UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS
DIVISION DE ESTUDIOS DE POSGRADO

EL PAPEL DE LA DUPLICACION GENICA EN LA EVOLUCION TEMPRANA DE LA VIDA

T E S I S

QUE PARA OBTENER EL GRADO ACADEMICO DE

DOCTOR EN CIENCIAS (BIOLOGIA)

P R E S E N T A :

ARTURO CARLOS II BECERRA BRACHO

DIRECTOR DE TESIS:
DR. ANTONIO LAZCANO ARAUJOREYES

MEXICO, D. F.

2000

A mi familia y amigos

Agradecimientos:

Especialmente al Dr. Lazcano y a todos los compañeros del laboratorio: Sara, Luis, Ana, Amanda, Erwin, Ulises, Beto, Hector, Rossana, los tallarines y demas anexos, por su amistad y apoyo

A mi comité tutorial: Dra. Valeria Souza, Dr. Luis Felipe Jiménez y al Dr. Antonio Lazcano

Al apoyo recibido por el proyecto PAPIIT IN213598

A mi universidad, la UNAM

Estructura de la tesis

Este trabajo de tesis está conformado por cuatro manuscritos, de los cuales dos son artículos de investigación y dos son capítulos de libros. Los textos han sido aceptados o publicados en revistas y libros con arbitraje internacional, con excepción de uno de los artículos, el cual próximamente será enviado al Comité Editorial del Journal Molecular Evolution. Los trabajos son:

- EARLY METABOLIC EVOLUTION: INSIGHTS FROM COMPARATIVE CELLULAR GENOMICS. Islas S., Becerra A., Leguina J.I., and Lazcano A. In J. Chela-Flores and F. Raulin (eds.), *Exobiology: Matter, Energy, and Information in the Origin and Evolution of Life in the Universe,* 167-174. 1998 Kluwer Academic Publisher. Printed in Netherlands

- MOLECULAR BIOLOGY AND THE RECONSTRUCTION OF MICROBIAL PHYLOGENIES: DES LAISIONS DANGEREUSES?. Becerra A., Velasco A., Silva E., Islas E. And Lazcano. In Chela-Flores J., Lemerchand G., and Oro J. (eds.). *Astrobiology: Origins from the Big-Bang to the civilisation.* Kluwer Academic Publisher. Printed in Netherlands (en prensa)

- PHYLOGENETIC DISTRIBUTION OF SIMPLE SEQUENCES: INSIGHTS FROM COMPARATIVE GENOMICS. Becerra A. And Lazcano A. Para ser enviado *a Journal of Molecular Evolution.*

- COMPARATIVE BIOCHEMISTRY OF CO_2 FIXATION AND THE EVOLUTION OF AUTOTROPHY. Peretó J.G., Velasco A. M., Becerra A., and Lazcano A.1999 *International Microbiology* 2:3-10

INTRODUCCION

I Evolución temprana de la vida y el último ancestro universal

El uso de la subunidad pequeña del rRNA como marcador molecular ha permitido desarrollar una filogenia universal, en donde todos los organismos quedan agrupados en alguno de los tres principales linajes celulares, eubacteria, arqueobacterias y eucariontes, ahora denominados como los dominios: Bacteria, Arquea y Eucarya (Woese et al., 1990). Este árbol evolutivo sin raíz, que resultó de la comparación de las secuencias de genes ortólogos de rRNA, se trifurca a partir de un ancestro común, al que Woese y Fox (1 977) denominaron *progenote*. Debido a que no se ha descubierto un organismo que pueda servir como grupo externo a estos tres linajes celulares, el *progenote* fue definido no sólo 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 fenotípo y genotipo aún no había tenido lugar (Woese y Fox, 1977). Años mas tarde, Woese (1983, 1987) continuó desarrollando su hipótesis y propuso que el *progenote* era un sistema donde el material hereditario estaba constituído por moléculas fragmentadas de RNA que aún no estaban integradas en un solo polímero genético.

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 contemporáneas, y propusieron que el árbol del rRNA se trifurcaba no a partir de un *progenote* sino de un organismo complejo al que denominaron *cenancestro*. Por otra parte, la comparación de las secuencias homólogas comunes a organismos de los tres linajes permitió proponer que el ancestro común del árbol de rRNA era, en realidad, una célula procarionte dotada

de los mismos rasgos biológicos de una bacteria contemporánea (Lazcano et al., 1992; Lazcano, 1995). A pesar de que no se puede excluir del todo ni la posibilidad de que hayan ocurrido fenómenos de transporte horizontal entre los tres linajes, ni de que hayan sucedido pérdidas secundarias durante la divergencia de las ramas (Becerra et al., 1997), la caracterización del ancestro común a estos tres grupos como una célula procarionte compleja sugiere la existencia de una fase de evolución biológica previa a la 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 más antiguo de todas las formas actuales de vida ha ido ganando una aceptación creciente (Brown y Dolittle, 1995), es igualmente cierto que existen anomalías aún no explicadas, entre las que se incluyen las filogenias construidas con secuencias de glutamato deshidrogenasas, glutamino sintetasas (Forterre et al., 1993), carbamoil-fosfato sintetasas, proteínas de choque térmico y otras más. Ello ha llevado a sugerir que en el pasado pudo haber ocurrido un transporte masivo de genes entre los ancestros de bacterias gram positivas y las arqueobacterias (Gogarten, 1994).

La descripción del último ancestro común a los tres linajes celulares puede ser inferida por la distribución de los caracteres homólogos entre sus descendientes. La disponibilidad de secuencias provenientes de genomas completos en bases con acceso publico ha aumentado las herramientas para la caracterización de la naturaleza del *cenancestro*. Sin embargo, las diferencias en el repertorio metabólico tanto como en los mecanismos de expresión de genes dentro de los tres dominios (Olsen and Woese, 1997), demuestra que la caracterización del ultimo ancestro común está aún lejos y

por el momento subsiste una fuerte controversia sobre su naturaleza (Doolittle 2000).

En principio, se pueden intertar reconstrucciones de estados ancestrales con una metodología relativamente simple. Dada la disponibilidad de secuencias de genomas completos de los tres dominios celulares, las características del cenancestro estarían definidas por las propiedades que se comparten en todos los organismos vivientes, menos aquellos que son el resultado de evolución convergente y los que son adquiridos por transporte horizontal (Figura 1). Sin embargo, el análisis de la intersección de genomas puede complicarse por las proteínas que no han sido identificadas, además de aquellas que han modificado su secuencia rápidamente y por ende no se reconoce su ancestría común. Además, se pueden hacer inferencias incorrectas si no se considera las pérdidas secundarias que se presentan en los descendientes, en especial aquéllos que son parásitos (Becerra et al., 1997).

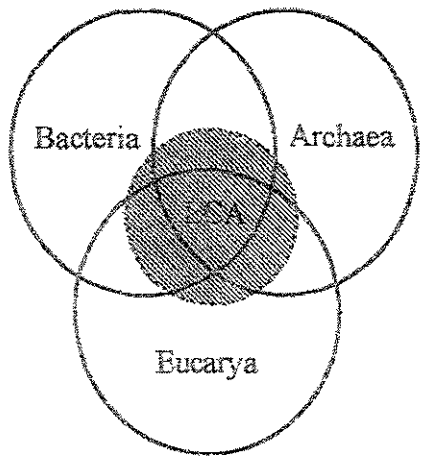


Figura 1. Caracterización del último Ancestro Común (LCA)

II El papel de las duplicaciones génicas en la evolución de rutas metabólicas

El primer intento por explicar el origen de las rutas metabólicas proviene del trabajo de Horowitz (1945), quien sugirió que el desarrollo y evolución de las vías biosintéticas es el resultado de una adquisición secuencias de enzimas, pero en un orden inverso al que actualmente poseen en una vía dada. Esta propuesta, basada en las ideas de Oparin (1938), establece una conexión evolutiva entre el ambiente primitivo y la emergencia del metabolismo biológico, al sugerir que los intermediarios bioquímicos de las rutas metabólicas básicas estaban ya presentes en el medio prebiótico. Tiempo después, el descubrimiento de los operones llevó a Horowitz (1965) a desarrollar su idea ya conocida para entonces como la hipótesis retrógradas, al proponer que el agrupamiento físico de algunos genes biosintéticos, es el resultado precisamente de una sucesión temprana de duplicaciones génicas en tandem.

