallcup WB, Koshland DE (1971) Half-of-the-sites reactivity and conformational states of cytidine triphosphate synthetase. Biochem 10:3371-3378

Li HC, Buchanan JM (1971) Biosynthesis of purines. J Biol Chem 246:4713-4719

Limauro D, Avitabile A, Cappellano C, Puglia AM, Bruni CB (1990) Cloning and characterization of the histidine biosynthetic cluster of *Streptomyces coelicolor* A3(2). Gene 90:31-41

Makaroff CA, Zalkin H, Switzer RL, Vollmer SJ (1983) Cloning of the *Bacillus subtilis* glutamine phosphoribosylprophosphate amidotransferase gene in *Escherichia coli.* J Biol Chem 258:10586-10593

Mantsala P, Zalkin H (1984a) Glutamine amidotransferase function. J Biol Chem 259:14230-14236

Mantsala P, Zalkin H (1984b) Glutamine nucleotide sequence of *Saccharomyces cerevisiae ADE4* encoding phosphoribosylpyrophosphate amidotransferase. J Biol Chem 259:8478-8484

Mantsala P, Zalkin H (1992) Cloning and sequencing of *Bacillus subtilis purA* and *guaA* involved in the conversion of IMP to GMP. J Bacteriol 174:1883-1890

Matsui K, Sano K, Ohtsubo E. (1986) Complete nucleotide and deduced amino acid sequences of the *Brevibacterium lactofermentum* tryptophan operon. Nuc Acids Res 14:10113-10114

Mei B, Zalkin H (1989) A cysteine-histidine-aspartate catalytic triad is involved in glutamine amide transfer function in *purF*-type glutamine amidotransferases. J Biol Chem 264:16613-16619

Mei B, Zalkin H (1990) Amino-terminal deletions define a glutamine amide transfer domain in glutamine phosphoribosylpyrophosphate amidotransferase and other *purF*-type amidotransferases. J Bacteriol 172:3512-3514

Meile L, Stettler R, Banholzer R, Kotik M, Leisinger T (1991) Tryptophan gene cluster of *Methanobacterium thermoautotrophicum* Marbug: molecular cloning and nucleotide sequence of a putative *trpEGCFBAD* operon. J Bacteriol 173:5017-5023

Miran SG, Chang SH, Raushel FM (1991) Role of four conserved histidine residues in the amidotransferase domain of carbamoyl phosphate synthetase. Biochem 30:7901-7907

Mullaney EJ, Hamer JE, Roberti KA, Yelton MM, Timberlake WE (1985) Primary structure of the *trpC* gene from *Aspergillus nidulans*. Mol Gen Genet 199:37-45

Nagano H, Zalkin H, Henderson EJ (1970) The anthranilate synthetase-anthranilate-5-phosphoribosyl-pyrophosphate phosphoribosyltransferase aggregate. J Biol Chem 245:3810-3820

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443-453

Nichols BP, Miozzari GF, van Cleemput M, Bennett GN, Yanofsky C (1980) Nucleotide sequences of the *trpG* regions of *Escherichia coli*, *Shigella dysenteriae*, *Salmonella typhimurium* and *Serratia marcescens*. J Mol Biol 142:503-517

Nyunoya H, Broglie KE, Lusty CJ (1985b) The gene for carbamoyl-phosphate synthetase I was formed by fusion of an ancestral glutaminase gene and a synthetase gene. Proc Nat Acad Sci USA 82:2244-2246

Nyunoya H, Broglie KE, Widgren EE, Lusty CJ (1985a) Charecterization and derivation of the gene coding for mitochondreal carbamyl phosphate synthetase I of rat. J Biol Chem 260:9346-9356

Nyunoya H, Lusty CJ (1984) Sequence of the small subunit of yeast carbamyl phosphate synthetase and identification of its catalytic domain. J Biol Chem 259:9790-9798

Orbach MJ, Sachs MS, Yanofsky C (1990) The *Neurospora crassa arg-2* locus. J. Biol. Chem. 265: 10981-10987.

Paluh JL, Zalkin H, Betsch D, Weith HL (1985) Study of anthranilate synthase function by replacement of cysteine 84 using site-directed mutagenesis. J Biol Chem 260:1889-1894

Paulus TJ, Switzer RL (1979) Characterization of pyrimidine-repressible and arginine-repressible carbamyl phosphate synthetase from *Bacillus subtilis*. J Bacteriol 137:82-91

Penalva MA, Sánchez F (1987) The complete nucleotide sequence of the *trpC* gene from *Penicillium chrysogenum*. Nuc Acids Res 15:1874-1874

Pierard A, Schroter B (1978) Structure-function relationships in the arginine pathway carbamylphosphate synthase of Saccharomyces cerevisiae. J Bacteriol 134:167-176

Pierson DL, Brien JM (1980) Human carbamylphosphate synthetase I. J Biol Chem 255:7891-7895

Piette J, Nyunoya H, Lusty CJ, Cunin R, Weyens G, Crabeel M, Charlier D, Glansdorff N, Pierard A (1984) DNA sequence of the *carA* gene and the control region of *carAB*: tandem promoters, respectively controlled by arginine and the pyrimidines, regulate the synthesis of carbamoyl-phosphate synthetase in *Escherichia coli* K-12. Proc Natl Acad Sci USA 81:4134-4138

Pons FW, Neubert U, Muller P (1988) Evidence for frameshift mutations in the *hisH* gene of *Escherichia coli* causing synthesis of a partially active glutamine amidostransferase. Genetics 120:657-665

Prusiner S (1973) Glutaminases in *E. coli*: properties, regulation, and evolution. In: Prusiner S, Stadtman RE (eds.) The enzymes of glutamine metabolism, Academic Press, New York, pp. 293-316

Quinn CL, Stephenson BT, Switzer RL (1991) Functional organization and nucleotide sequence of the *Bacillus subtilis* pyrimidine biosynthetic operon. J Biol Chem 266:9113-9127

Raijman L, Jones ME (1976) Purification, composition, and some properties of rat liver carbamyl phosphate synthetase. Arch Biochem Biophys 175:270-278

Sakamoto N, Hatfield GW, Moyed HS (1972) Physical properties and subunit structure of xanthosine 5'-phosphate aminase. J Biol Chem 247:5880-5887

Sato S, Nakada Y, Kanaya S, Tanaka T (1988) Molecular cloning and nucleotide sequence of *Thermus thermophilus* HB8 *trpE* and *trpG*. Biochim Biophys Acta 950:303-312

Schechtman MG, Yanofsky C (1983) Structure of the trifunctional *trp-1* gene from *Neurospora crassa* and its aberrant expression in *Escherichia coli*. J Mol Appl Genet 2:83-99

Schendel FJ, Mueller E, Stubbe J (1989) Formylglycinamide ribonucleotide synthetase from *Escherichia coli*: cloning, sequencing, overproduction, isolation, and characterization. Biochem. 28:2459-2471

Scofield MA, Lewis WS, Schuster SM (1990) Nucleotide sequence of *Escherichia coli asnB* and deduced amino acid sequence of asparagine synthetase B. J Biol Chem 265:12895-12902

Simmer JP, Kelly RE, Rinker AG, Scully JL, Evans DR (1990) Mammalian carbamyl phosphate synthetase. J Biol Chem 265:10395-10402

Slock J, Stahly DP, Han CY, Six EW, Crawford IP (1990) An apparent *Bacillus subtilis* folic acid biosynthesis operon containing *pab*, an amphibolic *trpG* gene, a third gene required for the synthesis of *para*-aminobenzoic acid, and dihydropteroate synthase gene. J Bacteriol 172:7211-7226

Souciet JL, Nagy M, Le Gouar M, Lacroute F, PotierS (1989) Organization of the yeast *URA2* gene: identification of a defective dihydroorotase-like domain in the multifunctional carbamoylphosphate synthetase-aspartate transcarbamylase complex. Gene 79:59-70

Surin BP, Downie JA (1988) Characterization of the *Rhizobium leguminosarum* genes *nodLMN* involved in efficient host-specific nodulation. Mol Microbiol 2:173-183

Tiedeman AA, Smith JM, Zalkin H (1985) Nucleotide sequence of the *guaA* gene encoding GMP synthetase of *Escherichia coli* K12. J Biol Chem 260:8676-8679

Trotta PP, Burt ME, Haschemeyer RJ, Meister A (1971) Proc Nat Acad Sci USA 68:2599-2603

Trotta PP, Pinkus LM, Haschemeyer RH, Meister A (1974) Reversible dissociation of the monomer of glutamine-dependent carbamyl phosphate synthetase into catalytically active heavy and light subunits. J Biol Chem 249:492-499

Tsai FY, Coruzzi GM (1990) Dark-induced and organ-specific expression of two asparagine synthetase genes in *Pisum sativum*. EMBO 9:323-332

Tso JY, Hermodson MA, Zalkin H (1980) Primary structure of *Serratia marcescens* anthranilate synthase component II. J Biol Chem 255:1451-1457

Tso JY, Hermodson MA, Zalkin H (1982a) Glutamine phosphoribosyl-pyrophosphate amidostransferase from cloned *Esclierichia coli purF*. J Biol Chem 257: 3532-3536

Tso JY, Zalkin H (1981) Chemical modifications of *Serratia marcescens* anthranilate synthase component I. J Biol Chem 256:9901-9908

Tso JY, Zalkin H, van Cleemput M, Yanofsky C, Smith JM (1982b) Nucleotide sequence of *Escherichia coli purF* and deduced amino acid sequence of glutamine phosphoribosyl-pyrophosphate amidotransferase. J Biol Chem 257:3525-3531

Tutino ML, Scarano G, Marino G, Sannia G, Cubellis MV (1993) Tryptophan biosynthesis genes *trpEGC* in the thermoacidophilic archaebacterium *Sulfolobus solfotaricus*. J Bacteriol 175:299-302

Van Lookeren Campagne MM, Franke J, Kessin RH (1991) Functional cloning of a *Dictyostelium discoideum* cDNA encoding GMP synthetase. J Biol Chem 266:16448-16452

Vollmer SJ, Switzer RL, Hermodson MA, Bower SG, Zalkin H (1983) The glutamine-utilizing site of *Bacillus subtilis* glutamine phosphoribosylpyrophosphate amidotransferase. J Biol Chem 258:10582-10585

Waley, SG (1969) Some aspects of the evolution of metabolic pathways. Comp Biochem Physiol 30:1-11

Walker JE, Gay NJ, Saraste M, Eberle AN (1984) DNA sequence around the *Escherichia coli unc* operon. Biochem J 224:799-815

Weng M, Makaroff CA, Zalkin H (1986) Nucleotide sequence of *Escherichia coli pyrG* encoding CTP synthetase. J Biol Chem 261 5568-5574

Weng M, Zalkin H (1987) Structural role for a conserved region in the CTP synthetase glutamine amide transfer domain. J Bacteriol 169:3023-3028

Werner M, Feller A, Pierard A (1985) Nucleotide sequence of yeast gene *CPA1* encoding the small subunit of arginine-pathway carbaomyl-phosphate synthetase. Eur J Biochem 146:371-381

Wong S, Abdelal A (1990) Unorthodox expression of a enzyme: evidence for an untranslated region within *carA* from *Pseudomonas aeruginosa*. J Bacteriol 172:630-642

Yanofsky C, Platt T, Crawford IP, Nichols BP, Christie GE, Horowitz H, van Cleemput M, Wu AM (1981) The complete nucleotide sequence of the tryptophan operon of *Escherichia coli*. Nuc Acids Res 9:6647-6668

Ycas M (1974) On earlier states of the biochemical system. J Theoret Biol 44:145-160

Yelton DB, Peng SL (1989) Identification and nucleotide sequence of the *Leptospira biflexa* Serovar *patoc trpE* and *trpG* genes. J Bacteriol 171:2083-2089

Zalkin H (1980) Anthranilate synthase: relationships betweeen bifunctional and monofunctional enzymes. In: Bisswanger H, Schmincke-Ott E (eds) Multifunctional proteins. John Wiley & Sons, New York, pp. 123-149

Zalkin H (1985) Glutamine amidotransferases. Meth Enzymol 113:263-264

Zalkin H, Argos P, Narayana SVL, Tiedman AA, Smith JM (1985) Identification of a trpG-related glutamine amide transfer domain in *Escherichia coli* GMP synthetase. J Biol Chem 260:3350-3354

Zalkin H, Paluh JL, van Cleemput M, Moye WS, Yanofsky C (1984) Nucleotide sequence of *Saccharomyces cerevisiae* genes *TRP2* and *TRP3* encoding bifunctional anthranilate synthase:indole-3-glycerol phosphate synthase. J Biol Chem 259:3985-3992

Zalkin H, Truitt CD (1977) Characterization of the glutamine site of *Escherichia coli* guanosine 5'-monophosphate synthetase. J Biol Chem 252:5431-5436

Zhou G, Dixon JE, Zalkin H (1990) Cloning and expression of avian glutamine phosphoribosylpyrophosphate amidotransferase. J Biol Chem 265:21152-21159

Zimmer W, Aparicio C, Elmerich C (1991) Relationship between tryptophan biosynthesis and indole-3-acetic acid production in *Azospirillum*: identification and sequencing of a *trpGDC* cluster. Mol Gen Genet 229:41-51

Zimmer W, Hundeshagen B (unpublished) A method for cloning fragments of the polymerase chain reaction directly to M13mp18 exemplified by the identification and sequence of the CTP-synthetase gene of *Azospirillum brasilense*. Genbank/EMBL DNA Sequence Database 1992, acc no x67216

Captions for Figures

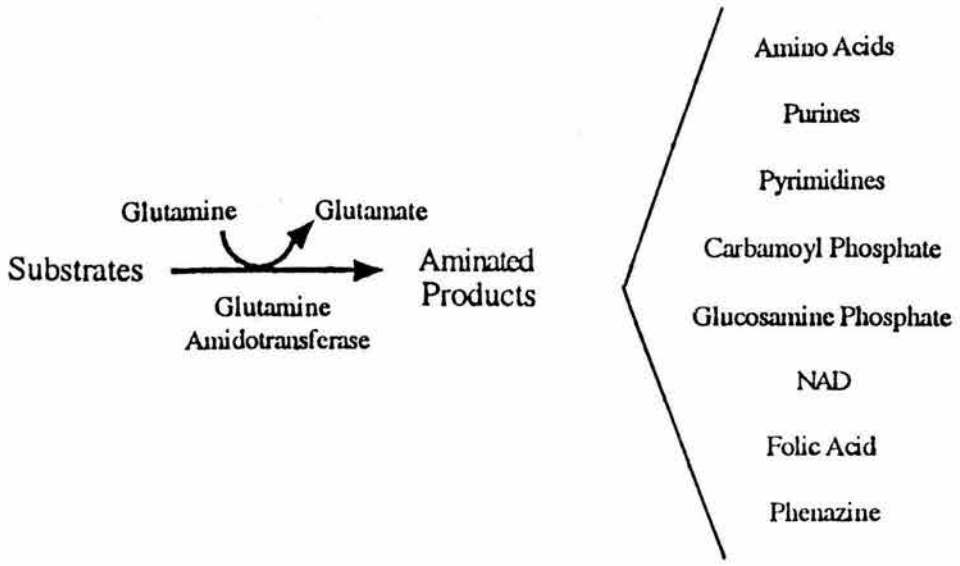
Figure 1. Generalized reaction of glutamine amidotransferases displaying various end-products and biosynthetic intermediates.

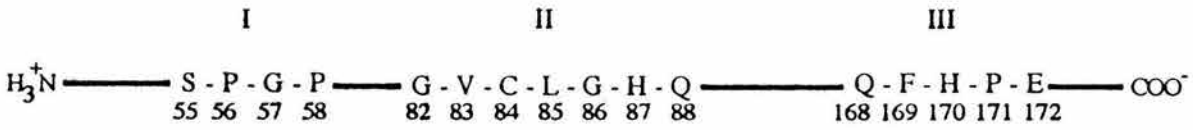
Figure 2. Schematic diagram displaying conserved regions in *trpG*-type GAT domains.

Figure 3. Progressive aligment of *trpG*-type QAT domains. Conserved residues are denoted with asterisk. Abbreviations as in Tables 2-5.

Figure 4. "Paralogous gene" *tree* derived from distance scores from *E. coli trpG*-type GAT domains.

Figure 5. Progressive aligment of *purF*-type GAT domains. Conserved residues are denoted with asterisk. Abbreviations as in Table 9.