Sin embargo, es fácil encontrar una serie de objeciones a la hipótesis retrógrada. La mayoría de los intermediarios metabólicos son químicamente inestables, y su acumulación en la sopa primitiva resulta difícil de explicar. Además, muchos de estos intermediarios son compuestos fosforilados, los cuales no podrían permear fácilmente las membranas primordiales en ausencia de un sistema de transporte especializado. Una propuesta alternativa sobre el papel de la duplicación génica en el establecimiento de rutas anabólicas fue desarrollada independientemente por Waley (1969), Ycas (1974) y Jensen (1976). Según estos autores, las rutas biosintéticas fueron ensambladas por un mecanismo de "patchwork" o "bricollage" en que participaron enzimas primitivas con una baja especificidad al sustrato. De acuerdo con esta idea, las rutas metabólicas estaban catalizadas por enzimas que podían usar diferentes substratos parecidos. De esta manera, la duplicación génica y la divergencia evolutiva de las secuencias resultantes fue el mecanismo que incrementó el tamaño de los genes

y aumentó la especificidad de las enzimas al substrato. Diferentes ejemplos de este mecanismo han sido descritos (Becerra y Lazcano 1998, Islas et al.,1998), lo que demuestra que la duplicación génica jugó un papel importante en la evolución temprana de la vida, al menos en el ensamblaje de las rutas metabólicas. Sin embargo, es claro que debieron existir mecanismos adicionales que participaron en el incremento del material genético, ya que no todos procesos metabólicos pueden ser explicados por eventos tipo *patchwork*.

III El papel de las secuencias simples en la evolución temprana de la vida

Uno de los problemas centrales en el estudio de la evolución temprana de la vida es entender cual es el origen del material genético a partir de lo que debieron haber sido genomas extremadamente reducidos. La disponibilidad de secuencias de genomas celulares completos provenientes de los tres linajes (Arquea, Bacteria, y Eucarya), ha permitido analizar este fenómeno, encontrando que la duplicación génica junto con el transporte horizontal y la simbiosis son los principales mecanismos que han incrementado el material genético. Sin embargo, pudieron existir otros mecanismos que incrementaron él numero y tamaño de los genes durante la evolución temprana de la vida. Uno de estos mecanismos pudo haber sido la generación de secuencias simples. Las secuencias simples son regiones de proteínas y ácidos nucleicos que presentan un sesgo composicional en sus residuos y típicamente contienen segmentos repetitivos, es decir, los llamados segmentos de baja complejidad. Los segmentos de baja complejidad son producidos por mutaciones tipo 'slipped-strand' durante la replicación del material genético (Fig.2). Si bien la alta frecuencia de secuencias simples ha sido detectada tanto en genomas eucariontes como en bases de datos (Britten and Kohne 1968; Tautz and Renz 1984; Wootton and Federhen 1993), recientemente se ha cobrado conciencia de que este fenómeno también ocurre en procariontes (Saunder et al. 1998; este trabajo). Debido a que las secuencias simples

y aumentó la especificidad de las enzimas al substrato. Diferentes ejemplos de este mecanismo han sido descritos (Becerra y Lazcano 1998, Islas et al.,1998), lo que demuestra que la duplicación génica jugó un papel importante en la evolución temprana de la vida, al menos en el ensamblaje de las rutas metabólicas. Sin embargo, es claro que debieron existir mecanismos adicionales que participaron en el incremento del material genético, ya que no todos procesos metabólicos pueden ser explicados por eventos tipo *patchwork*.

III El papel de las secuencias simples en la evolución temprana de la vida

Uno de los problemas centrales en el estudio de la evolución temprana de la vida es entender cual es el origen del material genético a partir de lo que debieron haber sido genomas extremadamente reducidos. La disponibilidad de secuencias de genomas celulares completos provenientes de los tres linajes (Arquea, Bacteria, y Eucarya), ha permitido analizar este fenómeno, encontrando que la duplicación génica junto con el transporte horizontal y la simbiosis son los principales mecanismos que han incrementado el material genético. Sin embargo, pudieron existir otros mecanismos que incrementaron él numero y tamaño de los genes durante la evolución temprana de la vida. Uno de estos mecanismos pudo haber sido la generación de secuencias simples. Las secuencias simples son regiones de proteínas y ácidos nucleicos que presentan un sesgo composicional en sus residuos y típicamente contienen segmentos repetitivos, es decir, los llamados segmentos de baja complejidad. Los segmentos de baja complejidad son producidos por mutaciones tipo 'slipped-strand' durante la replicación del material genético (Fig.2). Si bien la alta frecuencia de secuencias simples ha sido detectada tanto en genomas eucariontes como en bases de datos (Britten and Kohne 1968; Tautz and Renz 1984; Wootton and Federhen 1993), recientemente se ha cobrado conciencia de que este fenómeno también ocurre en procariontes (Saunder et al. 1998; este trabajo). Debido a que las secuencias simples

se han encontrado tanto en secuencias codificantes como no codificantes, se ha sugerido que juegan un papel importante como fuente de variabilidad genética y en la evolución del tamaño de genomas (Tautz et al., 1986; Hancock 1995). Debido a su hipermutabilidad, las secuencias simples son reconocidas como una fuente importante de variación fenotípica, especialmente dentro de los procariontes patógenos (Moxon et al., 1994; Moxon 1999). Sin embargo, el papel de las secuencias de baja complejidad pudo ser relevante en la evolución temprana de la vida, dado que es un mecanismo rápido de amplificación del material genético. Así, junto con la duplicación interna, la transferencia horizontal y la simbiosis, pueden haber jugado un papel importante en la evolución del tamaño y de las propiedades codificantes de los genes celulares.