Glutamine
Glutamate
Substrates
Aminated Products
Glutamine Amidotransferase
Amino Acids
Purines
Pyrimidines
Carbamoyl Phosphate
Glucosamine Phosphate
NAD
Folic Acid
Phenazine


I
II
III
H_3^+N
S - P - G - P
55 56 57 58
G - V - C - L - G - H - Q
82 83 84 85 86 87 88
Q - F - H - P - E
168 169 170 171 172
COO^-

```
                 *                                    *   *
ECTG         SPGP   GVP  S EAGCMPELL TRLRGK      LP    IIGICLGHQ AIV          EAYGGY
STTG         SPGP   GVP  S EAGCMPELL TRLRGK      LP    IIGICLGHQ AIV          EAYGGY
SDTG         SPGP   GVP  S EAGCMPELL TRLSGK      LP    IIGICLGHQ AIV          EAYGGY
SMTG         SPGP   GTP  S EAGCMPELL QRLRGQ      LP    IIGICLGHQ AIV          EAYGGQ
VPTG         SPGP   GAP  S EAGSMPELI QRMKGK      VP    MIGICLGHQ AIV          EAYGGT
PAPB         SPGP   GRP  E DAGCMLELL AWARGR      LP    VLGVCLGHQ ALA          LAAGGA
ECPA         SPGP   CTP  D EAGISLDVI RHYAGR      LP    ILGVCLGHQ AMA          QAFGGK
STPA         SPGP   CTP  N DAGISLAVI RHYAGR      IP    MLGVCLGHQ AMA          QAFGAS
KAPA         SPGP   CTP  D ESGISLAAI RHFSGQ      TP    ILGVCLGHQ AIA          QVFGAA
PATG         SPGP   CTP  N EAGVSLAVI ERFAGK      LP    LLGVCLGHQ SIG          QAFGGE
PPTG         SPGP   CTP  S EAGVSIEAI LHFAGK      LP    ILGVCLGHQ SIG          QAFGGD
SMPA         SPGP   CTP  N EAGISVAAI RHFAGK      LP    ILGVCLGHQ ALG          QAFGAE
BSPA         SPGP   CSP  D EAGISLEAI KHFAGK      IP    IFGVCLGHQ SIA          QVFGGD
ACTG         GPGP   CSP  T EAGISIPAI HHFAGR      IP    LLGVCLGHQ AIG          QAFGGN
TTTG         SPGP   CTP  F EAGLSVPLV QRYAPR      YP    ILGVCLGHQ AIG          AAFGGK
ABTG         SPGP   CDP  D KAGICLPLIDAAAKAA      VP    LMGVCLGHQ AIG          QPFGGT
PCTG         SPGP   GHP  DTDAGISNAVI KHFSGK      VP    IFGVCMGQQ CMI          TSFGGK
AGTG         SPGP   GHP  ETDAGISSAAI QYFSGK      IP    IFGVCMGQQ CII          TCFGGK
NCTG         SPGP   GHP  GTDSGISRDAI RHFAGK      IP    IFGVCMGQQ CIF          DVYGGD
ADTG         SPGP   GHP  KSDAGISNAAI QYFAGK      IP    IFGVCMGQQ CIH          HSFGGK
SCTG         SPGP   GHP  KTDSGISRDCI RYFTGK      IP    VFGICMGQQ CMF          DVFGGE
PBTG         SPGP   GHP  SHDAGVSRDVI SYFAGK      LP    ILGICMGEQ CIF          EVFGGT
HVTG         SPGP   GHPKNDRDVGVTNDVL TELSTE      IP    TLGVCLGLE AAV          YAYGGT
LBTG         SPGP   GHPADPAYFGVSADIL KELGKT      PP    VLGICLGMQ GMA          TVFGGE
MTTG         SPGP   GNPIKREDFGICSEVI GEFTDR       P    ILGVCLGHQ GIF          HYFGGV
SSTG         SPGP   GTPEKREDIGVSLDVI KYLGKR      TP    ILGVCLGHQ AIG          YAFGAK
LLTG         SPGP   GWP    ADAGKMETLI QQFAGQ      KP    ILGICLGFQ AIV          EVFGGK
SLPA         TPGP   CYP    AEAALNSCSIIGHLAGR      IP    ILGICLGQQ ALG          QARGGL
RMTG         SPGP   GTP    KDFDCKATIK KARARD      LP    IFGVCLGLQ ALA          EAYGGD
BLTG         SPGP   GYP    ADAGNMMALI ERTLGQ      IP    LLGICLGYQ ALI          EYHGGK
BSGA         SGGP   NSVYDENSFRCDEKIF  EL  D      IP    VLGICYGMQ LMT          HYLGGK
ECGA         SGGP   ESTTEENSPRAPQYVF  EA  G      VP    VFGVCYGMQ TMA          MQLGGH
DDGA         SGGP   ESVYGENAPKFDKSLFSEKL  N      LP    IFGICYGMQ LMN          YIFGGK
ECCA         SNGP   GDP    APCDYAITAIQKFLETD      IP    VFGICLGHQ LLA          LASGAK
STCA         SNGP   GDP    APCDYAITAIQKFLETD      IP    LFGICLGHQ LLA          LASGAK
PACA         ANGP   GDP    EPCDYAIRAIQEVLETD      IP    VFGICLGHQ LVA          LASGAI
BSCA         SNGP   GDP    KDVPEAIEMIKGVL GK      VP    LFGICLGHQ LFA          LACGAN
RNCA         AGGP   GNP    ALAQPLIQNVKKILESD  R  KEP    LFGISTGNI ITG          LAAGAK
HSCA         AGGP   GNP    ALAEPLIQNVQKILESD  R  KEP    LFGISTGNL ITG          LAAGAK
MACA         SNGP   GDP    ASYPGVVATLNRVLSEP  N  PRP    VFGICLGHQ LLA          LAIGAK
DMCA         SNGP   GNP    ESCDQIVQQVRKVIEEG     QKP    VFGICLGHQ LLA          KAIGCS
DDCA         SNGP   GDP    SLCGKAIENIRKVLALP  V  AKA    VFGVCMGNQ LLG          LAAGAQ
NCCA         SNGP   GDP    THCQETVYNLAKLMETS     PIP    IMGICLGHQ LLA          LAVGAK
YACA         SNGP   GNP    ELCQATISNVRELLNNPVYD  CIP    IFGICLGHQ LLA          LASGAS
YCCA         SNGP   GDP    SVLDDLSQRLSNVLEAK     KTP    VFGICLGHQ LIA          RAAVQS
ECHH         LPGV   GTAQAAMDQVRERELF DLIK   AC  TQP    VLGICLGMQ LLG          RRSEESN
STHH         LPGV   GTAQAAMDQVRERELI DLIK   AC  TQP    VLGICLGMQ LLG          RRSEETR
SCHH         VPGV   GAFAACMEGLKAARGD WIVDRRLSG  GRP    VMGICVGMQILFS          RGIEHDV
LLHH         LPGV   GAFPTAMNNLKKFNLI ELIQERAAA  GIP    ILGICLGMQVLFE          KGYEIE
ABPG         VPGG   FGS    RGTEGKIRAAQFARERK      VP    YFGICFGMQMAVIESARNMAGIVDAGSTE
ECPG         VPGG   FGY    RGVEGMITTARFARENN      IP    YLGICLGMQVALIDYARHVANMENANSTE
BSPQ         IPGGFSYGDYLRCGAI ARFANIMPAVKQAAAE  GKP    VLGVCNGFQILQE        L GLLPGA
ECPQ         ACGGFSYGDVLGAGEGWAKSILFNDRVRDEFATFFHRPQTLALGVCNGCQMMSN        LRELIPGSE
```

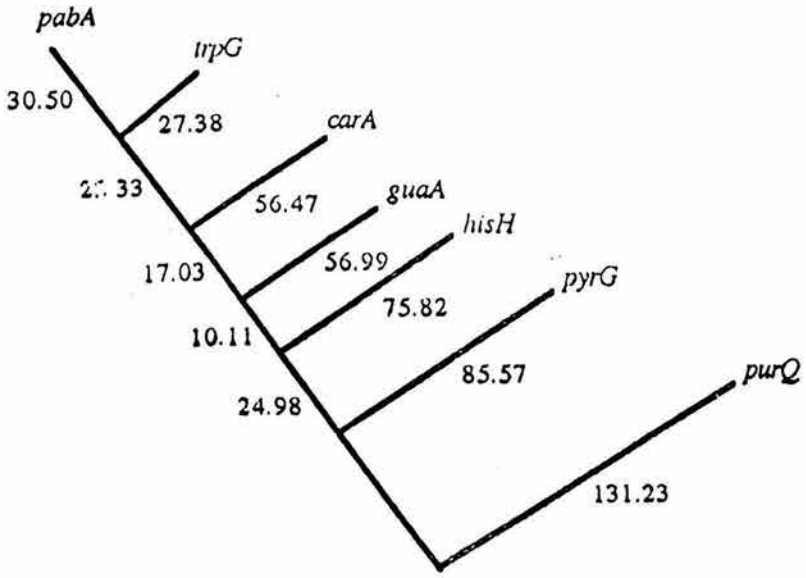
continued

```
ECTG        V   GQAGEILHGKASSIE  HDG QA MFAGLTNP            LPVA      RYH  SLV  GS
STTG        V   GQAGEILHGKASSIE  HDG QA MFAGLANP            LPVA      RYH  SLV  GS
SDTG        V   GQAGEILHGKASSIE  HDG QA MFAGLTNP            LPVA      RYH  SLV
SMTG        V   GQAGEILHGKASAIA  HDG EG MFAGMANP            LPVA      RYH  SLV  GS
VPTG        V   AGAGEIIHGKVSMME  HQD HA IYQNLPSP            LAIA      RYH  SLV
PAPB        V   GEARKPLHGKSTSLR  FDQRHP LFDGIAD             LRVA      RYH  SLV  VS
ECPA        V   VRAAKVMHGKTSPIT  HNG EG VFRGLANP            LTVT      RYH  SLVVEPD
STPA        V   VRAAKVMHGKTSPVT  HNG QG VFRGLPSP            LTVT      RYH  SLIVDPA
KAPA        I   VRAAKVMHGKTSPVS  HTG QG VFLGLNNP            LTVT      RYH  SLL  ID
PATG        V   VRARQVMHGKTSPIH  HKD LG VFAGLANP            LTVT      RYH  SLV  VK
PPTG        V   VRARQVMHGKTSPVH  HRD LG VFTGLNNP            LTVT      RYH  SLVVKRE
SMPA        V   VRARAVMHGKTSAIR  HLG VG VFRGLSDP            LTVT      RYH  SLV  LK
BSPA        V   VRAERLMHGKTSDIE  HDG KT IFEGLKNP            LVAT      RYH  SLIVKPE
ACTG        I   IRAKTVMHGRLSDMY  HTD KG IFSNLPSP            FSAT      RYH  SLVIEQE
TTTG        V   VPAPVLMHGKVSPIH  HDG TG VFRGLDSP            FPAT      RYH  SLA
ABTG        V   VRAPVPMHGKVDRMF  HQG RG VLKDLPSP            FRAT      RYH  SLI  VE
PCTG        V   DVTGEILHGKTSELK  HDS KG VYQGLPTS            LEVT      RYH  SLA  GT
AGTG        V   DVTGEILHGKTSALK  HDG KG AYEGLPDS            LALT      RYH  SLA  GT
NCTG        V   CFAGEILHGKTSPLR  HDG KG AYAGLSQD            LPVT      RYH  SLAGTHV
ADTG        V   DVTGEILHGKTSVLK  HDG RG AYEGLPPS            VIIT      RYH  SLAGTHS
SCTG        V   AYAGEIVHGKTSPIS  HDN CG IFRNVPQG            IAVT      RYH  SLA  GT
PBTG        V   SYAGDILHGKTSTIK  HDN RG LFKNVPQD            NQVT      RYH  SLA  GM
HVTG        I   GHAPDAIHGKAFPVD  HDG AG VFAGLEDG            FPAG      RY    LV  AT
LBTG        V   VRANIAMHGKLSPIE  HDG KG VFSGLTQG            IEIM        H  SLVAKEI
MTTG        V   GY GEPVHGKISEVF  HDG SE LFRGVPNP            FRAT      RYH  RCE  CS
SSTG        I   RRARKVFHGKISNIILVNNSPLS LYYGIAKE            FKAT      RYH  SLV  VD
LLTG        L   RLAHQVMHGKNSQVR  QTSGNL IFNHLPSK            FLVM      RYH  SIVMDEA
SLPA        V   IFA     HGKLSNIE  HNGIFA PLFNPPRA           LPAG      RYH  SLV
RMTG        L   RQLAIPMHGKPSRIR  VLEPGI VFSGLGKE            VTVG      RYH  SIFADPS
BLTG        V   EPCG PVHGTTDNMILTDAGVQSPVFAGLATDVEPDHPEVPGRKVPIG      RYH  SLG  CV
BSGA        V   EAATQREYGKANIRI  EGTPD  LFRDLPNE            QV        VWM  SHGDLVV
ECGA        V   EASNEREFGYAQVEV  VNDSA  LVRGIEDA            LTADGKPLLDVWM  SHGDKVT
DDGA        V   ESNSQREDGVHNIEILKDENQQ  LVSKLFKN            LNQTE     QVLL  THGDSVT
ECCA        T   VKMKFGHHGGNHPV   KDVEKNVVMITAQN                        H  GFAVDEA
STCA        T   VKMKFGHHGGNHPV   KDMDRNVVMITAQN                        H  GFAVDED
PACA        T   VKMPSGHHGANHPV   QDLETGVVMITSQN                        H  GFCADEA
BSCA        T   EKMKFGHRGSNHPV   KELATGKVALTSQN                        H  GYTVSSI
RNCA        S   YKMSMANRGQNQPV   LNITNRQAFITAQN                        H  GYALDN
HSCA        T   YKMSMANRGQNQPV   LNITNKQAFITAQN                        H  CYALDN
MACA        T   YKMRYGNRGHNQPC   LLVGTGRCFLTSQN                        H  GFAVDAD
DMCA        T   YKMKIGNRGHNLPC   LHRATGRCLMTSQN                        H  GYAVDLE
DDCA        T   HKMAFGNRGLNQPC   VDQISGRCHITSQN                        H  GFVIDSN
NCCA        T   IKLSMVIRAHNIPA   LDLTTGQCHITSQN                        H  GYAVDIS
YACA        T   HKLKYGNRAHNIPA   MDLTTGQCHITSQN                        H  GYAVDPE
YCCA        T   LKLKFGNRGHNIPC   TSTISGRCYITSQN                        H  GFAVDVD
ECHH        G   VDLLGIIDEDVPKMTDFGLPLPHMGWNRVYPQ            AGNR      LFQ  GIEDGAY
STHH        G   VDLLNIIEQDVPKMTDFGLPLPHMGWNRVYPQ            RGNR      LFQ  GIEDGAY
SCHH        E   AEGLDEWPGTVGPLE  ADVVPHMGWNTVEAP            ADSQ      LFA  GLDADAR
LLHH        E   RQGLGLLKGEVIPIKT NEKIPHMGWNQLNLA            KTSP      TTH  YLSGNDE
ABPG        L   GKPGNPVVGLLGLMTEWMRGN SLEKRTEGTDVGGTMRLGTYPAKLVPGSKVAEVYGTTDITERHR
ECPG        F     VPDCKYPVVALITEWRDENGNVEVRSEKSDLGGTMRLGAQQCQLVDDSLVRQLYNAPTIVERHR
BSPQ            MRRNKDLKFICRPVELIVQNDETLFTASYEKG          ESITIPVAHGEGNFYCDDET LAT
ECPQ        LWPRFVRNTSDRFEARFSLVEVTQSPSLLLQG MVG          SQMPIAVSHGEGRVEVRDAAHLAA
```