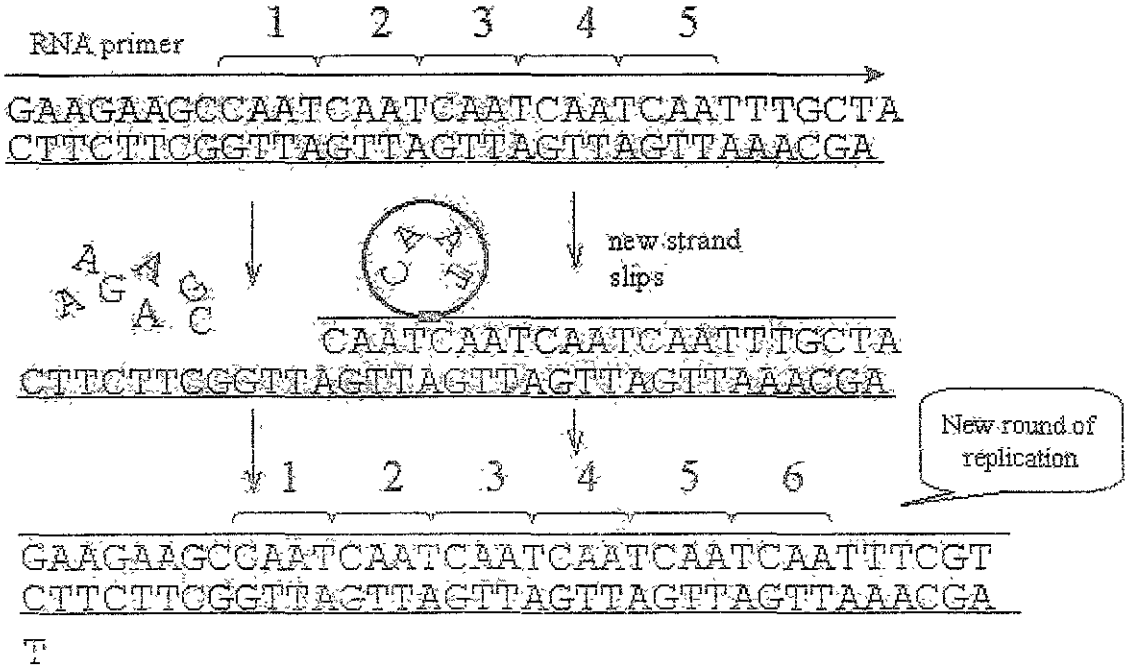


Figura 2. Proceso de "slipped-strand", donde un apareo incorrecto de las cadenas de DNA genera la amplificación de la secuencia simple.

Para conocer el papel de las secuencias simples en la evolución temprana de la vida se analizaron los siguientes proteomas completos de los siguientes organismos : *Escherichia coli* (Blattner et. al., 1997), *Haemophilus influenzae* (Fleischmann et. al. 1995), *Helicobacter pylori* 26695 (Tomb et. al. 1997), *H. pylori* J99 (Alm et. al. 1999), *Rickettsia prowazekii* (Anderson 1998), *Bacillus subtilis* (Kunst et al. 1997), *Mycoplasma genitalium* (Fraser et al. 1996), *M. pneumoniae* (Himmelreich et. al. 1996), *Mycobacterium tuberculosis* (Cole et al.1998), *Chlamydia trachomatis* (Stephens et al. 1998), *Campylobacter jejuni* (Parkhill et. al. 2000), *Borrelia burgdorferi* (Fraser et al. 1997), *Treponema pallidum* (Fraser et al. 1998), *Aquifex aeolicus* (Deckert et al. 1998), *Synechocystis sp.* (Kaneko et al. 1996), *Deinococcus radiodurans* (White et. al. 1999), *Thermotoga maritima* (Nelson et. al. 1999), *Methanococcus jannaschii* (Bult et al. 1996), *Methanobacterium thermoautotrophicum* (Smith et al. 1997), *Archaeoglobus fulgidus* (Klenk et al. 1997), *Pyrococcus horikoshii* (Kawarabayasi et al.1998), y los proteomas de *Saccharomyces cerevisiae* (Goffeau et. al. 1997), y *Caenorhabditis elegans.* Utilizando el programa SEG (segment sequences by local complexity program; Wootton and Federhen 1993), el cual cuantifica la complejidad de las secuencias. Para ello se utilizó el valor de granulidad: 12 2.0 2.2 donde el inicio de ventana (trigger) fue de W= 12, el valor de complejidad K1= 2.0; la extensión de complejidad K2= 2.2. El programa SEG se complemento con un programa en *shell unix*, lo que permitió la identificación de secuencias con segmentos de baja complejidad (programa 1). La composición de aminoácidos de las regiones con baja complejidad fue calculada usando el programa *aacomp* de la paquetería de FASTA (Pearson and Lipman 1988). Todo lo anterior permite conocer parte de la naturaleza de las secuencias simples y por lo tanto indagar en cual fue el papel de este fenómeno en el incremento del material genético en la evolución temprana de la vida.

IV La evolución de la autotrofía: bioquímica comparada de la fijación del CO_2

Diversos mecanismos de fijación biológica de CO_2, explican la diversidad y éxito evolutivo de los organismos autotróficos. Según la formulación clásica de la teoría heterótrofa del origen de vida (Oparin 1938), una vez que el suministro de compuestos orgánicos abióticos se volvió un factor limitante, las primeras células desarrollaron otras maneras de obtener carbono y energía. Esto llevó al desarrollo de los primeros organismos fotoautótrofos, y luego a la fotosíntesis (Oparin 1938). Existen diferentes evidencias que apoyan la antigüedad de la ruta de reducción de la pentosa-fosfato, o ciclo de Calvin-Benson. Ello incluye (a) la existencia de microfósiles tipo cianobacterias en sedimentos australianos con 3.5 x 10^9 años de edad, lo que sugiere que el ciclo Calvin-Benson apareció durante el Arqueano temprano (Schopf 1993); y (b) los perfiles de fragmentación isotópica del ciclo del carbono en el Arqueano temprano, que son consistente con el proceso de fijación del carbono catalizada por la RuBisco (Hayes, 1994). Sin embargo, la filogenia universal basada en el 16/18S rRNA indica que la fotosíntesis basada en clorofila se desarrolló relativamente tarde en la rama bacteriana (Woese 1987; Pace 1997). Por lo tanto los primeros organismos autótrofos debieron usar energía química en lugar de energía lumínica, por lo que la ruta de reducción de la pentosa-fosfato, debe haber sido precedida por mecanismos más antiguos de asimilación de CO_2.

La asimilación del dióxido de carbono es un muy extendido, pero las diferencias bioquímicas entre las rutas metabólicas sugiere que ésta habilidad evolucionó de manera convergente en grupos de procariontes ampliamente separados. Además del ciclo Calvin-Benson, hay otros mecanismos de asimilación de CO_2 que incluyen (a) la reducción del ácido cítrico, o ciclo de Arnon; (b) el ciclo de reducción de la acetil-CoA, (o ruta de Wood pathway); y (c) otro mecanismo menos común, la llamada ruta del

hidroxipropionato, descrita inicialmente en *Chloroflexus*, una cianobacteria fotosintética no sulfurosa.

Hace veinte años se argumentó que el ciclo de ribulosa-difosfato evolucionó a partir del ciclo de la ribulosa monofosfato en una población hetreotrófica ancestral (Quayle y Ferencia 1978). Recientemente, Wächtershäuser (1990) ha propuesto un esquema quimioatotrófico para el origen de vida, en donde la formación de pirita se relaciona con la fijación del CO_2. Nosotros proponemos que ninguna de estas dos alternativas es correcta, y que la distribución filogenéticas de las rutas Arnon y Wood-pathway no permite deducir cual de estos dos ciclos es más antiguo. También argumentamos que existen evidencias experimentales que sugieren que una ruta semi-enzimática tipo Wood pudo haber sido uno de los primeros procesos biológicos de fijación del carbón.

EARLY METABOLIC EVOLUTION: INSIGHTS FROM COMPARATIVE CELLULAR GENOMICS

S. ISLAS, A. BECERRA, J. I. LEGUINA, and A. LAZCANO
Facultad de Ciencias-UNAM
Apdo. Postal 70-407, Cd. Universitaria
04510 México D. F., MÉXICO

1. Introduction

The use of small subunit rRNAs as molecular markers has led to universal phylogenies, in which all known organisms can be grouped in one of three major cell lineages, the eubacteria, the archaeabacteria, and the eukaryotic nucleocytoplasm, now referred to as the domains Bacteria, Archaea, and Eucarya, respectively (Woese *et al.*, 1990). A description of the last common ancestor (LCA, i.e., the cenancestor), of these three primary kingdoms may be inferred from the distribution of homologous characters among its descendants. In conjunction with the fragmentary information available from other organisms, the complete genome sequences now available in the public databases allow such characterizations, and in some cases can even provide insights into the nature of the cenancestor predecessors. Here we discuss the basic assumptions and strategies used in such approaches, and apply them to the understanding of the evolutionary assemblage of arginine biosynthesis. Additional aspects of the evolution of metabolic routes have been discussed in Peretó *et al.* (1997).

2. Some problems in comparative genomic analysis

The distribution of many biosynthetic enzymes found in all three primary lines of descent before complete genome sequences became available had already led to the idea that the cenancestor was comparable to modern prokaryotes in its biological complexity, ecological adaptability, and evolutionary potential (Lazcano, 1995). However, the differences in the metabolic repertoire and gene expression mechanisms among the three primary domains (cf. Olsen and Woese, 1997) demonstrate that the characterization of the LCA is an unfinished task, and that strong statements and broad generalizations should be avoided.

J. Chela-Flores and F. Raulin (eds.),
Exobiology: Matter, Energy, and Information in the Origin and Evolution of Life in the Universe, 167–174.

In principle, backtrack reconstructions of ancestral states can be achieved with a simple, straightforward methodology. Given the availability of complete genome sequences from the three primary domains, the cenancestor is defined by properties shared by all living organisms, minus those that are the outcome of convergent evolution and those acquired by horizontal transfer (Figure 1). However, cross-genomic analysis can be difficulted by unidentified proteins encoded by rapidly evolving sequences, as well as from the properties of a given genomic dataset. Inferences on the nature of the LCA can also be biased by the reduced DNA content of parasites and pathogens such as the mycoplasma, which have been selected as model organisms because of their small, compact genomes (Becerra *et al.*, 1997). Although the application of shotgun sequencing has led to an impressive growth of the databases in a very short time, larger volumes of complete genome sequences reflecting a broader cross-section of biological diversity are still required.

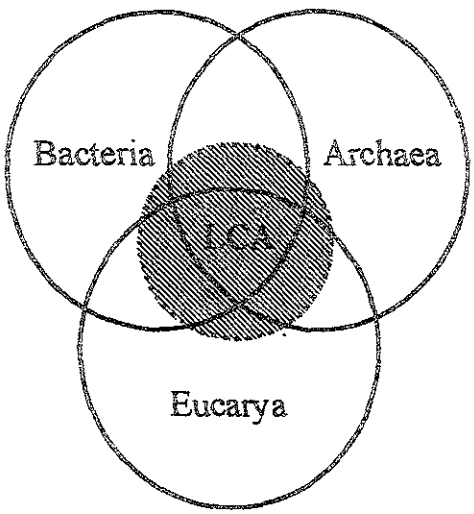


Figure 1. Intersection of the complete sequence spaces of the three domains defines the gene complement of the common ancestor (LCA). Identification of rapidly-evolving sequences would lead to a bigger set of ancestral genes (hatched areas).

The functions of many open reading frames (ORFs) derived from complete genome sequencing projects have been tentatively identified by computer searches based on structural similarities to known sequences in databases, but many more remain unidentified (30 to 50%, depending on the organism). Such databases are collections of the sequences that make up biological systems, but understanding how each component works is not enough for a proper description of how the entire system proceeds (Kanehisa, 1997). For instance, in the *Bacillus subtilis* tryptophan operon no sequence encodes the glutamine amido transferase required for anthranilate biosynthesis. This

would pose a problem in comparative genomic-based metabolic reconstructions, had biochemical experimentation not demonstrated that in *B. subtilis* the required gene is shared with the folate biosynthetic route, in whose operon it is located (Crawford, 1989).

As summarized in Table I, understanding of the evolutionary development of metabolism can be obscured by a complex series of changes involving enzymatic additions, secondary losses, pathway replacements, and functional redundancies. Additional complications can result from (a) intraespecific enzyme substitutions involving paralogous proteins; (b) that possibility that extant enzymes may have participated in alternative routes which no longer exist or remain to be discovered (Zubay, 1993; Becerra and Lazcano, 1997); (c) homologous enzymes endowed with widely different catalytic properties (see below); and (d) intracellular horizontal transfer within nucleated cells (Embley *et al.*, 1997).

TABLE I. Some processes in metabolic evolution.

process	examples	reference
addition of enzymatic step(s)	oxygen-dependent cholesterol biosynthesis	Bloch (1994), Ourisson and Nakatani (1994)
	archaeal biosynthesis of 2,3-di-*O*-phytanil *sn*-glycerol	Stetter (1996)
loss of routes and enzymes	purine biosynthesis in parasites	Becerra *et al.* (1997)
pathway replacement	aerobic instead of anaerobic biosynthesis of monounsaturated fatty acids	Bloch (1994)
	fungal lysine biosynthesis	Vogel (1960)
functional redundacies	phosphatidylcholine biosynthesis	Bloch (1994)
	imidazole biosynthesis in purine and histidine biosyntheses	Peretó *et al.* (1997)

3. Did metabolism evolve backwards?

The first attempt to explain the emergence of metabolic pathways was developed by Horowitz (1945), who suggested that biosynthetic enzymes had been acquired via gene duplications that took place in reverse order as found in extant pathways. This idea, also known as the retrograde hypothesis, established an evolutionary connection between the primitive soup and the development of metabolic pathways, and is frequently invoked in descriptions of early biological evolution (cf. Peretó *et al.*, 1997). Prompted by the discovery of operons, Horowitz (1965) restated his model, arguing

that it was supported not only by the overlap between the chemical structures of products and substrates of the enzymes catalyzing successive reactions, but also by the clustering of functionally related genes.

Although some operon-like gene clusters are found in both bacterial and archaeal genomes, whole genome comparisons between distant prokaryotes have shown that gene order can be easily eroded by extensive shuffling events (Mushegian and Koonin, 1996). This implies that the distribution in prokaryotic chromosomes of homologous genes encoding pathway enzymes cannot be used to (dis)prove the Horowitz hypothesis. However, if the enzymes catalyzing successive steps in a given metabolic pathway resulted from a series of gene duplication events (Horowitz, 1965), then they must share structural similarities. The known examples confirmed by sequence comparisons that satisfy this condition are limited to few pairs of enzymes and have been discussed elsewhere (cf. Pereto *et al.*, 1997).

4. The patchwork assemblage of biosynthetic routes

An alternative interpretation of role of gene duplication in the evolution of metabolism has been developed in the so-called patchwork hypothesis (cf. Jensen, 1976). According to this scheme, biosynthetic routes were assembled by primitive catalysts that could react with a wide range of chemically related substrates. The recruitment of enzymes from different metabolic pathways to serve novel catabolic routes under strong selective pressures is well document under laboratory conditions. Repeated occurrences of homologous enzymes in different pathways provide independent evidence of patchwork tinkering. Data derived from the ongoing genome projects has already demonstrated that a large portion of each organisms genes are related to each other as well as to genes in distantly related species. As discussed in the following section, the central role that gene duplication and recruitment have played in the assemblage of histidine anabolism (Alifano *et al.*, 1996) and purine nucleotide salvage pathways (Becerra and Lazcano, 1997) can also be extended to include arginine biosynthesis.