continued

```
                                                                  ***
ECTG          NIPA  GL          TINAHFN G       MVMAVRHDADR VCGFQFHPE
STTG          NVPA  GL          TINAHFN G       MVMAVRHDADR VCGFQFHPE
SDTG          NIPA  GL          TINAHFN G       MVMAVRHDADR ICGFQFHPE
SMTG          NIPA  DL          TVNARSG E       MVMAVRDDRRR VCGFQFHPE
VPTG          KVPD  SL          TITAEVD N       LVMSVVHEQDK VCGFQFHPE
PAPB          RLPE  GF          DCLADAD G       EIMAMADPRNR QLGLQFHPE
ECPA          SLPA  CF          DVTAWSE T     REIMGIRHRQWD LEGVQFHPE
STPA          TLPE  CF          EITAWSE T     QEIMGIRHREWD LEGVQFHPE
KAPA          PLPE  CF          EVTARSE E     GEIMGIRHRVFD LEGVQFHPE
PATG          RLPE  CL          EVTAWTQ HADGSLDEIMGVRHKTLN VEGVQFHPE
PPTG          TLPD  CL          EVTAWTA HEDGSVDEIMGLRHKTLN IEGVQFHPE
SMPA          ALPD  CF          EVTAWSE R DGVRDEIMGIRHRALA LEGVQFHPE
BSPA          TLPS  CF          TVTAQTK E     GEIMAIRHNDLP IEGVQFHPE
ACTG          SLPE  CL          EVTCWTN QNDGSIEEIMGVKHKTLP VEGVQFHPE
TTTG          VVEE  AL          VVNAWAE EAGG  RTVMGFRHRDYP THGVQFHPE
ABTG          TLPA  CL          EVTGETE D     GLIMALSHRELP IHGVQFHPE
PCTG          TIPD  CL          EVTSRVELGDASGKNIIMGVRHKEFA VEGVQFHPE
AGTG          HAPD  CL          EVSSSVQLTDDSNKDVIMGVRHKKLA VEGVQFHPE
NCTG          TLPE  CL          EVTSWIA KEDGSKGVIMGVRHKEYT IEGVQFHPE
ADTG          TIPE  CL          EVSSFAQLGEDADKTVIMGVRHKQFA VEGVQFHPE
SCTG          SLPS  CL          KVTASTENG       IIMGVRHKKYT VEGVQFHPE
PBTG          TLPE  VL          EVTATTDDG       VIMGVRHKKYT VEGVQFHPE
HVTG          DVPD  CF          DVSATTDHDGE   ALVMGVRHRDYP IECVQFHPE
LBTG          SLPN  DL          EITARVSAGEG  KGEIMGLRHKSLK IEGVQFHPE
MTTG          GVPE  DI          LVSASAPDG       TIMAIRHRQYP VYGLQFHPE
SSTG          EVHR  PL          IVDAISAED     NEIMAIHHEEYP IYGVQFHPE
LLTG          VALP  DF          AITAVATDD     GEIMAIENEKEQ IYGLQFHPE
SLPA          VEPA  RI          EVTGQCN QLEVVPQEIMAIRHRDLP VEGVQFHPE
RMTG          NLPR  EF          VITAESEDG       TIMGIEHSKEP VAAVQFHPE
BLTG          VAPD  GI          ESLGTCSSEIG   DVIMAARTTDGK AIGLQFHPE
BSGA          EVPE  GF          TVDAATSHH     CPNSAMSKGDKK WHGVQFHPE
ECGA          AIPS  DF           ITVASTES     CPFAIMANEEKR FYGVQFHPE
DDGA          KIAD  GF          KI ICKSDD     GIVSGIENERLG YYGVQFHPE
ECCA          TLPA  NL          RVTHKSLFD     GTLQGIHRTDKP AFSFQGHPE
STCA          SLPA  NL          RVTHKSLFD     GTLQGIHRTDKP AFSFQGHPE
PACA          AVPA  NL          RATHKSLFD     GTLQGIERTDKV AFSFQGHPE
BSCA          S KT  EL          EVTHIAIND     DTIEGLKHKTLP AFTVQYHPE
RNCA          TLPA  GW          KPLFVNVND     QTNEGIMHESKP FFAVQFHPE
HSCA          TLPA  GW          KPLFVNVND     QTNEGIMHESKP FFAVQFHPE
MACA          SLPA  GW          TPLFTNAND     CSNEGIVHDSLP FFSVQFHPE
DMCA          QLPD  GW          SELFVNAND     GTNEGIVHASKP YFSVQFHPE
DDCA          SLPAGSGW          KTYFINAND     ASNEGIYHESKP WFSVQFHPE
NCCA          TLPS  DF          KEYFVNLND     GSNEGMMHKTRP IFSTQFHPE
YACA          TLPK DQW          KPYFVNLND     KSNEGMIHLQRP IFSTQFHPE
YCCA          TLTS  GW          KPLFVNAND     DSNERFYHSELP YFSVQFHPE
ECHH        FYFVHSYAM           P VNPWTIAQCNYGE PFTAAVQKDN FYGVQFHPE
STHH        FYFVHSYAM           P VNPWTIAQCNYGE PFTAAVQKDN FFGVQFHPE
SCHH      FYFVHSYAVHEWTQESHNPLIAEPRVTWSTHGK PFVAAVENGA LWATQFHPE
LLHH      VYFVHSYQA             TCPDDELIAYTTYGEVKIPAIVGKNN VIGCQFHPE
ABPG      HRYEVNVYY       KDRLEKVGLLFSGLS PTQLPEIVEIPDHPWFIGVQFHPE
ECPG      HRYEVNNSL       LKQIEDAGLRVAGRSGDDQLVEIIEVPNHPWFVACQFHPE
BSPQ      LKENNQIAF             TYGSNIN  GSVSDIAGVVNEKGN VLGMMPHPE
ECPQ      LESKGLVALRYVDNFGKVTETYPANPN  GSPNGITAVTTESGR VTIMMPHPE
```

```
                *                          **
CLAB        cGIWALFGS  DD   CLSVQCLS AMKIAHRGPDAFRFENVNGYTNCCFGFHR LAVVDPLFGMQPIRV
MAAB        cGIWALFGS  DD   CLSVQCLS AMKIAHRGPDAFRFENVNGYTNCCFGFHR LAVVDPLFGMQPIRV
HSAB        cGIWALFGS  DD   CLSVQCLS AMKIAHRGPDAFRFENVNGYTNCCFGFHR LAVVDPLFGMQPIRV
AOAB        cGILAVLGC  SDDSQAKRVRVLELSRRLKHRGPD   WSGLCQHGDCFLSHQR LAIIDPASGDQPLYN
PAB2        cGILAVLGC  SDPSRAKRVRVLELSRRLKHRGPE   WSGLHQHGDCYLAQQR LAIVDPASGDQPLFN
PAB1        cGILAVLGC  SDDSQAKRVRILELSRRLKHRGPD   WSGLHQHGDNYLAHQR LAIVDPASGDQPLFN
ECAB        cSIFGVFDI  KTDAVELRKKALELSRLMRHRGPD   WSGIYASDNAILAHER LSIVDVNAGAQPLYN
RMGS        cGIVGIV      GHQPVSERLVEALEPLEYRGYDSAGVATMD AGT  LQRRRAEGKLGNLREKLKEAP
RLGS        cGIVGIV      GHKPVSERLIEALGRLEYRGYDSSGVATIF EGE  LHRRRAEGKLGNLKTRLKEAP
ECGS        cGIVGAI      AQRDVAEILLEGLRRLEYRGYDSAGLAVVDAEGH  MTRLRRLGKVQMLAQAAEEHP
ECPF        cGIVGIA      GVMPVNQSIYDALTVLQHRGQDAAGIITIDANNC  FRSLKANALVSDVFEARHMQR
SCPF        cGILGIVLA   NQTTPVAPELCDGCIFLQHRGQDAAGIATCGSRGR  VCQCKGNGMARDVFTQ RVSG
BSPF        cGVFGIW      GHEEAPQITYYGLHSLQHRGQEGAGIVATD GEK  LTAHKGQGLITEVFQNGELSK
GGPF        cGVFGCIAAGVWPTELDVPHVITLGLVGLQHRGQESAGIVTSDGESSQAFKVHKGMGLINHVFNADSLKK

                KKYPYLWLCYNGEIYNHKALQQRF  EFEY  QTNVDGEIILH
                KKYPYLWLCYNGEIYNHKALQQRF  EFEY  QTNVDGEIILH
                KKYPYLWLCYNGEIYNHKKMQQHF  EFEY  QTKVDGEIILH
                EDKSIV VTVNGEIYNHEELRRRLPDH KY  RTGSDCEVIAH
                EDNPSI VTVNGEIYNHEDLRKQLSNH TF  RTGSDCDVIAH
                EDKSII VTVNGEIYNHEELRKQLPNH KF  FTQCDCDVLAH
                QQKTHV LAVNGEIYNHQALRAEYGDRYQF  QTGSDCEVILA
                L SGT  IGLAHTRWATHGAPTERNAHPHF  TE   GVAVVH
                L SGT  VGLAHTRWATHGAPTECNAHPHF  TD   GVAVVH
                L HGG  TGLAHTRWATHGEPSEVNAHPHV  SE   HIVVVH
                L QGN  MGIGHVRYPTAGSSSASEAQPFY  VNSPYGITLAH
                L AGS  MGIAHLRYPTPGLRLILEAQPFY  VNSPYGINLAH
                V KGK  GAIGHVRYATAGGGGYENVQPLLFRSQNNGSLALAH
                LYVSN  LGIGHTRYSTSGISELQNCQPFV VETLHGKIAVAH
```


pabA
trpG
30.50
27.38
carA
56.47
guaA
56.99
hisH
17.03
75.82
pyrG
10.11
85.57
purQ
24.98
131.23

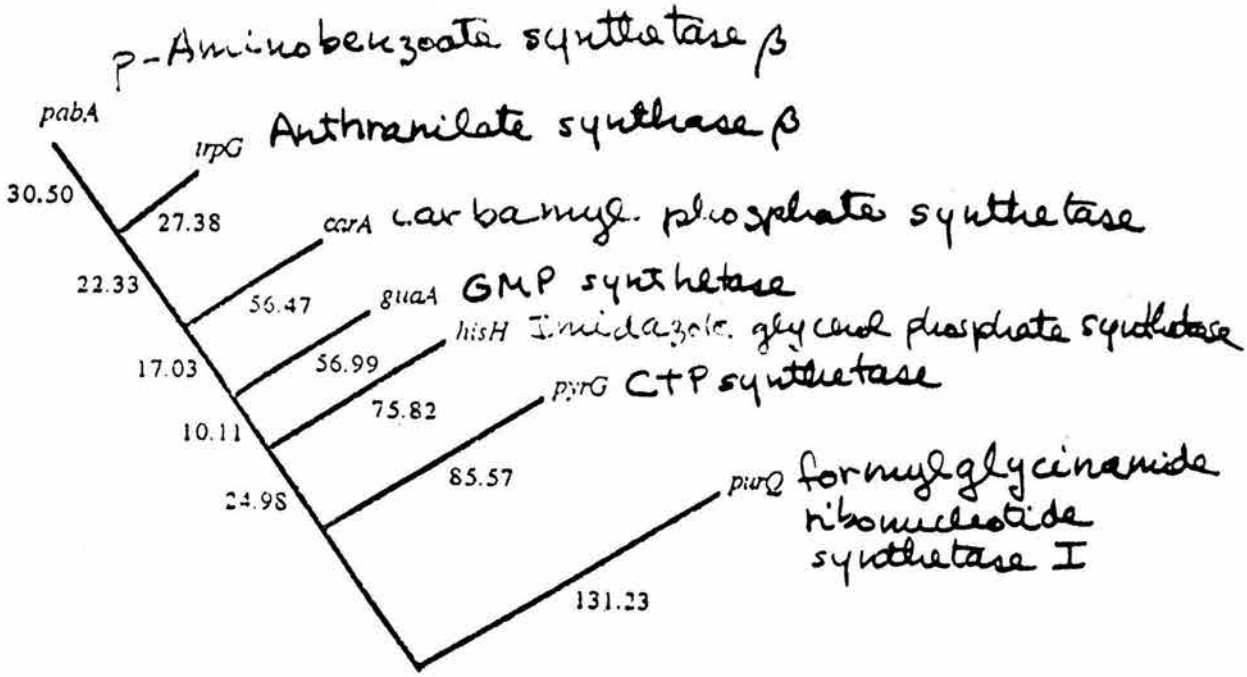
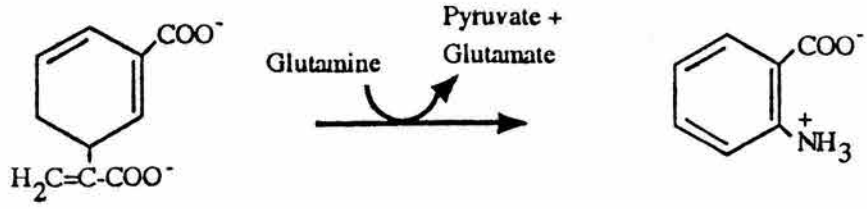

p-Aminobenzoate synthetase β
pabA
trpG Anthranilate synthase β
30.50
27.38
carA carbamyl. phosphate synthetase
22.33
56.47
guaA GMP synthetase
hisH Imidazole. glycerol phosphate synthetase
17.03
56.99
pyrG CTP synthetase
10.11
75.82
24.98
85.57
purQ formylglycinamide ribonucleotide synthetase I
131.23

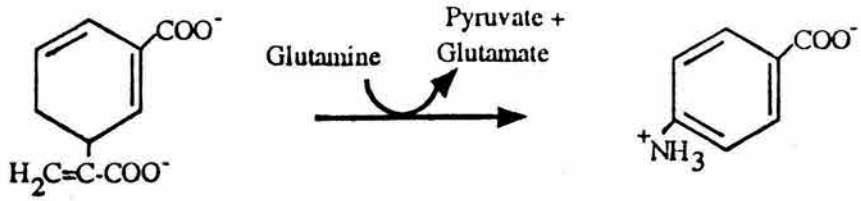
Table 1. Reactions catalyzed by enzymes containing *trpG*-type GAT domains. (table continued on next two pages).

ANTHRANILATE SYNTHASE



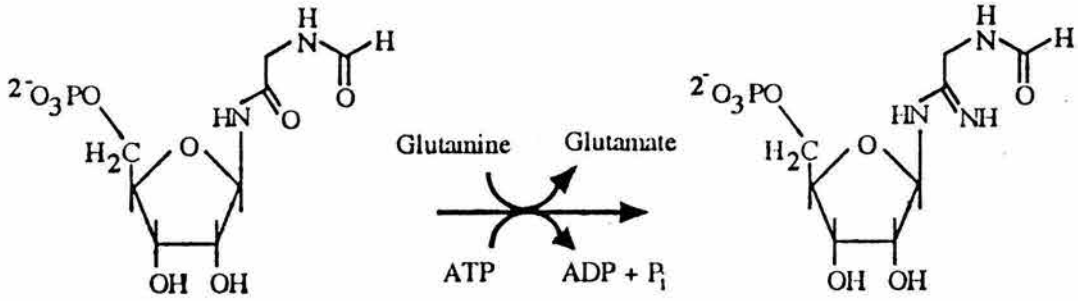
Chorismate — Anthranilate

p-AMINOBENZOIC ACID SYNTHASE



Chorismate — *p*-Aminobenzoate

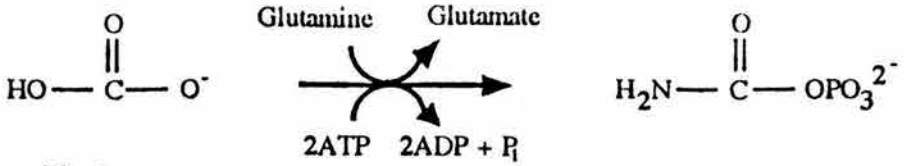
FORMYLGLYCINAMIDINE RIBONUCLEOTIDE SYNTHETASE



Formylglycinamide Ribonucleotide — Formylglycinamidine Ribonucleotide

continued

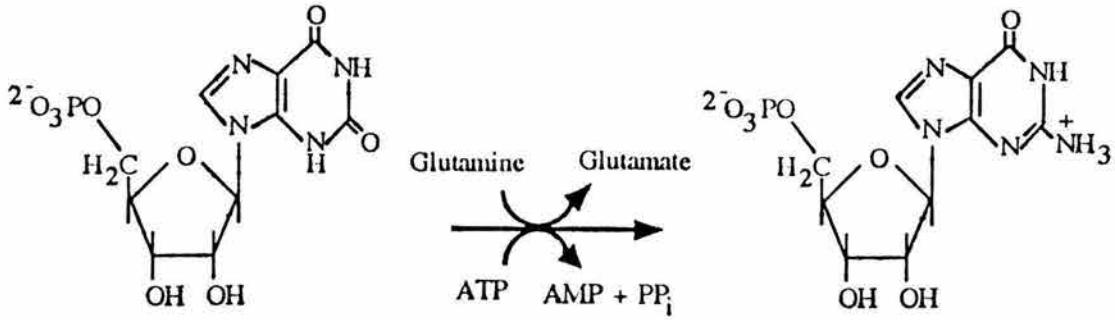
CARBAMOYL PHOSPHATE SYNTHETASE



Bicarbonate

Carbamoyl Phosphate

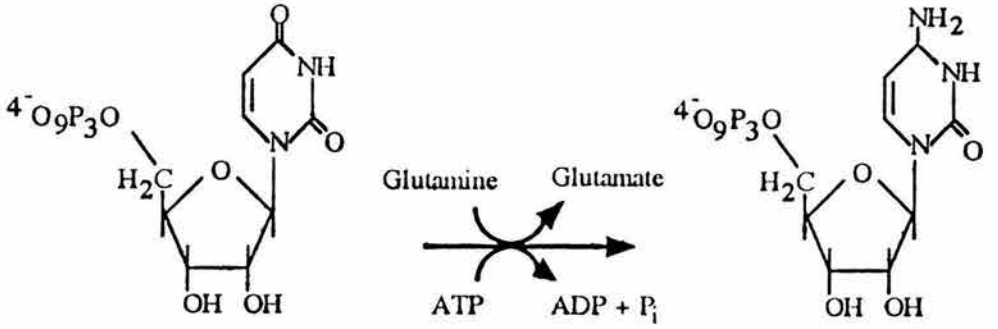
GMP SYNTHETASE



Xanthosine Monophosphate

Guanosine Monophosphate

CTP SYNTHETASE

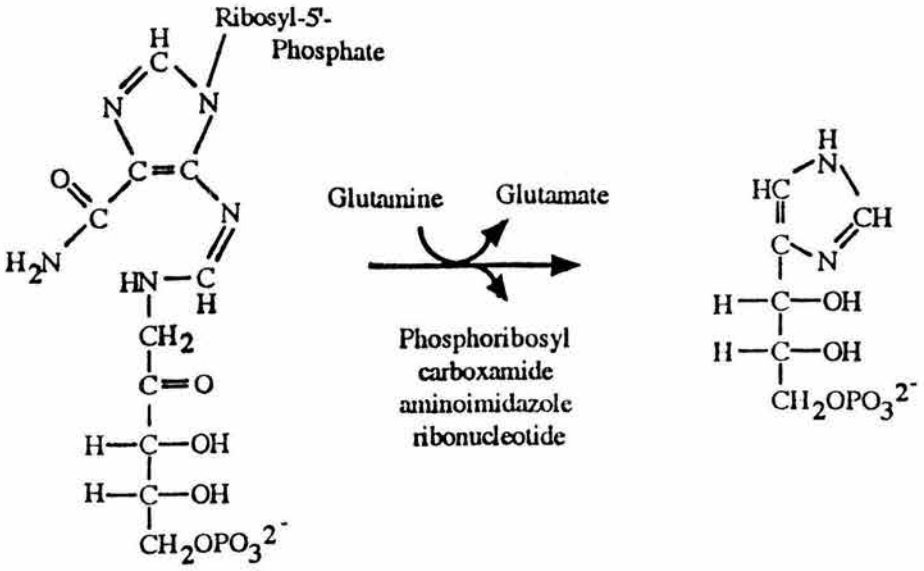


Uridine Triphosphate

Cytidine Triphosphate

continued

IMIDAZOLE GLYCEROL PHOSPHATE SYNTHASE



Aminoimidazole
Carboxamide
Ribonucleotide

Imidazole Glycerol
Phosphate

Table 2. GAT domain DNA sequences from anthranilate synthase and p-aminobenzoate synthase.

Organism	Abrev. Code	Length (a.a.)	Gene Organization	Subunit Configuation†	Reference	Accession Number
ANTHRANILATE SYNTHASE						
Archaebacteria						
Haloferax volcanii trpG	hvtg	204	*trpD·F·E·G*	?	Lam et al. 1992	m83788
M. thermoautotrophicum trpG	mttg	196	*trpE·G·C·F·B·A·D*	?	Meile et al. 1991	m65060
Sulfolobus solfataricus trpG	sstgt	195	*trpE·G·C*(1)	?	Tutino et al. 1993	m98048
Eubacteria						
Acinetobacter calcoaceticus trpG(2)	actg	194	*trpG·D·C*	αβ	Kaplan et al. 1984	m36636
Azospirillum brasilense trpG	abtg	196	*trpG·D·C*	?	Zimmer et al. 1991	x57853
Brevibacterium lactofermentum trpG	bltg	208	*trpL·E·G·D·C(F)·B·A*	?	Matsui et al. 1986	x04960
Escherichia coli trp(G)D	ectg	531(3)	*trpL·E·(G)D·C(F)·B·A*	$\alpha_2\beta_2$	Yanofsky et al. 1981	j01714
Lactococcus lactis trpG	lltg	198	*trpE·G·D·C·F·B·A*	?	Bardowski et al. 1992	m87483
Leptospira biflexa trpG	lbtg	201	*trpE·G*	?	Yelton & Peng 1989	m22468
Pseudomonas aeruginosa trpG	patg	201	*trpG·D·C*	αβ	Essar et al. 1990a	m33814
Pseudomonas putida trpG	pptg	197	*trpG·D·C*	αβ	Essar et al. 1990b	m33799
Rhizobium meliloti trpE(G)	rmtg	729(4)	*trpE(G)*	?	Bae et al. 1989	m22983
Salmonella typhimurium trp(G)D	sttg	531(3)	*trpE·(G)D·C(F)·B·A*	$\alpha_2\beta_2$	Horrowitz et al. 1983	m30285
Serratia marcescens *trpG*	smtg	193	*trpE·G*	$\alpha_2\beta_2$	Nichols et al. 1980	j01792
Shigella dysenteriae trp(G)D	sdtg	>194(3)	*trpE·(G)D*	?	Nichols et al. 1980	j01787
Thermus thermophilus trpG	tttg	204	*trpE·G*	?	Sato et al. 1988	x07744
Vibrio parahaemolyticus trpG	vptg	196	*trpE·G·D·C(F)·B·A*	?	Crawford et al. 1991	x17149
Eukaryotes(3)						
Aspergillus nidulans trpC	adtg	768	*trp(G)C(F)*	?	Mullaney et al. 1985	x02390
Aspergillus niger trpC	agtg	770	*trp(G)C(F)*	?	Adams & Royer 1990	x53576
Neurospora crassa trp1	nctg	762	*trp(G)C(F)*	$\alpha_2\beta_2$	Schechtman et al. 1983	j01252
Penicillium chrysogenum trpC	pctg	752	*trp(G)C(F)*	?	Penalva & Sanchez 1987	x05033
Phycomyces blakesleeanus trp1	pbtg	765	*trp(G)C(F)*	?	Choi et al. 1988	m23177
Saccharomyces cerevisiae TRP3	sctg	484	*trp(G)C*	?	Zalkin et al. 1984	k01386

continued

PABA SYNTHASE

Eubacteria						
Bacillus subtilis pabA(2)	bspa	194	*pabB·pabA·pabC*(5)	?	Slock et al. 1990	m34053
Escherichia coli pabA	ecpa	187	*pabA*	?	Kaplan & Nichols 1983	k00030
Klebsiella aurogenes pabA	kapa	187	*pabA*	?	Kaplan et al. 1985	x02604
Salmonella typhimurium pabA	stpa	187	*pabA*	?	Kaplan et al. 1985	x02603
Serratia marcescens pabA	smpa	191	*pabA*	?	Kaplan et al. 1985	x02605
Streptomyces lividins pabA	slpa	192	*pabB·A*	?	Arhin & Vining, unpub.	m64859

PHENAZINE PATHWAY ANTHRANILATE SYNTHASE

Eubacteria						
Pseudomonas aeruginosa phnB	papb	199	*phnA·B*	?	Essar et al. 1990c	m33811

---

† from Zalkin 1980

(1) *trpEGC* cluster in *S. sulfolobus* is proposed and awaits the elicidation of the positions of other *trp* genes

(2) amphibolic in that it serves as the β subunit for both anthranilate synthase and PABA synthase

(3) *trpG* domain is composed roughly of the 200 N-terminal amino acids

(4) *trpG* domain is composed roughly of the 200 C-terminal amino acids

(5) the given loci names are *pab*, *trpG* (or GAT) and *pabC* respectively , here it is suggested that the more reasonable names *pabB*, *pabA*, and *pabC* are used for the operon in line with the loci names used in other species

Table 3. GAT domain DNA and cDNA sequences from *trpG* homologs

Organism	Abrev. Code	Length (a.a.)	Gene Organization	Subunit Configuation	Reference	Accession Number
IMIDAZOLE GLYCEROL PHOSPHATE SYNTHASE						
Eubacteria						
Azospirillum brasilense hisH	bhh	161	*hisA·B·F·H*	?	Fani et al. 1989	x61207
Escherichia coli hisH	echh	196	*hisG·D·C·B·H·A·F·I(E)*	?	Carlomagno et al. 1988	x13462
Lactococcus lactis hisH	llhh	201	*hisC·G·D·B·H·A·F·I(E)*	?	Delorme et al 1992	m90760
Streptomyces coelicolor hisH	schh	206	*hisD·C·B·H·A*	?	Limauro et al. 1990	m31628
Salmonella typhimurium hisH	sthh	196	*hisG·D·C·B·H·A·F·I(E)*	?	Carlomagno et al. 1988	x13464
FORMYLGLYCINAMIDE RIBONUCLEOTIDE SYNTHETASE						
Eubacteria						
Bacillus subtilis purQ	bspq	227	*purE·K·B·C·Q·L..F·M·N·H(J)·D*	?	Ebbole & Zalkin 1987	j02732
Escherichia coli pur(L)Q	ecpq	1295	*pur(L)Q*	α	Schendel et al. 1989	m19501
GMP SYNTHETASE						
Eubacteria[1]						
Bacillus subtilis guaA	bsga	513	*guaA*	?	Mantsala & Zalkin 1992	m83691
Escherichia coli guaA	ecga	525	*guaB·A*	α_2[2]	Tiedeman et al. 1985	m10101
Eukaryotes						
Dictyostelium discoideum cDNA	ddga	718	*?*	?	V.L. Campagne et al. 1991	m64282

continued

CTP SYNTHETASE

Eubacteria						
Azospirillum brasilense pyrG	abpg	463	*pyrG*	?	Zimmer et al., unpub.	x67216
Escherichia coli pyrG	ecpg	544	*pyrG•eno*	α_4(3)	Weng et al. 1986	m12843

(1) *guaA* exists as part of *guaAB* operon in *E. coli*, but *guaA* and *guaB* are separate genes in *B. subtilis*
(2) Sakamoto et al. 1972
(3) Levitzki et al. 1971

Table 4. GAT domain DNA sequences from eubacterial carbamyl phosphate synthetases

Organism	Abrev. Code	Length (a.a.)	Gene Organization	Subunit Configuration	Reference	Accession Number
CARBAMYL PHOSPHATE SYNTHASE AP(1)						
Escherichia coli carA	ecca	380	*carA·B*	$\alpha\beta$(2)	Piette et al. 1984	j01597
Pseudomonas aeruginosa carA	paca	379	*carA·B*	$\alpha\beta$(3)	Wong & Abdelal 1990	m33818
Salmonella typhimurium carA	stca	381	*carA·B*	$\alpha\beta$(4)	Kilstrup et al. 1988	m36540
CARBAMYL PHOSPHATE SYNTHASE P(5)						
Bacillus subtilis pyrAA	bsca	364	*pyrB·C·AA·AB·D·F·E*	?	Quinn et al. 1991	m59757

(1) glutamine dependent activity (β) and NH_3 dependent activity (α) are encoded by *carA* and *carB* respectively of *carAB* operon
(2) Trotta et al. 1974
(3) Abdelal et al. 1983
(4) Abdelal et al. 1975
(5) glutamine dependent activity and NH_3 dependent activity are encoded by *pyrAA* and *pyrAB* respectively of *pyr* operon

Table 5. GAT domain DNA and cDNA sequences from eukaryotic carbamyl phosphate synthetases

Organism	Abrev. Code	Length (a.a.)	Gene Organization	Subunit Configuation	Reference	Accession Number
CARBAMYL PHOSPHATE SYNTHASE A[1]						
Saccharomyces cerevisiae CPA1	yaca	411	*cpa1*	$\alpha\beta$[2]	Nyunoya & Lusty 1984	x01764
Neurospora crassa arg2	ncca	453	arg2	$\alpha\beta$[3]	Orbach et al. 1990	j05512
CARBAMYL PHOSPHATE SYNTHASE I (NH_3-dependent)						
Homo sapien cDNA	mca	1461	?	α_x[4]	Haraguchi et al. 1991	s73956
Rattus norvegicus (rat) cDNA	hsca	1461	?	α_2[5]	Nyunoya et al. 1985a	m12318
CARBAMYL PHOSPHATE SYNTHASE II[6]						
Dictyostelium discoideum PYR1-3	ddca	?	*PYR1-3*	?	Faure et al. 1989	x14633
Drosophila melanogaster cDNA	dmca	2236	?	?	Freund & Jarry 1987	x04813
Mesocricetus auratus (hamster) cDNA	maca	?	?	?	Simmer et al. 1990	j05503
Saccharomyces cerevisiae URA2	yeca	2212	*URA2*	?	Souciet et al. 1989	m27174

(1) glutamine dependent activity is encoded by *CPA1* in yeast and *arg2* in *N. crassa*, while NH_3 dependent activity is encoded by *CPA2* and *arg3* in yeast and *N. crassa* respectively
(2) Pierard & Schroter 1978
(3) Davis et al. 1980
(4) Pierson & Brien 1980
(5) Raijman & Jones 1976
(6) multifucntional proteins: *D. discoideum*, *D. melonogaster* & Syrian hamster have GATase-CPSase-DHOase-ATCase structure; Yeast has GATase-CPSase-ATCase structure, with separate DHOase

Table 6. **Average inter- and intra-group distances between each sub-group of *trpG*-type GAT domains. *PhnB* had a only a single sequence, so no intra-group average is shown.**

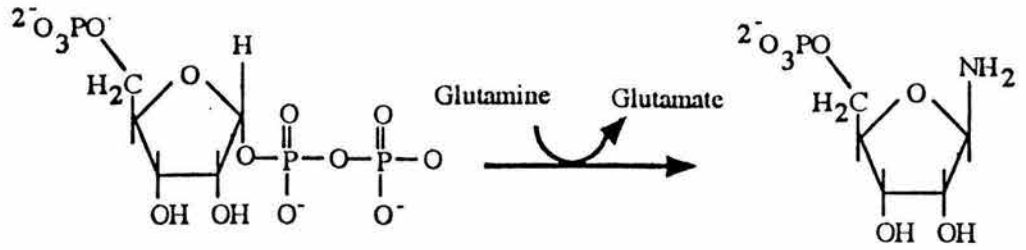
	trpG	*phnB*	*pabA*	*guaA*	*carA*	*hisH*	*pyrG*	*purQ*
trpG	66.6057							
phnB	79.2652	NA						
pabA	62.5582	79.4183	39.3913					
guaA	136.481	131.883	141.146	70.5933				
carA	143.681	137.106	136.14	157.291	57.7129			
hisH	174.593	170.98	179.916	205.023	196.055	72.09		
pyrG	194.66	194.84	194.857	203.383	215.204	195.121	52.35	
purQ	201.863	189.985	208.927	198.862	220.208	228.86	250.148	102.01

Table 7. **Distance scores from alignment of *trpG*-type GAT domains from *E. coli*.**

	pabA	*trpG*	*carA*	*guaA*	*hisH*	*pyrG*	*purQ*
pabA	0.00						
trpG	57.88	0.00					
carA	105.76	109.72	0.00				
guaA	116.58	126.64	137.85	0.00			
hisH	159.37	150.04	156.31	145.11	0.00		
pyrG	187.62	178.34	189.24	181.61	194.30	0.00	
purQ	248.33	237.28	239.53	216.15	223.09	215.80	0.00

Table 8. . Reactions catalyzed by enzymes containing *purF*-type GAT domains.

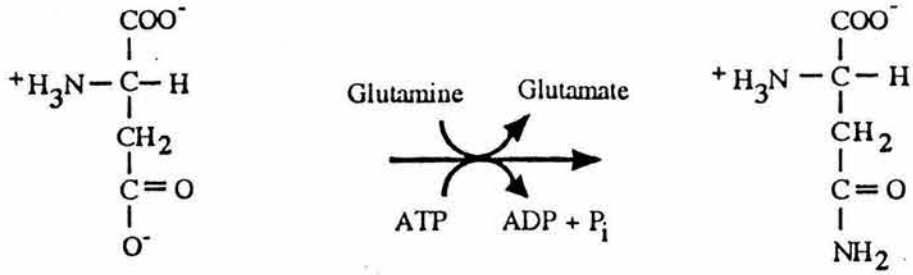
AMIDOPHOSPHORIBOSYLTRANSFERASE



Phosphoribosylpyrophosphate

Phosphoribosylamine

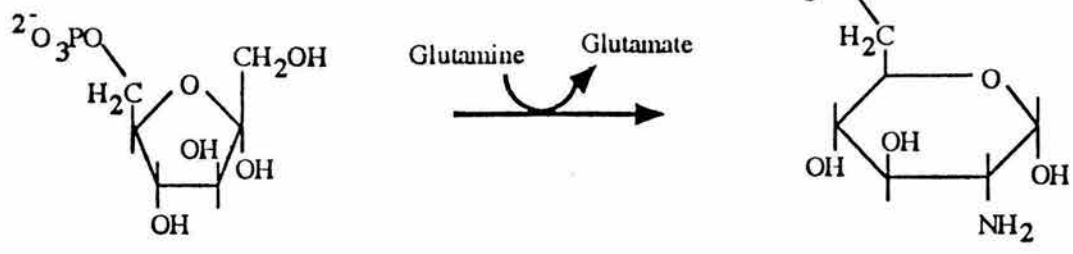
ASPARAGINE SYNTHETASE



Aspartate

Asparagine

GLUCOSAMINE SYNTHASE



Fructose 6-phosphate

Gucosamine 6-phosphate

Table 9. DNA and cDNA sequences from *purF*-type GAT domains.*

	Organism	Abrev. Code	Polypeptide Length	Reference	Accession Number
GLUTAMINE PRPP AMIDOTRANSFERASE					
Eubacteria	*Bacillus subtilis purF*	bspf	465	Makaroff et al. 1983	j02732
	Escherichia coli purF	ecpf	503	Tso et al. 1982b	v00322
Eukaryotes	*Gallus gallus* (chicken) cDNA	ggpf	499	Zhou et al. 1990	m60069
	Saccharomyces cerevisiae ADE4	scpf	509	Mantsala & Zalkin 1984b	k02203
ASPARAGINE SYNTHETASE					
Eubacteria	*Escherichia coli asnB*	ecab	553	Scofield et al. 1990	j05554
Eukaryotes	*Asparagus* officinalis (Asparagus plant) cDNA	aoab	509	Davies et al., unpub.	x67958
	Cricetulus longicaudatus (Chinese hamster) cDNA	clab	560	Andrulis et al. 1989	m27898
	Homo sapien cDNA	hsab	560	Andrulis et al. 1987	m27396
	Mesocricetus auratu (Syrian hamster) cDNA	maab	488	Gong & Basilico 1990	x52130
	Pisum sativum (pea plant) cDNA	p1ab	585	Tsai & Coruzzi 1990	x52179
	Pisum sativum (pea plant) cDNA	p2ab	582	Tsai & Coruzzi 1990	x52180
GLUCOSAMINE SYNTHASE					
Eubacteria	*Escherichia coli glmS*	ecgs	608	Walker et al. (1984)	x01631
	Rhizobium meliloti nodM	rmgs	604	Baev et al. 1991	x58632
	Rhizobium leguminosarum nodM	rlgs	607	Surin & Downie 1988	m13658

* in all cases, the GAT domain occupies positions 1 through approximately 200 starting at the amino-terminus

J Mol Evol (1994) 39:546–554

Journal of Molecular Evolution

How Long Did It Take for Life to Begin and Evolve to Cyanobacteria?