5. Gene duplication and arginine anabolism

The phylogenetic distribution of arginine biosynthetic genes suggest that this route was already present in the LCA. Hence, its absence in both *Helicobacter pylori* and the mycoplasma probably reflects polyphyletic secondary losses. Although the same chemical steps involved in arginine biosynthesis have been found in all organisms studied, two different strategies for the deacetylation of the intermediate *N*-acetylornithine have been described. In enterobacteria, the genus *Bacillus*, and the archaeon *Sulfolobus solfataricus* this reaction is catalyzed by *N*-acetylornithinase, the

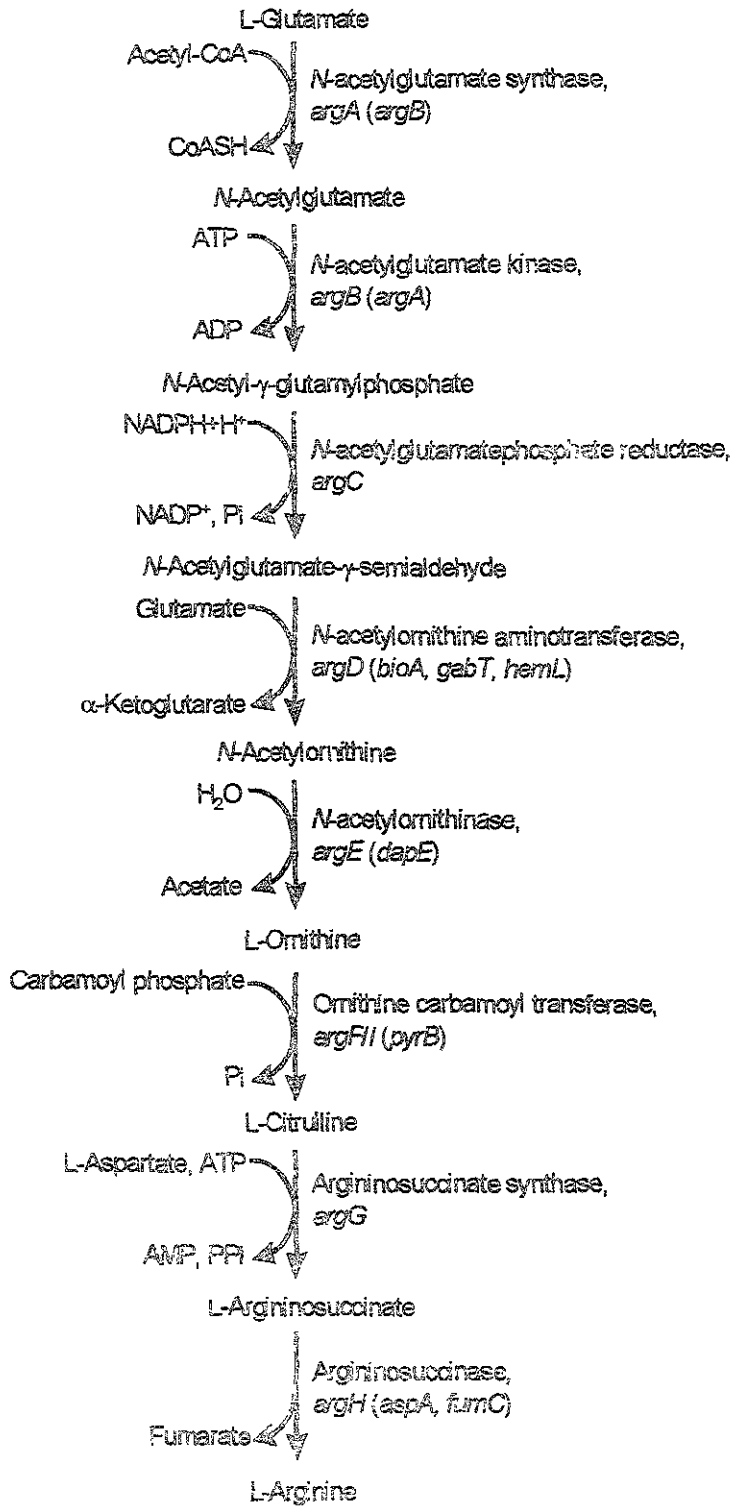


Figure 2. Arginine biosynthesis. The arginine biosynthetic genes paralogs are indicated within parenthesis.

gene product of *argE* (Figure 2), while in other prokaryotes and in fungi the acetyl group is removed by ornithine-glutamate acetyltransferase. There is no evidence of phylogenetic relationship between these two different enzymes. Another variation in this pathway occurs in the *E. coli* K12 strain, where two homologous genes (*argI*, *argF*) encode a family of four trimeric isoenzymes, that bind to L-ornithine and carbamoyl-phosphate to produce L-citrulline (Glansdorff, 1996).

Arginine biosynthesis consists of eight steps, five of which are mediated by enzymes that belong to different paralogous families (Figure 2). The list includes the pairs *argA/argB*, *argE/dapE*, and *argI/argF*, and the three- and four member families *argH/aspA/fumC* and *argD/bioA/gabT/hemL* (Riley and Labedan, 1996). Although the first two consecutive reactions in the pathway are catalyzed by the gene products of homologous sequences (*argA* and *argB*), we do not consider this as conclusive proof of the retrograde mechanism. Both reactions are chemically equivalent, and during the early evolution of this route they may have been catalyzed by an ancestral less-specific enzyme. Arginine biosynthesis thus provides additional evidence of the role of enzyme recruitment in metabolic evolution.

6. Homologous enzymes can have different catalytic properties

With the exception of proteins in which the evolutionary accretion of a functional motif or module has led new catalytic or binding properties, all enzymes encoded by paralogous genes can be expected to be endowed with comparable biochemical properties. However, reports on the existence of homologous enzymes that catalyze separate and mechanistically different reactions (Neidhart *et al.*, 1990) prompted us to search for additional examples in the available databases.

This analysis was performed using the database assembled by Riley and Labedan (1996), who compared the *E. coli* 1,862 protein sequences available as of April 1996 in the SwissProt databank. They concluded that 52.17% of all studied protein sequences had resulted from gene duplications, and classified them in paralogous families defined by sequence similarity. Their list includes 112 small families with only two sequences, 38 with three, 41 with three to seven, and 13 large families. As noted by Riley and Labedan, most of the members of paralogous families share comparable biochemical properties, with a scarce 1.23% of homologous protein pairs displaying what appear to be different functions.

We have repeated this analysis by looking exhaustively at all the characterized paralogous genes, and excluding from our sample 88 ORFs reported as hypothetical proteins. The resulting set was cross-checked with experimental data and the corresponding Enzyme Comission (EC) number. We have found a higher number of homologous genes with different EC numbers, which will be described elsewhere. An example is shown in Figure 3. It includes argininosuccinate lyase, which catalyzes the last step in arginine biosynthesis (Figure 2), and its homologs aspartate ammonia-lyase

at takes part in the synthesis and interconversion of aspartate and asparagine), a
narate hydratase (which participates in the tricarboxylic acid cycle). As denoted by
eir corresponding EC number, these enzymes catalyze different reversible reactions
ion-hydrolytic cleavage, (de)amination, and a hydration reaction, respectively).
lowever, all three of them use fumarate as substrate, which suggest that the structural
basis for their sequence similarity may be a large homologous binding site for this
compound.

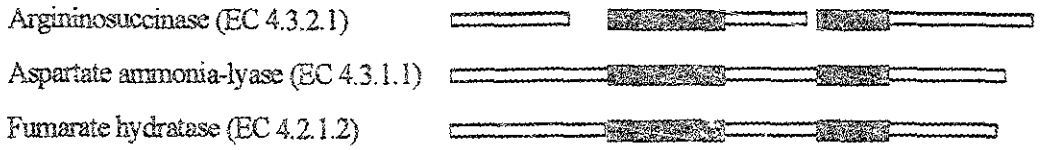


Figure 3. A three-member family of *E. coli* paralogous enzymes which different catalytic properties. The sequences were aligned using the Macaw program. The regions with statistically significant sequence similarity are shown in black.