Antonio Lazcano,[1] Stanley L. Miller[2]

[1] Facultad de Ciencias, UNAM, Apdo. Postal 70-407, Cd. Universitaria, México 04510, D.F., Mexico
[2] Department of Chemistry, University of California, San Diego, La Jolla, CA 92093-0317, USA

Received: 19 March 1994 / Revised and accepted: 5 July 1994

Abstract. There is convincing paleontological evidence showing that stromatolite-building phototactic prokaryotes were already in existence 3.5×10^9 years ago. Late accretion impacts may have killed off life on our planet as late as 3.8×10^9 years ago. This leaves only 300 million years to go from the prebiotic soup to the RNA world and to cyanobacteria. However, 300 million years should be more than sufficient time. All known prebiotic reactions take place in geologically rapid time scales, and very slow prebiotic reactions are not feasible because the intermediate compounds would have been destroyed due to the passage of the entire ocean through deep-sea vents every 10^7 years or in even less time. Therefore, it is likely that self-replicating systems capable of undergoing Darwinian evolution emerged in a period shorter than the destruction rates of its components (<5 million years). The time for evolution from the first DNA/protein organisms to cyanobacteria is usually thought to be very long. However, the similarities of many enzymatic reactions, together with the analysis of the available sequence data, suggest that a significant number of the components involved in basic biological processes are the result of ancient gene duplication events. Assuming that the rate of gene duplication of ancient prokaryotes was comparable to today's present values, the development of a filamentous cyanobacterial-like genome would require approximately 7×10^6 years—or perhaps much less. Thus, in spite of the many uncertainties involved in the estimates of time for life to arise and evolve to cyanobacteria, we see no compelling reason to assume that this process, from the beginning of the primitive soup to cyanobacteria, took more than 10 million years.

Key words: Prebiotic synthesis — Early gene duplication — Time for life to arise

Introduction

There has been considerable speculation over the years as to how long it took for life to arise from the prebiotic soup. Most of the estimates have emphasized the long periods of time (several billions of years) thought to be available for the process (Oparin 1938; Urey 1952; Wald 1954; Huang 1959a; Simpson 1964; Cloud 1968; Rutten 1971; Dickerson 1978; Schidlowski 1992), although there have been some dissenting opinions (Miller 1982; Oberbeck and Fogleman 1989; Awramik 1992; Lazcano et al. 1992). In recent years the time available has been considerably compressed. There are compelling microfossils in the Warrawoona Formation at 3.5×10^9 years (Schopf 1983, 1993), and it is clear that life may have been killed off as late as 3.8×10^9 years ago if impacts from large asteroids were important (Sagan 1974; Maher and Stevenson 1988; Sleep et al. 1989; Chyba et al. 1990; Oberbeck and Fogleman 1990). This leaves only 300×10^6 years to go from the prebiotic soup to the cyanobacterialike microfossils in the Warrawoona Formation. A number of workers feel cornered by the very short time allowed for the origin and early evolution of

Correspondence to: A. Lazcano

life. Attempts to deal with this disarray have led some to propose solutions within the framework of panspermia (Brooks and Shaw 1978; Hoyle 1983; Schidlowski 1990). We show here that the probable rates of the chemical steps leading to the first self-replicating entities are rapid. If the rates of gene amplification and fixation of duplicate copies of genes were comparable to present-day values, then there is more than adequate time to go from the prebiotic soup to the RNA world to the DNA/protein world, and then to cyanobacteria. The whole process may have taken place in less than 10 million years, and possibly in much less time. We will take the beginning of the process of originating life as the point after the last large destructive event, when the prebiotic soup began to accumulate.

The Rates of Syntheses and Destruction of Prebiotic Organic Compounds

The chemistry of prebiotic reactions is fast and robust. The rates of the prebiotic synthesis of amino acids, almost all of which are the result of the Strecker synthesis, have been worked out in some detail. The rates depend on temperature, pH, and HCN, NH_3, and aldehyde concentrations. At pH 8 and 0°C, the slow step in amino acid synthesis, the hydrolysis of the corresponding amino nitrile to the amide, is ~40 years (Miller and Van Trump 1981). Adenine synthesis is also rapid from concentrated solutions of NH_4CN—hours at 100°C (Oró 1960) and a few weeks at 0°C. The reaction is so robust that samples of liquid HCN containing a trace of basic catalyst have been known to explosively polymerize (Walker and Eldred 1925). The reaction occurs in months in dilute aqueous solution frozen at −20°C (Sanchez et al. 1966). Even though sugars are currently out of favor as prebiotic reagents (Shapiro 1988; Joyce et al. 1987), we note that their synthesis from formaldehyde is also rapid. This is not to imply that all the monomers were formed in a few years. The formation and buildup of the prebiotic soup took place over millions of years, but the individual reactions have short half-lives, and there are no known slow steps. An example of a rapid prebiotic synthesis is that of amino acids on the Murchison meteorite asteroid, where it apparently took place in less than 10^5 years (Peltzer et al. 1984).

The accumulation of organic compounds is limited by destructive processes. The destruction rate that applies to all organic compounds on the Earth is controlled by the pyrolysis of organics in the submarine vents. Large amounts of water go through the vents, with the whole ocean on the average passing through them in 10 million years (Edmonds et al. 1982), during which most the organic compounds are destroyed at temperatures of 350°C or higher (Miller and Bada 1988). The approximate passage time through the vents is inversely proportional to the heat flow and to the ocean size. The passage rate through the vents may have been considerably faster on the early Earth because the heat flow was about ten times greater and the oceans may have been smaller. We will use the conservative figure of 5 million years. Not all organic compounds on the Earth will be destroyed by the vents. Even assuming a well-mixed ocean, there is a fraction of organics that will escape the vents for long periods of time. In addition, organic material in places such as some inland sediments, salt deposits, evaporites, and in the nonmixed parts of the oceans would not go through the vents. Even with these protected areas, most of the organic molecules, especially the water-soluble ones, would have been decomposed by the vents in geologically short times.

If it is assumed that the organic compounds were produced over a short period of time in a short-lived reducing atmosphere or came to the Earth in comets or interstellar dust, then vent destruction sets a limit on the availability of the organic compounds. This availability will follow first-order kinetics with a half-life for passage through the vents of 5 million years. If the production of organic compounds is continuous over extended periods of time, then a steady-state oceanic concentration will result from the continuous input balanced by loss in the vents (Stribling and Miller 1987), and there will be no buildup of slowly synthesized molecules. The steady-state concentrations reach half their maximum values in about 5 million years. If prebiotic synthesis or extraterrestrial input were to stop for some reason (e.g., a change in atmospheric composition), then the constraint on the time for life to arise would be the half-life for decomposition of the essential compounds. This is likely to be considerably shorter than the 5-million-year constraint due to passage through the vents.

We are not aware of any relevant examples of slowly synthesized molecules. Some prebiotic reactions that could produce critically important compounds could have high activation energies and so proceed very slowly, but we know of no such example. Another possibility is that a polymer such as a peptide is synthesized at random and that a sequence of low probability is needed to initiate a self-replicating system. We consider these examples to be unlikely, but such possibilities cannot be excluded.

There are numerous other loss channels besides the vents, the most important of which is decomposition in aqueous solutions, which is significant even at 0°C (Miller and Orgel 1974). Although some compounds such as acetic acid and alanine are very stable, sugars and many amino acids decompose in hundreds or a few thousand years, as do peptides and the phosphodiester bonds in RNA and DNA (Lindhal 1993). Thus the proposals that the essential prebiotic compounds and polymers formed as a result of unspecified very slow reactions do not take into account the numerous loss channels.

The Origin of Self-Replicating Systems

It is difficult to estimate the rate of organization of polymers into self-replicating systems because the chemical steps are unknown. There are at least two stages in this process. The first is the transition from the prebiotic soup to the pre-RNA world, in which polymers are capable of self-replication but lack the ribose phosphate RNA backbone. The second stage is the replacement of the precursor backbone by ribose phosphate. The bases in the pre-RNA world may not have been adenine, uracil, guanine, and cytosine, and alternative bases may have been used in the early RNA world. The second stage seems relatively easy compared to the development of self-replicating molecules, and the difficult steps may only be the biosynthesis of ribose and possibly the bases.

The major unknown lies in our almost-complete ignorance of the origin of self-replicating pre-RNA molecules. We are clearly unable to give a time scale for this period, but there are some constraints. An informational polymer must have a lifetime comparable to that of the organisms (Westheimer 1987) or, at least, to the time required for its replication. If a slow addition of monomers to polymers is assumed, the rate of polymer formation must nonetheless be rapid compared to hydrolysis rates, especially if a considerable amount of genetic information is to be contained in the polymer. Thus, a 100-base-long RNA molecule needs to be synthesized 100 times faster than the hydrolysis rate of a single phosphodiester bond. Even if highly stable precursors to the ribose phosphate backbone of RNA are proposed, i.e., the bases attached to polyethylene (Pitha 1977), the bases themselves will decompose over long periods of time. For example, cytosine hydrolyzes to uracil in 300 years (pH 7, 25°C) in single-stranded DNA (Lindhal 1993). Adenine, which is usually thought to be very stable, deaminates to hypoxanthine at a rate only 20 times slower than cytosine (Frick et al. 1987; Shapiro 1994).

The second argument is that all the prebiotic chemistry known so far is rapid on the geological time scale (Miller and Orgel 1974), and so it is at least plausible that the unknown reactions are also rapid. The idea that life emerged rapidly becomes even more compelling if one accepts the possibility of a hot emergence of life (Pace 1991), since a high-temperature environment would greatly lower the half-life of organic compounds, especially nucleic acids, prior to the emergence of replicating systems.

How Long Did the RNA World Last?

The above considerations provide some constraints on the time necessary for life to arise even though we do not have a mechanism for which a realistic calculation might be made. The hydrolysis and decomposition rates indicate very short periods of time, but the destruction in the vents places an upper limit. As discussed above, we take 5 million years for this figure, recognizing that the actual time may have been much shorter.

It has been suggested that the RNA world extended up to $2.0–2.5 \times 10^9$ years ago (Benner et al. 1989), but this would require that the microfossils in the Warrawoona and Gunflint formations were RNA organisms, and we consider this unlikely. We believe that the RNA world could not have lasted for a long period, since it would have quickly exhausted the prebiotically synthesized purines, pyrimidines, and ribose, some of which were probably never very abundant. The contemporary biosynthesis of adenine, uracil, guanine, and cytosine from simple precursors requires elaborate metabolic pathways. There are 13 steps in the present biosynthesis of adenosine monophosphate and six steps in the biosynthesis of cytidine monophosphate. This may not be achievable with ribozymes. It is possible that another pathway was used, e.g., a ribozyme-catalyzed synthesis of adenine from HCN. The biosynthesis of ribose would need to begin particularly early in the RNA world, since large amounts of ribose, if any, could not have accumulated in the primitive ocean because of its instability. The same would be true of cytidine, which is hydrolized to uridine with a half-life at 0°C of less than 2×10^4 years (Miller and Orgel 1974).

The second major bottleneck is the origin of protein synthesis, i.e., the transition from the RNA world to the DNA/protein world. The process seems very complex, but several factors point to a rapid emergence of protein synthesis: (1) Protein synthesis must have developed before the prebiotic amino acids were destroyed in the vents or by thermal decomposition, unless they were available due to amino acid biosynthetic routes accomplished by ribozymes; (2) the *E. coli* 52 ribosomal proteins are the result of gene duplications (Jue et al. 1980), and so may have developed very rapidly; and (3) the aminoacyl tRNA synthetases are also the result of gene duplication events of one or possibly two starter types (Nagel and Doolittle 1991).

Given the rapid exhaustion of prebiotically synthesized RNA components (purines, pyrimidines, and ribose) discussed above, as well as the apparent versatility of proteins over ribozymes, the available evidence points to a rapid transition to a nucleic acid/protein world. The emergence of protein synthesis is sometimes thought to be a chicken-and-egg problem because of the large number of proteins involved in the process. However, there is evidence suggesting that ribosomal proteins may not be needed for protein biosynthesis, so a rudimentary system may have arisen entirely using ribozymes (Noller et al. 1992), and therefore not involving a series of improbable events.

Size Increase of Early Genomes

The time for evolution of the first DNA/protein organisms to the cyanobacteria is usually thought to be very long because the latter are generally considered to be very complex. We assume that the first DNA/protein organism had about 100 enzymes with ~1,000 bp/gene, 80 of which may have resulted from gene duplication of 20 "starter types." A primordial DNA genome encoding 100 enzymes is only one-fourth the size of *Mycoplasma capricolum*'s genome, which is one of the smallest-known free-living prokaryotes and contains approximately 400 genes, of which 350 encode for proteins (Muto et al. 1986). The assumed primordial genome size may be reasonable, since primitive enzymes were probably less specific than modern ones and so able to catalyze a wider range of chemically related metabolic reactions (Ycas 1974; Jensen 1976).

The filamentous cyanobacteria *Oscillatoria* spp., which are morphologically similar to the Warrawoona microfossils, have genome sizes of $\sim 5.9 \times 10^3$ kb (Herdman 1985). The question is how to estimate the time needed to go from a primordial organism with 100 genes to an oscillatorianlike cyanobacteria with approximately 7,000 enzymes. If the Warrawoona microfossils are remains of anoxygenic photosynthetic bacteria with only photosystem I (Olson and Pierson 1986; Ward et al. 1989), then even fewer genes and less time would be required for their emergence.

The evolution of 7,000 enzymes over the 5 million years estimated above on average amounts to one enzyme per 700 years. This may be fast for enzymes like mammalian hemoglobins, but it is very easy to understand such a rate for haploid organisms under metabolic stress conditions. The major source of new genes under such conditions is gene duplication, which occurs under a wide variety of different environmental conditions (Hoffman 1985; Schimke et al. 1986; Sonti and Roth 1989). Under starvation conditions, bacteria are known to undergo duplications of large segments of their genome. This provides more enzymes that allow them to overcome the lowered concentrations of nutrients. These duplications are lost on the removal of the selection pressure (Rigby et al. 1974; Sonti and Roth 1989). It has been observed, however, that under the starvation conditions used in some directed evolution experiments, only a few weeks are required to fix duplicate copies. Examples of this include the copies of the gene encoding a ribitol dehydrogenase in *Klebsiella aerogenes* (Rigby et al. 1974; Hartley 1974). Comparable rapidity is observed in the case of a haploid strain of *Saccharomyces cerevisiae*, in which duplicates of a gene for an acid phosphatase involved in the hydrolysis of β-glycerophosphate (Hansche 1975) were also quickly fixed.

We assume that life arose in a prebiotic soup containing most, if not all, of the necessary small molecules, and that the evolution of the basic biosynthetic pathways occurred in a stepwise fashion (Horowitz 1945) or by means of a related mechanism. It is clear that this provides both for the growth and the energy supply of a large number of organisms, and it quickly results in the depletion of the available nutrients. The selective pressure to develop these basic biosynthetic pathways would be so strong that the process could be called *explosive metabolic evolution*, analogous to the explosive evolution of metazoa in the late Precambrian. The process would have been over as soon as the biosynthetic pathways for all the basic components were developed. After this, there was very little to add except for the fine tuning of the biochemical syntheses and the appearance of regulatory mechanisms.