7. Conclusions

The discovery that homologous enzymes that catalyze similar biochemical reactions are found in many different anabolic pathways supports the idea that enzyme recruitment took place at a massive scale during the early development of anabolic pathways. This conclusion is supported by analysis of the available genomic databases, which suggest that approximately 50% of cellular DNA is the outcome of paralogous duplications that may have preceded the divergence of the three primary domains. Such high levels of redundancy suggest that the wealth of phylogenetic information older than the cenancestor itself may be larger than realized, and that this information may provide fresh insights into a crucial but largely unexplored stage of early biological evolution.

Acknowledgements

The work of J. I. L. has been supported in part by the Consejo Superior de Investigaciones Cientificas (CSIC, Madrid, Spain). This paper was completed during a leave of absence of one of us (A. L.) as Visiting Professor at the Institut Pasteur (Paris), during which he enjoyed the hospitality of Professor Henri Buc and his associates at the Unité de Physicochimie des Macromolécules Biologiques. Support from the Manlio Cantarini Foundation (A. L.) is gratefully acknowledged. A.L. is an Affiliate of the NSCORT (NASA Specialized Center for Research and Training) in Exobiology at the University of California, San Diego.

References

Alifano, P., Fani, R., Lio, P., Lazcano, A., Bazzicalupo, M., Carlomagno, M. S., and Bruni, C. B. (1996) Histidine biosynthetic pathway and genes:structure, regulation, and evolution, *Microbiol. Rev.* 60, 44-69.

Becerra, A. and Lazcano, A. (1997) The role of gene duplication in the evolution of purine nucleotide salvage pathways, *Origins Life Evol. Biosph.* (in press)

Becerra, A., Islas, S., Leguina, J. I., Silva, E., and Lazcano, A. (1997) Polyphyletic gene losses can bias backtrack characterizations of the cenancestor, *J. Mol. Evol.* 45, 115 118.

Bloch, K. (1994) *Blondes in Venetian Paintings, the Nine Banded Armadillo, and other Essays in Biochemistry*, Yale University Press, New Haven.

Crawford, I. P. (1989) Evolution of a biosynthetic pathway: the tryptophan paradigm, *Annu Rev. Microbiol.* 43, 567-600.

Embley, T. M., Horner, D. A., and Hirt, R. P. (1997) Anaerobic eukaryote evolution: hydrogenosomes as biochemically modified mitochondria? *TREE* 12, 437-441.

Glansdorff, N. (1996) Biosynthesis of arginine and polyamines, in F.C. Neidhardt (ed.), *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, AMS Press, Washington, D C, pp. 408-433.

Horowitz, N. H. (1945) On the evolution of biochemical synthesis, *Proc. Natl. Acad. Sci. USA* 31, 153-157.

Horowitz, N. H. (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, pp. 15-23.

Jensen, R. A. (1976) Enzyme recruitment in the evolution of new function, *Annu. Rev. Microbiol.* 30, 409-425.

Kanehisa, M. (1997) A database for post-genome analysis, *Trends Genet.* 13, 375-376.

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

Mushegian, A. R. and Koonin, E. V. (1996) Gene order is not conserved in bacterial evolution, *TIGS* 12, 289-290.

Neidhart, D. J., Kenyon, G. L., Gertl, J. A., and Petsko, G. A. (1990) Mandelate racemase and muconate lactonizing enzyme are mechanistically different and structurally homologous, *Nature* 347, 692-694.

Olsen, G. J. and Woese, C. R. (1997) Archaeal genomics: an overview, *Cell* 89, 991-994.

Ourisson, G. and Nakatani, Y. (1994) The terpenoid theory of the origin of cellular life: the evolution of terpenoids to cholesterol, *Chemistry & Biology* 1, 11-23.

Peretó, J., Fani, R., Leguina, J. I., and Lazcano, A. (1997) Enzyme evolution and the development of metabolic pathways, in A. Cornish-Bowden (ed.), *Cell-Free Fermentation and the Growth of Biochemistry: Essays in Honour of Eduard Buchner*, Publicacions de la Universitat de Valencia, Valencia, Spain, (in press)

Riley, M. and Labedan, B. (1996) *Escherichia coli* gene products: physiological functions and common ancestries, in F.C. Neidhardt (ed.), *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, AMS Press, Washington, D.C., pp. 2118-2202.

Stetter, K. O. (1996) Hyperthermophiles in the history of life, in G. R. Bock and J. A. Goode (eds.) *Evolution of Hydrothermal Ecosystems on Earth (and Mars?)*, Wiley, Chichester, pp. 1-10.

Vogel, H. J. (1960) Two modes of biosynthesis among lower fungi: evolutionary significance, *Biochem. Biophys. Acta* 41, 172-173.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, *Proc. Natl. Acad. Sci. USA* 87, 4576-4579.

Zubay, G. (1993) To what extent do biochemical pathways mimic prebiotic pathways?, *Chemtracts - Biochem. Mol. Biol.* 4, 317-323.

MOLECULAR BIOLOGY AND THE RECONSTRUCTION OF MICROBIAL PHYLOGENIES: Des Liaisons Dangereuses?

A. BECERRA, E. SILVA, L. LLORET, S. ISLAS,
A. M. VELASCO, and A. LAZCANO

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

1. Introduction

Only half-a-century after the DNA double chain model was first suggested, molecular biology has become one of the most provocative, rapidly developing fields of of scientific research, that has led not only to tantalizing new findings on processes and mechanisms at the molecular level, but also to major conceptual revolutions in life sciences. Is there any hope of developing methodological approaches and theoretical frameworks not only to make sense of the overwhelming growing body of data that this relatively new field is producing, but also to use them to develop a more integrative, truly multidisciplinary understanding of biological phenomena? As Peter Bowler wrote a few years ago, Charles Darwin and his followers were accutely aware that "evolutionism's strength as a theory came fom its ability to make sense out of a vast range of otherwise meaningless facts" (Bowler, 1990). This situation has not changed. Evolutionary biology may be in a state of major turmoil, but its unifying powers have not diminished at all. In fact, they probably represent one of the most promising possibilities of overcoming the perils of reductionism that have plagued molecular biology since its inception.

Molecular approaches to evolutionary issues are a century old. The possibility of developing a sucessful blending between them may have been first suggested by the American-born British biologist and physician George H. F. Nuttall, who in 1904 published a book summarizing the results of the detailed comparison of blood proteins that he had used to reconstruct the evolutionary relationships of animals. "In the absence of palaeontological evidence", wrote Nuttall (1904), "the question of of the interrelation-ship amongst animals is based upon similarities of structure in existing forms. In judging of these similarities, the subjective element may largely enter, in evidence of which we need but look at the history of the classification of the Primates". Such subjective element, Nuttall believed, could be succesfully overcomed by

overcomed by constructing a phylogeny based not on form but on the inmunological reactions of blood-related proteins.

Although the comparative analysis of biochemical properties, metabolic pathways and, in few cases, morphological characteristics, had provided some useful insights on the evolutionary relationships among certain microorganisms, until a few years ago the reconstruction of bacterial phylogenies and the understanding of microbial taxonomy were both viewed with considerable skepticism. This situation has undergone dramatic changes with the recognition that proteins and nucleic acid sequences are historical documents of unsurpassed evolutionary significance (Zuckerkandl and Pauling, 1965), and has led to a radical renovation of the phylogeny, classification, and systematics of prokaryotic and eukaryotic microbes (Woese, 1987).