There was a large potential energy supply from different fermentations. The observed production of cells from a variety of fermentation reactions is 10 g of dry weight per mole of ATP (Thauer et al. 1977). This corresponds to 5.6×10^{11} cells per mole of ATP, assuming 2 μ cells and 70% water content. The usual example cited is the fermentation of glucose, but is unlikely that large quantities of this sugar were available because of its instability. A more likely early fermentation reaction is that suggested by Clarke and Elsden (1980):

$$\text{glycine} + \text{NADH} + \text{ADP} + \text{P}_\text{i} = \text{acetate} + \text{NH}_4^+ + \text{NAD}^+ + \text{ATP}$$

Thus an ocean with a volume $V_{\text{ocean}} = 1.5 \times 10^{21}$ l with a 1×10^{-4} M concentration of glycine would have contained 1.5×10^{17} moles of glycine. At 1 mol ATP per mole of glycine, this would be equivalent to 8.4×10^{28} cells. A glycine concentration of 10^{-4} M is a rich prebiotic soup, but producible under reducing conditions (Stribling and Miller 1987). The concentration may have been much lower, perhaps 10^{-8} M if the organic compounds were only of extraterrestrial origin (Anders 1989; Chyba et al. 1990). In the latter case, this would still represent 10^{25} cells. Other fermentations would have increased the mass of the primordial biosphere perhaps by a factor of ten, but there would still have been an exponential decrease in the concentration of the available fermentable organic compounds of prebiotic origin. Such a decrease would bring about a metabolic energy crisis, and it would have developed rapidly. While 10^{29} cells is a huge population, it still corresponds to only 7×10^4 cells/cm^3 of ocean. Although osmotic energy sources and some inorganic sources (e.g., methane production from $CO_2 + H_2$, and reduction of sulfate) could take the place of fermentations, the only really abundant energy source was visible light. Thus, there would be a strong selection pressure for the development of photosynthesis.

Nei (1969) and Li (1982) have estimated that the rate of increase of genomes is 7.0–7.5 nucleotide pairs per

year. If it is assumed that gene duplications are neutral, at any given time in a population there should be numerous duplicates on their way to fixation or loss. Under the assumption that duplications can be treated as random mutations, the number of neutral duplicate genes that will become established per generation in a given population will be equal to their rate of occurrence (Kimura 1968; Doolittle 1979; Li 1982). The rate of spontaneous duplication of bacterial genes shows considerable variation, but it is generally estimated to have values of 10^{-5}–10^{-3} gene duplications per gene per cell generation (Anderson and Roth 1977; Stark and Wahl 1984; Tlsty et al., 1984). We will use the most conservative figure of 10^{-5}, and for the sake of simplicity will assume that during the early Archean there were ten generations of cells per year. Thus, the rate of accumulation of duplicons in a primordial genome encoding 100 enzymes would be 100 genes × 10 cell generations per year × 10^{-5} gene duplications per gene per cell generation, which equals 0.01 gene duplications per year. If 10% of the duplications are neutral and 90% are deleterious (Li 1982), the rate of accretion of duplicons would be 0.001 gene per year. Since we have assumed that primordial genes were 1,000 nucleotides long, the genomes would be increasing their DNA content at a rate of one nucleotide pair per year. This figure can be seen to be very conservative in that it is almost an order of magnitude smaller than the ~7 nucleotide pairs per year given by Nei (1969) and Li (1982). However, one nucleotide pair per year implies that the maximum time required to go from the assumed 100-gene-DNA-protein organism to a 7,000-gene filamentous cyanobacteria, assuming only neutral duplications, would be 7×10^6 years. The time would be a factor of 17 less if it is assumed that the rate of genome growth is proportional to its size.

The situation under strong positive selection should be much faster. If the fixation of new duplicates increases the overall fitness of the population, as during starvation conditions, then the rate of accretion of new genes may be expected to increase significantly (Spofford 1969; Mayo 1970; Li 1982). Theoretical estimates suggest that even for a small selective advantage, the rate of fixation of new duplicates per generation can be increased by a factor of four (Mayo 1970). Such an advantage will develop when one of the duplicate gene copies undergoes favorable mutations or internal recombinations, and it will then spread rapidly throughout the population. The rapid divergence of a fixed duplicate is exactly what would happen as the prebiotic soup became exhausted, compound by compound. The heterotrophs that developed abilities to metabolize the remaining organic compounds would have been strongly selected over others. The selective pressure favouring those bacteria that had developed an enzymatic apparatus allowing the use of photochemical energy sources would have been even stronger.

To the best of our knowledge there are no published experimental data on the rate of formation of new enzyme activities resulting from gene duplication events under either neutral or positive selection conditions. However, it is known that under stress conditions in directed evolution experiments the conversion of existing enzymes to new substrate specificities for different amides (Clarke 1986) and for galactosylarabinose (Hall and Zuzel 1980) may take place even in a few weeks. If the ability of early Archean bacteria to adapt to new carbon, nitrogen, and energy sources was comparable to those of present prokaryotes, then the explosive metabolic evolution of metabolism becomes easy to understand.

There is, however, a third bottleneck in this process. Although the divergence of related enzymes from a starter type should be very rapid, the emergence of the original enzyme (e.g., the first NAD or FAD enzyme) would be much more difficult. The number of starter types is a matter of debate (Zuckerkandl 1975; Doolittle 1981), but it was probably very small in the 100-enzyme organism discussed above, perhaps as few as 20. In addition, the starter types may stem from slow nonenzymatic reactions in which the protein improves on a previously sluggish process. An example would be pyridoxal-catalyzed transaminations.

The above estimates imply that no duplicons are lost. Extant merodiploid bacteria, i.e., partially diploid prokaryotes, are known to be unstable, and populations rapidly revert to the haploid wild type (Anderson and Roth 1977; Roth and Schmid 1981). However, experimental stress conditions can be designed in order to retain the merodiploid phenotype and can be maintained for many generations (Straus and Straus 1976). If such duplicates diverge into sequences with new selectable functions, they would rapidly be fixed. Many duplicate genes may be expected to be lost by streamlining selection (Zamenhof and Eichhorn 1967; Smith 1970) or by chance elimination, but it is known that additional copies of a gene encoding enzymes with identical or overlapping biochemical properties can be fixed even if they do not undergo major divergence (Khan and Haynes 1972; Riley and Anilionis 1978; Dykstra et al. 1984; Wang and Walker 1993).

Gene Duplications and the Origin of Oxygen-Releasing Photosynthesis

The calculations presented here would seem to ignore the complexity of some basic metabolic processes. However complex, the enzymes involved in these processes are mostly the result of multiple gene duplications, as indicated by the fact that more than 30% of the *E. coli* proteins whose sequences are available are the result of duplication events (Riley 1993). Examples include elon-

gation factors, [aa]tRNAs, F-type ATPases, ferredoxins, dehydrogenases, carbamoyl-phosphate synthetases, glutamate synthetases, glutamine synthetases, DNA polymerases, and DNA topoisomerases (Forterre et al. 1993; García-Meza et al. 1994; Gogarten-Boekels and Gogarten 1994). All these duplications took place before the divergence of the three cell lines.

In the case of chlorophyll-dependent photosynthesis, evidence of duplication and double-duplication events has been preserved in the sequences of ferredoxins (Otaka and Ooi 1989), F-type ATPases (Gogarten et al. 1989), the reductases involved in chlorophyll and bacteriochlorophyll biosynthesis (Burke et al. 1993), the bacterial photosynthetic reaction center (Feher et al. 1989; Blankenship 1992), the two sets of light-harvesting antennae (Youvan and Ismail 1985), and photosystems I and II (Youvan and Marrs 1984; Blankenship 1992). A gene duplication of photosystem I and the subsequent development of oxygen-producing photosynthesis still requires the linkage via electron flow of the two photosystems, but this could involve a surprisingly small number of modifications (Chapman and Schopf 1983). It is possible that photosystems I and II developed in different organisms and combined by horizontal transfer or a fusion event (Büttner et al. 1992; Blankenship 1992), but this would only have accelerated the sequence of evolutionary events.

Proof-Reading and Fidelity

There are additional mechanisms that could increase the rate of evolution. These include recombination within genes (Hall and Zuzel 1980), the modular assembly of new proteins (Gilbert 1987), gene sharing (Piatigorsky and Wistow 1989), gene fusion events (Confalonieri et al. 1993), horizontal transfer (Campbell 1981), and duplication of the entire primitive genome (Herdman 1985). In addition, the error rate and lower repair efficiency of early organisms may have been the most important factor in the evolution of new metabolic activities. This is consistent with the average mutation rate per base pair being approximately inversely proportional to the genome size for DNA viruses, bacteria, and ascomycetes (Drake 1991). Based on this combined evidence, it is reasonable to assume that the rate of evolution during the early Archean would have been considerably faster than that of present-day microorganisms, and that evolution to cyanobacteria may have been much faster than the 7×10^6 years period calculated here based on genome size growth under neutral conditions.

The absence of proofreading refinements in early cells was probably the most significant factor limiting the development of larger genomes. As pointed out by Eigen (1971), there is a threshold relationship between the mean error rate in replication and the genome size. It is possible, of course, that repair and proofreading began in the RNA world and were subsequently transferred into DNA polymerases. However, there are no known viral or cellular RNA polymerases with proofreading abilities. Since RNA replication is limited by an error rate of about 10^{-4}, the genome size is restricted to approximately ten enzymes (Holland 1993), and polymerases with editing properties must have developed early. Although the repair and proofreading processes are complex, the major evolutionary obstacle was probably the development of the first exonuclease. This may have been derived from the active site of a polymerase, since the exonuclease reaction is in part the reverse of the polymerization step. The set of different DNA polymerases endowed with editing properties appears to have arisen by gene duplication (Forterre et al. 1993).

Conclusions

We are aware that many of the figures presented here are uncertain. It is not possible to know how fast these historical events actually took place. We only suggest that the chemistry and the bacterial genetics are such that it could have taken place in 10 million years. In spite of the many uncertainties that plague current descriptions of the origin and early evolution of life, the data summarized here suggest that the most important bottlenecks slowing down the process leading from the prebiotic soup to the RNA world and to cyanobacteria may have been (1) the origin of replicating systems; (2) the emergence of protein biosynthesis; and (3) the evolutionary development of the starter types from which latter proteins evolved through gene duplication and divergence. It is possible, of course, to imagine a process in which periods of rapid evolution such as those described here alternated with more slowly evolving intermediate stages during which ribosome-mediated protein biosynthesis was developed. In fact, such episodic schemes are now recognized to have taken place during the evolution of a number of animal and plant lineages. However, we feel that because of the rapid exhaustion of prebiotically synthesized compounds, it is extremely unlikely that unstable, intermediate stages during early cellular evolution could have persisted for many millions of years. We thus see no compelling reason to assume that the origin and early evolution of life took more than 10 million years. We believe that all our estimates of time required correspond to upper limits, and that revision and refinement of the calculations presented here will give time periods considerably less than 10 million years, rather than significantly higher figures.

Finally, these estimates also suggest that if the early environmental conditions on Mars were comparable to those of the Earth, life may have also started there. Moreover, O, B, and A stars, which are usually omitted in calculations of the abundance of life (Huang 1959a,b), because of their short main-sequence lifetimes of

approximately 1.5, 15, and 500 million years, respectively (Limber 1960), may have life on planets surrounding them. Double stars, which are also omitted from SETI calculations because they do not have long-term stable orbits (Huang 1960), may also have microbial life on their planets.

Acknowledgments. We thank C. Gómez Eichelman for several useful references, as well as M. Gogarten-Boekels and J.P. Gogarten for providing us with their results prior to publication. We thank the NSCORT (NASA Specialized Center for Research and Training) in Exobiology at the University of California, San Diego, for a visiting professor fellowship (A.L.) and grant support (S.L.M.).

References

Anders E (1989) Pre-biotic organic matter from comets and asteroids. Nature 342:255–257

Anderson RP, Roth JR (1977) Tandem genetic duplications in phage and bacteria. Ann Rev Microbiol 31:473–505

Anderson RP, Roth JR (1981) Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between ribosomal RNA (*rrn*) cistrons. Proc Natl Acad Sci USA 78:3113–3117

Awramik SM (1992) The oldest records of photosynthesis. Photosynt Res 33:75–89

Benner SA, Ellington AD, Tauer A (1989) Modern metabolism as a palimpsest of the RNA world. Proc Natl Acad Sci USA 86:7054–7058

Blankenship RE (1992) Origin and early evolution of photosynthesis. Photosynt Res 33:91–111

Brooks J, Shaw G (1978) A critical assessment of the origin of life. In: H. Noda (ed) Origin of life. Center for Academic Publications/ Japan Sci. Soc. Press, Tokyo, pp 597–606

Burke DH, Hearst JE, Sidow A (1993) Early evolution of photosynthesis: clues from nitrogenase and chlorophyll iron proteins. Proc Natl Acad Sci USA 90:7134–7138

Büttner M, Xie D-L, Nelson H, Pinther W, Hauska G, Nelson N (1992) Photosynthetic reaction centers genes in green sulfur bacteria and in photosystem I are related. Proc Natl Acad Sci USA 89:8135–8139

Campbell A (1981) Evolutionary significance of accessory DNA elements in bacteria. Ann Rev Microbiol 35:55–83

Chapman DJ, Schopf JW (1983) Biological and biochemical effects of the development of an aerobic environment. In Schopf JW (ed) Earth's earliest biosphere: its origin and evolution. Princeton University Press, Princeton, pp 302–320

Chyba CF, Thomas PJ, Brookshaw L, Sagan C (1990) Cometary delivery of organic molecules to the early Earth. Science 249:366–373

Clarke PH (1986) Experiments on the evolution of bacteria with novel enzyme activities. Chemica Scripta 26B:337–342

Clarke PH, Elsden SR (1980) The earliest catabolic pathways. J Mol Evol 15:333–338

Cloud PE (1968) Atmospheric and hydrospheric evolution of the primitive Earth. Science 160:729–736

Confalonieri F, Elie Ch, Nadal M, Bouthier de La Tour C, Forterre P, Duguet M (1993) Reverse gyrase: a helicase-like domain and a type I topoisomerase in the same polypeptide. Proc Natl Acad Sci USA 90:4753–4758

Dickerson RE (1978) Chemical evolution and the origin of life. Sci Am 239(3):70–86

Doolittle RF (1979) Protein evolution. In: Neurath H, Phill RL (eds) The proteins. Academic Press, New York, vol 4, pp 1–118

Doolittle RF (1981) Similar amino acid sequences: chance or common ancestry? Science 214:149–159

Drake JW (1991) A constant rate of spontaneous mutations in DNA-based microbes. Proc Natl Acad Sci USA 88:7160–7164

Dykstra CC, Prasher D, Kushner SR (1984) Physical and biochemical analysis of the cloned *recB* and *recC* genes of *Escherichia coli* K-12. J Bacteriol 157:21–27

Edmonds JM, Von Damm KL, McDuff RE, Measures CI (1982) Chemistry of hot springs on the East Pacific Rise and their effluent dispersal. Nature 297:187–191

Eigen M (1971) Self-organization of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465–523

Feher G, Allen JP, Okamura MY, Rees DC (1989) Structure and function of bacterial photosynthetic reaction centers. Nature 339:111–116

Forterre P, Benachenhou-Lahfa N, Confalonieri F, Duguet M, Elie Ch, Labedan B (1993) The nature of the last universal ancestor and the root of the tree of life, still open questions. Biosystems 28:15–32

Frick L, Mac Neela JP, Wolfenden R (1987) Transition state stabilization of deaminases: rates of nonenzymatic hydrolysis of adenosine and cytidine. Bioorg Chem 15:100–108

García-Meza V, González-Rodríguez V, Lazcano A (1994) Ancient paralogous duplications and the search for Archean cells. In: Fleischaker GR, Colonna S, Luisi P-L (eds) Self production of supramolecular structures: from synthetic structures to models of minimal living systems Kluwer Academic Publishers, Dordrecht

Gilbert W (1987) The exon theory of genes. Cold Spring Harbor Symp Quant Biol 52:901–905

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, Konishi J, Denda K, Yoshida M (1989) Evolution of the vacuolar H^+-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci USA 86: 6661–6665

Gogarten-Boekels M, Gogarten JP (1994) The effects of heavy meteorite bombardment on the early evolution of cellular life–a new look at the molecular record. Orig Life Evol Biosph (submitted)

Hall B, Zuzel T (1980) Evolution of a new enzymatic function by recombination within a gene. Proc Natl Acad Sci USA 77:3529–3533

Hansche PE (1975) Gene duplication as a mechanism of genetic adaptation in *Saccharomyces cerevisiae*. Genetics 79:661–674