But these changes have also sparked new debates, and have led to an increased appreciation that the scope and limits of molecular cladistic methodologies require clarification. As shown by the current controversies on the characteristics of the first organisms, the origin of the different components of the eukaryotic cell, and the soundness of traditional taxonomic systems, the development of the full potential of molecular cladistics will depend not only on methodological refinements to improve the algorithms used for reconstructing evolutionary history from molecular data, but also on the critical reexamination of its theoretical framework, which includes a number of central concepts, most of which were grafted from classical evolutionary theory into molecular biology. Here we discuss some of these issues, and review briefly some of the major contributions that they have promoted in our understanding of previously uncharacterized early periods of biological evolution.

2. On the nature of eukaryotic cells

The awareness that genomes are extraordinarily rich historical documents from which a wealth of evolutionary information can be retrieved has widened the range of phylogenetic studies to previously unsuspected heights. The development of efficient nucleic acid sequencing techniques, which now allows the rapid sequencing of complete cellular genomes, combined with the simultaneus and independent blossoming of computer science, has led not only to an explosive growth of databases and new sophisticated tools for their exploitation, but also to the recognition that different macromolecules may be uniquely suited as molecular chronometers in the construction of nearly universal phylogenies.

A major achievement of this approach has been the evolutionary comparison of small subunit ribosomal RNA (rRNA) sequences, which has allowed the construction of a trifurcated, unrooted tree in which all known organisms can be grouped in one of three major (apparently) monophyletic cell lineages: the

eubacteria, the archaebacteria, and the eukaryotic nucleocytoplasm, now referred to as new taxonomic categories, i.e., the domains *Bacteria*, *Archaea*, and *Eucarya*, respectively (Woese et al., 1990). There is strong evidence that the identification of these lineages is not an artifact based solely upon the reductionist extrapolation of information derived from one single molecule. While trees based on whole genome information have confirmed at a broad level rRNA-based phylogenies (Snel et al., 1999; Tekaia et al., 1999), it is also true that the congruence between rRNA genes and other molecules is not always ideal, and anomalous phylogenies have been reported (Rivera and Lake, 1992; Gupta and Golding, 1993). At the time being there is no general explanation to account for these peculiar topologies, and the possibility that we may have to restrict ourselves to empirical characterizations of such cases should be kept in mind. However, a large variety of phylogenetic trees constructed from DNA and RNA polymerases, elongation factors, F-type ATPase subunits, heat-shock and ribosomal proteins, and an increasingly large set of genes encoding enzymes involved in biosynthetic pathways, have confirmed the existence of the three primary cellular lines of evolutionary descent (Doolittle and Brown, 1994), between which extensive horizontal transfer events have taken place (Doolittle, 1999).

The ensuing tripartite taxonomic description of the living world fostered by Woese and his followers has been disputed by a number of workers, who contend that both eubacteria and archaebacteria are *bona fide* prokaryotes, regardless of the pecularities that separate that separate them at the molecular level, both are prokaryotes (Mayr, 1990; Margulis and Guerrero, 1991; Cavalier-Smith, 1992). Furthermore, because of their very nature, molecular dichotomous phylogenetic trees cannot be drawn which include anastomozing branches corresponding to the lineages which gave rise to the different components of eukayotic cells. Accordingly, Margulis and Guerrero (1991) have argued that although molecular cladistics is now a prime force in systematics, phylogenetically accurate taxonomic classifications should be based not only on the evolutionary comparison of macromolecules, but also on metabolic pathways, chromosomal cytology, ultrastructural morphology, biochemical data, life cycles, and, when available, paleontological and geochemical evidence.

While molecular phylogenies have confirmed the endosymbiotic origin of plastids and mitochondria, a number of trees also suggest that a major portion of the eukaryotic nucleocytoplasm originated from an archaebacteria-like cell whose descendants form the monophyletic eucaryal branch (Gogarten-Boekels and Gogarten, 1994). As asserted by Woese and his collaborators, although the presence of endosymbionts is of critical importance to the eukaryotes, it is undeniable that the latter "have a unique, meaningful phylogeny" (Wheelis et al., 1992). While such view assumes an absolute continuity between the nucleocytoplasm and its direct ancestor, the holistic arguments advocated by Margulis and Guerrero (1991), Cavalier-Smith (1992), and others, emphasize the evolutionary emergence of an novel type of cell as a result of endosymbiotic events. According to the latter, the key

transitional event leading to eukaryosis was the evolutionary acquisition of heritable intracellular symbionts, and the eucaryal branch does not represent eukaryotic cells as a whole, any more than fungal hyphae or phycobionts like the *Trebouxia* algal cells exhibit, by themselves, all the phenotypic and genetic characteristics of a lichen thallus.

Of course, antagonistic taxonomies have coexisted more or less peacefully along the history of biology. However, the urgent need to critically revise current classificatory systems cannot be underscored. Modern taxonomic schemes need to acknowledge not only the existence of three major cell lineages, but also the eukaryotic divergence patterns, which appear to be the result of rapid bursts of speciation (Sogin, 1994). Any such modifications in biological classification require the recognition of the functional and anatomical continuity between the eukaryotic cytoplasm and the intranuclear environment, as well as the likelihood that the evolution of membrane-bounded nuclei is indeed a byproduct of permanent intracellular associations. In fact, extant amitochondrial eukaryotes such as *Giardia* and *Trichomonas* appear to have had mitochondria in the past (Germont et al., 1997), and still harbor permanent intracellular bacterial endosymbionts (Margulis, 1993). These amitochondrial cells, which may include the microaerophilic, amitotic, multinucleated giant amoeba *Pelomyxa palustris*, are all located in the lowest branches of the eucarya, and contain several types of intracellular prokaryotes which may be the functional equivalents of mitochondria. The ubiquity of endosymbionts suggests that they may have played a critical role in the evolutionary development of nucleated cells. This hypothesis is amenable to observational and experimental designs, and may be supported by studying the possible bacterial affinities of membrane-bounded hydrogenosomes that are known to multiply by binary division in the *Trychomonas* cytoplasm (Müller, 1988), as well as by searching for prokaryotic endosymbionts in species of Parabasalia, Retortomonads, Diplomonads, Calonymphids, and other protist taxa, some of which may have evolved prior to mitochondrial acquisition.

3. The root of the tree or the tip of the trunk?

The construction of the unrooted rRNA tree showed that no single major branch predates the other two, and all three derive from a common ancestor. It was thus concluded that the latter was a progenote, which was defined as a hypothetical entity in which phenotype and genotype still had an imprecise, rudimentary linkage relationship (Woese and Fox, 1977). According to this view, the differences found among the transcriptional and translational machineries of eubacteria, archaebacteria, and eukaryotes, were the result of evolutionary refinements that took place separately in each of these primary banches of descent after they have diverged from their universal ancestor (Woese, 1987).

From an evolutionary point of view it is reasonable to assume that at some point in time the ancestors of all forms of life must have been less complex than even the simpler extant cells, but our current knowledge of the characteristics shared between the three lines has shown that the conclusion that the last common ancestor was a progenote was premature. This interpretation, based on rRNA-based trees for which no outgroups have been discovered, has been definitively superseded (Woese, 1993). A partial description of the last common ancestor of eubacteria, archaebacteria, and eukaryotes may be inferred from the distribution of homologous traits among its descendants. The set of such genes that have been sequenced and compared is still small, but the sketchy picture that has already emerged suggests that the most recent common ancestor of all extant organisms, or *cenancestor*, as defined by Fitch and Upper (1987), was a rather sophisticated cell with at least (a) DNA polymerases endowed with proof-reading activity; (b) ribosome-mediated translation apparatus with an oligomeric RNA polymerase; (c) membrane-associated ATP production; (d) signalling molecules such as cAMP and insulin-like peptides; (e) RNA processing enzymes; and (f) biosynthetic pathways leading to amino acids, purines, pyrimidines, coenzymes, and other key molecules in metabolism (cf. Lazcano, 1995).