Hartley BS (1974) Enzymes families. In: Carlile MJ, Skehel JJ (eds) Evolution in the microbial world: 24th symposium of the Society for General Microbiology, Cambridge University Press, Cambridge, pp 151–182

Herdman M (1985) The evolution of bacterial genomes. In: Cavalier-Smith T (ed) The evolution of genome size. John Wiley and Sons, London, pp 37–68

Hoffman GR (1985) Genetic duplication in bacteria and their relevance for genetic toxicology. Mutat Res 150:107–117

Holland J (1993) Replication error, quasispecies populations, and extreme evolution rates of RNA viruses. In: Morse SS (ed) Emerging viruses. Oxford University Press, New York, pp 203–218

Horowitz NH (1945) On the evolution of biochemical synthesis. Proc Natl Acad Sci USA 31:153–157

Hoyle F (1983) The intelligent universe. Michael Joseph Ltd, London

Huang S-S (1959a) Occurrence of life in the Universe. Am Sci 47: 397–402

Huang S-S (1959b) The problem of life in the Universe and the mode of star formation. Publ Astron Soc Pacific 71:421–424

Huang S-S (1960) Life supporting regions in the vicinity of binary systems. Publ Astron Soc Pacific 72:106–114

Jensen RA (1976) Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409–425

Joyce GF, Schwartz AW, Miller SL, Orgel LE (1987) The case for an ancestral genetic system involving simple analogues of the nucleotides. Proc Natl Acad Sci USA 84:4398–4402

Jue RA, Woodbury NW, Doolittle RF (1980) Sequence homologies

among *E. coli* ribosomal proteins: evidence for evolutionary related groupings and internal duplications. J Mol Evol 15:129–148

Khan NA, Haynes RH (1972) Genetic redundancy in yeast: non-identical products in a polymeric gene system. Mol Gen Genet 118:279–285

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Lazcano A, Fox GE, Oró J (1992) Life before DNA: the origin and evolution of early Archean cells. In: Mortlock RP (ed) The evolution of metabolic function. CRC Press, Boca Raton, pp 237–295

Li W-H (1982) Evolutionary change of duplicate genes. In: Rattazzi MC, Scandalios JG, Whitt GS (eds) Isozymes: current topics in biological and medical research. Alan R. Liss, New York, vol 6, pp 55–92

Limber DN (1960) The universality of the initial luminosity function. Astrophys J 131:168–201

Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362:709–715

Maher KA, Stevenson DJ (1988) Impact frustation of the origin of life. Nature 331:612–614

Mayo O (1970) The role of duplications in evolution. Heredity 25:543–553

Miller SL (1982) Prebiotic synthesis of organic compounds. In: Holland HD, Schidlowski M (eds) Mineral deposits and the evolution of the biosphere: a Dahlem konferenzen. Springer-Verlag, Berlin, pp 155–176

Miller SL, Bada JL (1988) Submarine hot springs and the origin of life. Nature 334:609–611

Miller SL, Orgel LE (1974) The Origins of life on earth. Prentice Hall, Englewood Cliffs

Miller SL, Van Trump JE (1981) The Strecker synthesis in the primitive ocean. In: Wolman Y (ed) Origin of life. Reidel, Dordrecht, pp 135–141

Muto A, Yamao F, Hori H, Osawa S (1986) Gene organization of *Mycoplasma capricolum*. Adv Biophys 21:49–56

Nagel GM, Doolittle RF (1991) Evolution and relatedness in two aminoacyl-tRNA synthetase families. Proc Natl Acad Sci USA 88: 8121–8125

Nei M (1969) Gene duplication and nucleotide substitution in evolution. Nature 221:40–42

Noller HF, Hoffarth V, Zimniak L (1992) Unusual resistance of peptidyl transferase to protein extraction procedures. Science 256: 1416–1419

Oberbeck VR, Fogleman G (1989) Estimates of the maximum time required to originate life. Orig Life Evol Biosph 19:549–560

Oberbeck VR, Fogleman G (1990) Impact constraints on the environment for chemical evolution and the continuity of life. Orig Life Evol Biosph 20:181–195

Olson JM, Pierson BK (1986) Photosynthesis 3.5 thousand million years ago. Photosynt Res 9:251–259

Oparin AI (1938) The origin of life. MacMillan, New York

Oró J (1960) Synthesis of adenine from ammonium cyanide. Biochem Biophys Res Commun 2:407–412

Otaka E, Ooi T (1989) Examination of protein sequence homologies: V. New perspectives on evolution between bacterial and chloroplast-type ferredoxins inferred from sequence evidence. J Mol Evol 29:246–254

Pace NR (1991) Origin of life—facing up to the physical setting. Cell 65:531–533

Peltzer ET, Bada JL, Schlesinger G, Miller SL (1984) The chemical conditions on the parent body of the Murchison meteorite: some conclusions based on amino-, hydroxy-, and dicarboxylic acids. Adv Space Res 4:69–74

Piatigorsky J, Wistow GJ (1989) Enzyme/crystallins: gene sharing as an evolutionary strategy. Cell 57:197–199

Pitha J (1977) Vinyl polymer analogues of nucleic acids. Polymer 18:425–430

Rigby PW, Burleigh BD, Hartley BS (1974) Gene duplication in experimental enzyme evolution. Nature 251:200–204

Riley M (1993) Functions of the gene products of *Escherichia coli*. Microbiol Rev 57:862–952

Riley M, Anilionis A (1978) Evolution of the bacterial genome. Ann Rev Microbiol 32:519–560

Roth JR, Schmid MB (1981) Arrangement and rearrangement of the bacterial chromosome. Stadler Symp 13:53–70

Rutten MG (1971) The origin of life by natural causes. Elsevier, Amsterdam

Sagan C (1974) The origin of life in a cosmic context. Orig Life 5:497–505

Sanchez RA, Ferris JP, Orgel LE (1966) Cyanoacetylene in prebiotic synthesis. Science 154:784–785

Schidlowski M (1990) Life on the early Earth: bridgehead from cosmos or autoctonous phenomenon? In: Gopalan K, Gaur VG, Somayajulu BLK, Macdougall JD (eds) From mantle to meteorites. Indian Academy of Sciences, Bangalore, pp 189–199

Schidlowski M (1992) The initiation of biological processes on the Earth: summary of empirical evidence. Adv Space Res 12:143–156

Schimke RT, Sherwood SW, Hill AB (1986) The rapid generation of genomic change as a result of over-replication of DNA. Chemica Scripta 26B:305–307

Schopf JW (ed) (1983) The earth's earliest biosphere: its origin and evolution. Princeton University Press, Princeton

Schopf JW (1993) Microfossils of the early Archean apex chert: new evidence of the antiquity of life. Science 260:640–646

Shapiro R (1988) Prebiotic ribose synthesis: a critical analysis. Orig Life Evol Biosphere 18:71–85

Shapiro R (1994) The prebiotic role of adenine: a critical analysis. Abstracts of the 207th American Chemical Society National Meeting, San Diego, California, March 13–17, 1994, Division of Geochemistry, Abstract 29

Simpson GG (1964) The nonprevalence of humanoids. Science 143: 769–775

Sleep NH, Zahnle KJ, Kasting JF, Morowitz HJ (1989) Annihilation of ecosystems by large asteroid impacts on the early Earth. Nature 342:139–142

Smith EL (1970) Evolution of enzymes. In: Boyer PD (ed) The enzymes. Academic Press, New York, 3rd ed, vol I, pp 267–339

Sonti RV, Roth JR (1989) Role of gene duplications in the adaptation of Salmonella typhimurium to growth on limiting carbon sources. Genetics 123:19–28

Spofford JB (1969) Heterosis and the evolution of duplicates. Am Nat 103:407–432

Stark GR, Wahl GM (1984) Gene amplification. Ann Rev Biochem 53:447–491

Straus DS, Straus LD (1976) Large overlapping tandem duplications in *Salmonella tiphymurium*. J Mol Biol 103:143–153

Stribling R, Miller SL (1987) Energy yields for hydrogen cyanide and formaldehyde syntheses: the HCN and amino acid concentrations in the primitive oceans. Orig Life 17:261–273

Thauer RK, Jungermann K, Decker K (1977) Energy conservation in chemotrophic anaerobic bacteria. Bacteriol Rev 41:100–180

Tlsty TD, Albertini AM, Miller JH (1984) Gene amplification in the *lac* region of *E. coli*. Cell 37:217–224

Urey HC (1952) On the early chemical history of the Earth and the origin of life. Proc Natl Acad Sci USA 38:351–363

Wald G (1954) The origin of life. Sci Am 191:44–53

Walker M, Eldred DN (1925) The decomposition of liquid hydrocyanic acid. Ind Eng Chem 17:1074–1081

Wang Y, Walker PJ (1993) Adelaide River rhabdovirus expresses consecutive glycoprotein genes as polycistronic mRNAs: new evidence of gene duplication as an evolutionary process. Virology 195:719–731

Ward DM, Weller R, Shiea J, Castenholz RW, Cohen Y (1989) Hot

spring microbial mats: anoxygenic and oxygenic mats of possible evolutionary significance. In: Cohen Y, Rosenberg E (eds) Microbial mats: physiological ecology of benthic microbial communities. American Soc Microbiol Washington, DC, pp 3–15

Westheimer FH (1987) Why nature chose phosphates. Science 235: 1173–1178

Yeas M (1974) On earlier states of the biochemical system. J Theor Biol 44:145–160

Youvan DC, Ismail S (1985) Light-harvesting II (B800-850-complex) structural genes from *Rhodopseudomonas capsulata*. Proc Natl Acad Sci USA 82:58–62

Youvan DC, Marrs BL (1984) Molecular genetics and the light reactions of photosynthesis. Cell 39:1–3

Zamenhof S, Eichhorn HH (1967) Study of microbial evolution through loss of biosynthetic functions: establishment of "defective" mutants. Nature 216:456–458

Zuckerkandl E (1975) The appearance of new structures and functions in proteins during evolution. J Mol Evol 7:1–57

CONCLUSIONES

1. En los últimos años se ha podido caracterizar tanto los mecanismos que pudieron haber llevado a la síntesis prebiótica de diversos monómeros químicos de importancia biológica, como el papel que parecen haber jugado las ribozimas en etapas tempranas de la evolución. Sin embargo, huelga decir que aun subsisten grandes lagunas en nuestro conocimiento sobre el origen y la evolución temprana de la vida. Es posible, por ejemplo, que el llamado mundo del RNA haya sido precedido por sistemas biológicos mas simples, cuya existencia estuviera basada en polímeros catalíticos y replicativos que carecieran de esqueletos fosfodiestéricos. De ser así, es necesario reconocer allí el punto de origen de las primeras rutas biosintéticas, y de cuya evolución habría de resultar eventualmente el RNA mismo. Nada sabemos, sin embargo, de estas etapas tempranas, ni de la continuidad o no que pudo haber existido entre estas vías anabólicas primordiales y las que son comunes a todos los seres vivos contemporáneos. Por lo tanto, los modelos, la metodología, y resultados discutidos en este trabajo solamente son aplicables a fases de la evolución biológica posteriores al surgimiento de proteínas con propiedades catalíticas.

2. A partir de la idea de que las enzimas son los catalizadores biólogicos por antonomasia, y cuarenta años antes de que se reconociera el papel del RNA en la evolución biológica temprana, N. H. Horowitz (1945) sugirió la llamado hipótesis retrógrada para explicar el origen de las rutas biosintéticas. La aplicabilidad de esta propuesta parece ser bastante limitada. El análisis cladístico de los operones conocidos no apoya la idea de Horowitz que supone que los genes adyacentes son homólogos y codifican para enzimas que catalizan pasos sucesivos en la misma ruta biosintéticas. Es cierto que en algunas vías metabólicas se pueden identificar reacciones sucesivas catalizadas por enzimas homólogas, pero estos ejemplos son pocos y no bastan, por si mismos, para validar la hipótesis retrógrada. Se puede suponer, por ejemplo, que los pasos son químicamente equivalentes y que la enzima original era menos específica. En este caso, la evidencia de la

homología de las enzimas podía ser vista como apoyo a las ideas de Waley (1969), Ycas (1974) y Jensen (1976).

3. La hipótesis de Horowitz, en su formulación original, parece no ser aplicable para explicar la aparición de la biosíntesis de la valina, la isoleucina, y la leucina. El análisis de las secuencias del operon ilvGMEDA demostró que los genes que lo forman no son el resultado de duplicaciones sucesivas (Cox et al., 1987). Por otra parte, los precursores bioquímicos de la valina, la isoleucina, y la leucina son extraordinariamente inestables, y es poco probable que se hubieran acumulado en la sopa primitiva (Keefe et al., 1994). Sin embargo, es posible que el anabolismo de estos aminoácidos hubiera comenzado con la carboxilación de ácidos grasos de cadena corta, seguida de reacciones de transaminación no-enzimáticas. Este esquema, que está basado en una variante de la hipótesis retrógada que supone la existencia en el medio prebiótico de los productos de degradación de los compuestos orgánicos de origen abiótico, se puede extender para explicar el origen de la biosíntesis de serina y la treonina apelando al reclutamiento de enzimas ancestrales poco específicas que originalmente participaban en la biosíntesis de los aminoácidos alifáticos. Este esquema sugiere que aunque la hipótesis Waley-Ycas-Jensen parece tener un valor considerable, algunas premisas de la idea de Horowitz son rescatables y tienen un cierto valor heurístico.

4. La existencia de rutas anabólicas que sintetizan de distinta manera los mismo compuestos muestra que no todo el metabolismo básico es de origen monofilético. La lista de estas vias incluye, entre otros, la biosíntesis de los tetrapirroles, del ácido mevalónico, y a la lisina (Chapman y Ragan, 1980). El origen de estas rutas alternas ha sido poco estudiado, pero muestra la conveniencia de llevar a cabo un análisis concienzudo de la distribución filogenética de las vias metabólicas. Aunque se han reportado distintas vías para la isoleucina, la metionina, la tirosina, y la lisina (Chapman y Ragan, 1980), este no parece ser el caso de la histidina. La biosíntesis de este aminoácido parece haberse establecido antes de la separación de los tres linajes celulares, y está extraordinariamente conservada. El análisis cladístico de los genes de la histidina ha permitido probar el papel que distintos eventos de elongación y duplicación génica

jugaron en el ensamblaje de esta ruta. Estas comparaciones sugieren que en un principio la biosíntesis de la histidina estuvo mediada originalmente por una serie de enzimas ancestrales poco específicas (incluyendo la glutamin-amido transferasa, la aminotransferasa, la ciclasa, y la isomerasa que participan en distintos pasos de la ruta) que probablemente podían intervenir en otras rutas (Fani et al., 1995; Alifano et al., 1996). Este descubrimiento apoya la hipótesis del "patchwork" en contraposición de las ideas de Horowitz, y contradice al menos en parte la idea de que la biosíntesis de la histidina es un vestigio del mundo del RNA.

5. Aunque el análisis filogenético de los genes que participan en la biosíntesis de la histidina apoyan la hipótesis de Waley-Ycas-Jensen, la cercanía filogenética que muestran las secuencias de los genes hisC y hisI de bacterias Gram positivas con los de arqueobacterias es consistente con los resultados obtenidos con otras enzimas metabólicas como la glutamato deshidrogenasa (Benachenhou-Lahfa et al., 1993), la glutamina sintasa (Tiboni et al., 1993), y la carbamolfosfato sintasa (Lazcano, Puente, y Gogarten, en prep), y parece apoyar la idea de una transferencia masiva de genes entre los ancestros de estas líneas celulares.

6. La demostración del papel que la duplicación génica jugó en el ensamblaje de diversas rutas biosintéticas no solamente es consistente con la hipótesis de Waley-Ycas-Jesen, sino que permite explicar la aparente rapidez con la que pudieron haber evolucionado las rutas anabólicas durante el Arqueano temprano. Esta conclusión abre una serie de preguntas hasta ahora poco exploradas. Se sabe los estados merodiploides son poco estables, y que en los procariontes contemporáneos la probabilidad de fijar en forma permanente una secuencia génica duplicada es extraordinariamente reducida. Por lo tanto, la presencia de enzimas homólogas en distintas rutas biosintéticas sugiere que hace mas de 3.5×10^9 años (i) un número considerable de duplicones eludió a los fenómenos de conversión génica; y (ii) que aunque las tasas de duplicación hayan tenido valores comparable a las actuales, antes de la trifurcación de los tres linajes las tasas de divergencia y fijación de genes deben haber tenido valores mayores que los contemporáneos.

DESIDERATA

El trabajo presentado en esta tesis forma parte de un proyecto mas ambicioso que pretende arrojar algunas pistas sobre el origen y la evolución temprana de la rutas biosintéticas comunes a todos los seres vivos. Sin embargo, existe una larga lista de saldos pendientes, entre los que se incluyen las siguientes cuestiones:

1. Si bien es cierto que los resultados presentado en este trabajo apoyan la idea de que las rutas biosintéticas se ensamblaron a partir del reclutamiento de de enzimas poco específicas, es evidente que este proceso solo pudo haber tenido lugar luego de la aparición de proteínas catalíticas. El origen de estas últimas sigue siendo una pregunta abierta. Sin embargo, es posible que haya habido etapas mas antiguas, durante las cuales los sistemas biológicos pueden haber dependido no solo de tanto de enzimas ambiguas y poco eficientes, como de reacciones químicas espontáneas (Lazcano y Miller, en prep). El análisis crítico de esta hipótesis abre la posibilidad de postular una época de la evolución metabólica mas antigua aún que la que se ha discutido en los textos de Fani et al. (1995), Alifano et al. (1996), y Mills et al (en prep), y que se podría denominar como una fase semi-enzimática. Actualmente se está aplicando este enfoque al estudio de la biosíntesis de la histidina (Leguina, Fani, Lazcano, y Miller, en prep), en donde se sabe que en presencia de concentraciones elevadas de NH_4^+, la adición del nitrógeno para formación del imidazol-glicerol-fosfato no requiere de la presencia de la glutaminamido-transferasa correspondiente (Martin et al., 1971; Klem y Davisson, 1993). En principio, otras reacciones equivalentes podrían existir, especialmente aquellas que involucran la isomerización y ruptura del grupo imidazol del ATP precursor.

2. El reconocimiento del papel central que el RNA y los ribonucleótidos parecen haber jugado en etapas tempranas de la evolución sugiere que las biosíntesis de los diversos componentes de los ácidos nucleicos deben ser consideradas como una de las rutas anabólicas mas antiguas. Por ello, se ha hipotetizado que el mecanismo sugerido por Waley-Ycas-Jensen pudo haber jugado un papel importante en su evolución temprana. Existen varias evidencias de que esto puede haber sido el caso. Sabemos, por ejemplo, que

las nucleosidasas que participan en las rutas de salvamento son homólogas a las fosforibosiltransferasas (Mushegian y Koonin, 1994). Por otra parte, el análisis preliminar de las secuencias de las enzimas que participan en las llamadas rutas de salvamento ha mostrado que, en efecto, la duplicación génica subyace el origen de los siguientes conjuntos de proteínas: (a) las purinil-fosforibosil transferasas; (b) la guanina reductasa y la inosina monofosfato deshidrogenasa; y (c) la adenosina desaminasa y la adenosina monofosfato desaminasa (Becerra y Lazcano, en prep.).

3. No existe ninguna razón para suponer que el proceso de ensamblaje de rutas metabólicas sugerido por Waley-Ycas-Jensen operó solo durante las épocas anteriores a la divergencia de los tres grandes linajes celulares. De hecho, la biosíntesis de la lisina parece constituir un buen candidato para probar la validez de este mecanismo. Es sabido que mientras que las eubacterias, las metanógenas, las plantas y muchos protistas sintetizan la lisina vía la utilización del ácido meso-diaminopimélico como un intermediario, los hongos, los critidios, y los blastocladiales utilizan una ruta distinta en donde participa el ácido α-aminoadípico (AAA) (Léjohn, 1974; Ragan, 1989). La ruta del AAA es tan característica de los hongos, que se considera un rasgo diagnóstico de este grupo. Su presencia en critidios y blastocladiales es congruente con las filogenias basadas en las secuencias de los 18S rRNAs de estos organismos y que demuestran su afinidad con los hongos (Sogin, 1994). Debido a que los organismos que sintetizan la lisina siguiendo la ruta del AAA son claramente posteriores a laos procariontes, se dispone de un modelo que en principio permitía explicar el surgimiento de una ruta alterna de biosíntesis de la lisina mediante la hipótesis de Waley-Ycas-Jensen. El análisis de las secuencias de las enzimas que participan en ambas rutas se está llevando a cabo con el propósito de analizar esta hipótesis (Leguina, Lazcano y Vogel, en prep.).

4. La disponibilidad pública de secuencias de los genomas completos de varios microorganismos nos ha permitido iniciar el análisis de operones con el propósito de comprobar el alcance de la hipótesis de Horowitz. En particular, el análisis de la estructura de los operones y las secuencias de las enzimas que determinan las biosintesis de aminoácidos, purinas, pirimidinas, nucleosidos, y nucleótidos en H. influenzae (Fleischmann et al.,

1995), y de la serina, el ácido fólico, las porfirinas, y el grupo heme en Mycoplasma genitalium (Fraser et al., 1995), ha permitido demostrar las limitaciones de la hipótesis retrógrada, por una parte y, por otra, permitirá detectar procesos de duplicación y elongación anteriores a la divergencia de las bacterias Gram positivas y Gram negativas (Islas, Silva y Lazcano).

5. Por último, hay varios aspectos de la evolución de los conjuntos de genes biosintéticos que hasta ahora no han sido explorados. Por ejemplo, la idea de que en el pasado tuvo lugar un reclutamiento masivo de enzimas poco específicas, o de que la especificidad hacia un nuevo substrato es resultado de procesos de duplicación y divergencia, sugiere que el manetnimiento de la nueva ruta metabólica es resultado de una coevolución de las distintas secuencias que la conforman. Este proceso es equivalente a la coevolución que sufren de dos o mas poblaciones, y que ha sido objecto de análisis detallados (cf. Roughgarden, 1983). Pretendo eventualmente acercarme a este problema, así como a las cuestiones ya mencionadas de las tasas de duplicación, fijación y conversión génica. Es igualmente interesante mencionar que el aumento de la especificidad enzimática y del número de componentes a nivel molecular en un proceso biológico implica, por si mismo, un mecanismo de regulación (Galas et al., 1986). Ello debe llevar asociado una serie de aspectos de costo/beneficio cuyo análisis está también pendiente y al que probablemente valdría la pena aproximarse.

Referencias

- Alifano, P., Fani, R., Liò, P., Lazcano, A., Bazzicalupo, M., Carlomagno, M. S., and Bruni, C. B. (1996) Histidine biosynthetic pathway and genes: structure, regulation, and evolution. Microbiol Rev. **(en prensa)**

- Anderson, R. P. and Roth, J. R. (1977) Tandem genetic duplications in phage and bacteria. Ann. Rev. Microbiol. **31**: 473-505

- Belfaiza, J., Parsot, C., Martel, A., Bouthier de la Tour, C., Maragarita, D., Cohen, G. N., and Saint-Girons, I. (1986) Evolution in biosynthetic pathways: two enzymes catalyzing consecutive steps in methionine biosynthesis originate from a common ancestor and possess a similar regulatory region. Proc. Natl. Acad. Sci. USA **83**: 867-871

- Brenner, S. E., Hubbard, T., Murzin, A., and Chothia, C. (1995) Gene duplications in H. influenza. Nature **378**: 140

- Brown, J. R. and Doolittle, W. F. (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthetases. Proc. Natl. Acad. Sci. USA **92**: 2441-2445

- Burke, D. H., Hearst, J. E., and Sidow, A. (1993) Early evolution of photosynthesis: clues from nitrogenase and chlorophyll proteins. Proc. Natl. Acad. Sci. USA **90**: 7134-7138

- Cánovas, J. L., Ornston, L. N., and Stanier, R. Y. (1967) Evolutionary significance of metabolic control systems. Science **156**: 1695-1698

- Chapman, D. J. and Ragan, M. A. (1980) Evolution of biochemical pathways: evidence from comparative biochemistry. Ann. Rev. Plant Physiol. **31**: 639-678

- Chyba, C. F., Thomas, P. J., Brookshaw, L., and Sagan, C. (1990) Cometary delivery of organic molecules to the early Earth. Science **249**: 366-373

- Clarke, P. H. (1983) Experimental evolution. In D. S. Bendall (ed), Evolution from Molecules to Man (Cambridge University Press, Cambridge), pp. 283-292

- Cox, J. L., Cox, B. J., Fidanza, V., and Calhoun, D. H. (1987) The complete nucleotide sequence of the ilvGMEDA cluster of Escherichia coli K12. Gene **56**: 185-198

- Doolittle, R. F. (1995) Of Archaea and Eo: what's in a name? Proc. Natl. Acad. Sci. USA **92**: 2421-2423

- Fani, R., Liò, P., Chiarelli, I., and Bazzicalupo, M. (1994) The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the hisA and HisF genes. J. Mol. Evol. **38**: 489-495

- Fani, R., Liò, P., and Lazcano, A. (1995) Molecular evolution of the histidine biosynthetic pathway. J. Mol. Evol. **41**: 760-774

- Fitch, W. M. and Upper, K. M. (1987) The phylogeny of tRNA sequences provides evidence of ambiguity reduction in the origin of the genetic code. Cold Spring Harbor Symp. Quant. Biol. **52**: 759-767

- Fleishmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedlom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brrandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Georghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., and Venter, J. C. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269**: 496-512

- Forterre, P., Benachenhou-Lahfa, N., Confalonieri, F., Duguet, M., Elie, C., and Labedan, B. (1993) The nature of the last common ancestor and the root of the tree of life, still open questions. BioSystems **28**: 15-32

- Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison III, C. A., and Venter, J. C. (1995) The minimal gene complement of Mycoplasma genitalium. Science **270**: 397-403

- Galas, D. J., Kirkwood, T. B. L., and Rosenberger, R. F. (1986) An introduction to the problem of accuracy. In T. B. L. Kirkwood, R. F. Rosenberger, and D. J. Galas (eds), Accuracy in Molecular Processes (Chapman and Hall, London), pp. 1-16

- Gogarten, J. P. (1994) Which is the most conserved group of proteins? Homology, orthology, paralogy, and the fusion of independent lineages. J. Mol. Evol. **39**: 541-543

- Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, J., Date, T., Oshima, T., Konishi, L., Denda, K., and Yoshida, M. (1989) Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA **86**: 6661-6665

- Hall, B. G. and Hauer, B. (1993) Acquisition of new metabolic activities by microbial populations. In E. A. Zimmer, T. J. White, R. L. Cann, and A. C. Wilson (eds), Molecular evolution: producing the biochemical data. Methods in Enzymology **224**: 603-613

- Horowitz, N. J. (1945) On the evolution of biochemical synthesis. Proc. Natl. Acad. Sci. USA **31**: 153-157

- Horowitz, N. J. (1965) The evolution of biochemical synthesis --retrospect and prospect. In V. Bryson and H. J. Vogel (eds) Evolving Genes and Proteins (Academic Press, New York), 15-23

- Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T. (1989) Evolutionary relationship of archeabacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA **86**: 9355-9359

- Jensen, R. A. (1976) Enzyme recruitment in evolution of new function. Ann. Rev. Microbiol. **30**: 409-425

- Joyce, G. F., Schwartz, A. W., Miller, S. L., and Orgel, L. E. (1987) The case for an ancestral genetic system involving simple analogues of the nucleotides. Proc. Natl. Acad. Sci. USA **84**: 4398- 4403

- Keefe, A. D. and Miller, S. L. (1995) Are polyphosphates or phosphate esters prebiotic reagents? J. Mol. Evol. **41**: 693-702

- Keefe, A. D., Lazcano, A., and Miller, S. L. (1995) Evolution of the biosynthesis of the branched-chain amino acids. Origins of Life and Evol. Biosph. **25**: 99-110

- Klem, T. J. and Davisson, V. J. (1993) Imidazole glycerol phosphate synthase: the glutamine amidotransferase in histidine biosynthesis. Biochemistry **32**: 5177-5186

- Labedan, B. and Riley, M. (1995) Widespread protein sequence similarities: origins of Escherichia coli genes. J. Bacteriol. **177**: 1585-1588

- Larralde, R., Roberston, M. P., and Miller, S. L. (1995) Rates of decomposition of ribose and other sugars: implications for chemical evolution. Proc. Natl. Acad. Sci. USA **92**: 8158-8160

- Lazcano, A. (1993) The significance of ancient paralogous genes in the study of the early stages of microbial evolution. In R. Guerrero and C. Pedrós-Alió (eds), Trends in Microbial Society (Spanish Society for Microbiology, Barcelona), 559-562

- Lazcano, A. (1994a) The RNA world, its predecessors and descendants. In S. Bengtson (ed), Early Life on Earth: Nobel Symposium No. 84 (Columbia University Press, New York), pp. 70-80

- Lazcano, A. (1994b) The transition from non-living to living. In S. Bengtson (ed), Early Life on Earth: Nobel Symposium No. 84 (Columbia University Press, New York), pp. 60-69

- Lazcano, A. (1995a) Alexandr I. Oparin: the man and his theory. In. B. F. Plogazov, B. I. Kurganov, M. S. Kritsky, and K. L. Gladilin (eds) Frontiers in Physicochemical Biology and Biochemical Evolution (Bakh Institute of Biochemistry and ANKO, Moscow), pp. 49-56

- Lazcano, A. (1995b) Alexander I. Oparin: apuntes para una biografía intelectual. In F. Morán, J. Peretó, y A. Moreno (eds), Orígenes de la Vida: en el centenario del nacimiento de A. I. Oparin (Editorial Complutense, Madrid), 15-39

- Lazcano, A. (1995c) Cellular evolution during the early Archean: what happened between the progenote and the cenancestor? Microbiologia SEM **11**: 185-198

- Lazcano, A. and Miller, S. L. (1994) How long did it take for life to appear and evolve to cyanobacteria? J. Mol. Evol. **39**: 546-554

- Lazcano, A., Fox, G. E., and Oró, J. (1992) Life before DNA: the origin and evolution of early Archean cells. In R. P. Mortlock (ed), The Evolution of Metabolic Function (CRC Press, Boca Ratón), pp. 237-339

- Lazcano, A., Díaz-Villagómez, E., Mills, T., and Oró, J. (1995) On the levels of enzymatic substrate specificity: implications for the early evolution of metabolic pathways. Adv. Space Res. **15**: 345-356

- Lazcano, A., Guerrero, R., Margulis, L., and Oró, J. (1988) The evolutionary transition from RNA to DNA in early cells. J. Mol. Evol. 27: 283-290

- Léjohn, H. B. (1974) Biochemical parameters of fungal phylogenetics. In T. Dobzhansky, M. K. Hecht, and W. C. Steere (eds), Evolutionary Biology **7** (Plenum Press, New York): 79-123

- Lewis, E. B. (1951) Pseudoallelism and gene evolution. Cold Spring Harbor Symp. Quant. Biol. **16**: 159-174

- Martin, R. G., Berberich, M. A., Ames, B. C., Davis, W. W., Goldberger, R. F., and Yourno, J. D. (1971) Enzymes and intermediates of histidine biosynthesis in Salmonella typhimurium. Methods Enzymol. **17B**: 3-44

- Mortlock, R. P. (1984) Microorganisms as Model Systems for Studying Evolution (Plenum Press, New York)

- Mushegian, A. R. and Koonin, E. V. (1994) Unexpected sequence similarity between nucleosidases and phosphoribosyltransferases of different specificity. Protein Science **3**: 1081-1088

- Oparin, A. I. (1924) Proizkhozhdenie Zhisni (Moscowkij Rabotchij, Moscow)

- Oparin, A. I. (1938) Origin of Life (Macmillan, New York)

- Ornston, L. N. (1971) Regulation of catabolic pathways in Pseudomonas. Bacteriol. Rev. **35**: 87-98

- Ragan, M. A. (1989) Biochemical pathways and the phylogeny of the eukaryotes. In B. Fernholm, K. Bremer, and H. Jörnvall (eds), The Hierarchy of Life (Elsevier Science Publ., Dordrecht), pp. 145-160

- Roughgarden, J. (1983) The theory of coevolution. In D. J. Futuyma and M. Slatkin (eds), Coevolution (Sinauer Associates, Sunderland), pp. 33-64

- Schopf, J. W. (1993) Microfossils of the early Archean Apex chert: new evidence of the antiquity of life. Science **260**: 640-646

- Sleep, N. H., Zahnle, K. J., Kasting, J. F., and Morowitz, N. H. (1989) Annhilation of ecosystems by large asteroid impacts on the early Earth. Nature **342**: 139-142

- Sogin, M. L. (1994) The origin of eukaryotes and evolution into major kingdoms. In S. Bengtson (ed), Early Life on Earth: Nobel Symposium No. 84 (Columbia University Press, New York), pp. 181-192

- Waley, S. G. (1969) Some aspects of the evolution of metabolic pathways. Comp. Biochem. Physiol. **30**: 1-7

- Walsh, J. B. (1987) Sequence-dependence gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics **117**: 543-557

- Woese, C. R. (1983) The primary lines of descent and the universal ancestor. In D. S. Bendall (ed), Evolution from Molecules to Man (Cambridge University Press, Cambridge), pp. 209-233

- Woese, C. R. (1987) Bacterial evolution. Microbiol. Rev. **51**: 221-271

- Woese, C. R. and Fox, G. E. (1977) The concept of cellular evolution. J. Mol. Evol. **10**: 1-6

- Ycas, M. (1974) On the earlier states of the biochemical system. J. Theor. Biol. **44**: 145-160