Although the possibility of horizontal transfer should always be kept in mind, the traits listed above are far to numerous and complex to assume that they evolved independently or that they are the result of massive multidirectional horizontal transfer events which took place before the earliest speciation events recorded in each of the three lineages. Their presence suggests that the cenancestor was not a direct, immediate descendant of the RNA world, a protocell or any other pre-life progenitor system. Very likely, it was already a complex organism, much akin to extant bacteria, and must be considered the last of a long line of simpler earlier cells for which no modern equivalent is known.

Unfortunately, the characteristics of evolutionary predecessors of the cenancestor cannot be inferred from the plesiomorphic traits found in the space defined by rRNA sequences. Although trees constructed from such universally shared characters appear to be free of internal inconsistencies, the lack of outgroups leads to topologies that specify branching relationships but not the position of the ancestral phenotype. Thus, such trees cannot be rooted. This phylogenetic *cul-de-sac* may be overcomed by using paralogous genes, which are sequences that diverge not through speciation but after a duplication event. As noted over twenty years ago by Schwartz and Dayhoff (1978), rooted trees can be constructed by using one set of paralogous genes as an outgroup for the other set, a rate-independent cladistic methodology that expands the monophyletic grouping of the sequences under comparison.

This approach was used independently a few years ago by Iwabe et al (1989) and Gogarten et al (1989), who analyzed paralogous genes encoding (a) the two elongation factors (EF-G and EF-Tu) that assist in protein biosynthesis; and (b) the alpha and beta hydrophilic subunits of F-type ATP synthetases. Using different tree-

constructing algorithms, both teams independently placed the root of the universal trees between the eubacteria, on the one side, and archaebacteria and eukaryotes on the other. Their results imply that eubacteria are the oldest recognizable cellular phenotype, and imply that specific phylogenetic affinities exist between the archaea and the eucarya.

This branching order, which was promptly adopted by Woese et al (1990), appears to be consistent with structural and functional similarities which are known to exist in the translation and replication machineries of both archaebacteria and eukaryotes (Ouzonis and Sander, 1992; Kaine et al., 1994). However, the issue is far from solved, and has in fact been further complicated by the availability of completely sequenced genomes. The situation is further aggravated by the fact that the phylogenetic analysis of sets of ancestral paralogous genes other than the elongation factors and the ATPase hydrophilic subunits has challenged the conclusion that universal trees are rooted in the eubacterial branch (cf. Forterre et al., 1993). While the sequences of the products of genes involved in the transcription/transcriptional molecular machinery of eukaryotes appear to be closer to those of the archaea than to the eubacteria, other sequences such as those encoding heat-shock proteins and several enzymes suggest the existence of phylogenetic affinities between archaebacteria and Gram positive bacteria. No support for a particular topology was detected when mean interdomain distance analysis was used to analize a set of approximately forty genes common to the three lineages (Doolittle and Brown, 1994).

The lack of congruency between different universal phylogenies may be the result not only of the statistical problems involved in the aligment and comparison of a large number of sequences that may have diverged more than 3.5×10^9 years ago, but also of even older additional paralogous duplications (Forterre et al., 1993), and of horizontal gene tranfer events (Doolittle, 1999), both of which may be obscuring the natural relationships between the lineages. Given the likelihood that microbial phylogenetic analysis will increase its reliance on paralogous duplicates to define outgroups and character polarities (Sidow and Bowman, 1991), detailed studies should be devoted to assess the validity and limits of this cladistic methodology.

Minor differences in the basic molecular processes of the three main cell lines can be distinguished, but all known organisms, including the oldest ones, share the same essential features of genome replication, gene expression, basic anabolic reactions, and membrane-associated ATPase mediated energy production. The molecular details of these universal processes not only provide direct evidence of the monophyletic origin of all extant forms of life, but also imply that the sets of genes encoding the components of these complex traits were frozen a long time ago, i. e., major changes in them are very strongly selected against and are lethal. Biological evolution prior to the divergence of the three domains was not a continuous, unbroken chain of progressive transformation steadily proceeding towards the

cenancestor. However, no evolutionary intermediate stages or ancient simplified version of the basic biological processes have been discovered in extant organisms.

Nevertheless, clues to the genetic organization and biochemical complexity of the earlier entities from which the cenancestor evolved may be derived from the analysis of paralogous sequences. Their presence in the three cell lineages implies not only that their last common ancestor was a complex cell already endowed, among others, with pairs of homologous genes encoding two elongation factors, two ATPase hydrophilic subunits, two sets of glutamate dehydrogenases, and the A and B DNA polymerases, but also that the cenancestor itself must have been preceded by simpler cells in which only one copy of each of these genes existed. In other words, Archean paralogous genes provide evidence of the existence of ancient organisms in which ATPases lacked the regulatory properties of its alpha subunit, protein synthesis took place with only one elongation factor, and the enzymatic machinery involved in the replication and repair of DNA genomes had only one polymerase ancestral to the *E. coli* DNA polymerase I and II.

By definition, the node located at the bottom of the cladogram is the root of a phylogenetic tree, and corresponds to the common ancestor of the group under study. But names may be misleading. The recognition that basic biological processes like DNA replication, protein biosynthesis, and ATP production require today the products of pairs of genes which arose by paralogous duplications during the early Archean, implies that what we have been calling the root of universal trees is in fact the tip of a trunk of unknown length in which the history of a long (but not necessarily slow) series of archaic evolutionary events may still be recorded. The inventory of paralogous genes that duplicated during this previously unchacterized stage of biological evolution appears to include, in addition to elongation factors, ATPase subunits, and DNA polymerases, the sequences encoding heat shock proteins, ferredoxins, dehydrogenases, DNA topoisomerases, several pairs of aminoacyl-tRNA synthetases, and enzymes involved in nitrogen metabolism and amino acid biosynthesis. It is noteworthy that this list includes also aspartate transcarbamoyl transferase, an enzyme which together with carbamyl phosphate synthetase (whose large subunit is itself the product of an internal, i.e., partial, paralogous duplication) catalyzes the initial steps of pyrimidine biosynthesis (García-Meza et al, 1995).

Thus, prior to the early duplication events that led to what may be a rather large number of cenancestral paralogous sequences, simpler living systems existed which lacked the large sets of enzymes and the sophisticated regulatory abilites of contemporary cells. Although lateral transfer of coding sequences may be almost as old as life itself, gene duplication followed by divergence probably played a dominant tole in the accretion of complex genomes, and may have led to a rapid rate of microbial evolution. If its is assumed that the rate of gene duplicative expansion of ancient cells was comparable to today's present values, which are of 10^{-5} to 10^{-3}

gene duplications per gene per cell generation (Stark and Wahl, 1984), the maximum time required to go from an hypothetical 100-gene organism to one endowed with a filamentous cyanobacterial-like genome of approximately 7000 genes would be less than ten million years (Lazcano and Miller, 1994).

Although there are no published data on the rate of formation of new enzymatic activities resulting from gene duplication events under either neutral or positive selection conditions, the role of duplicates in the generation of evolutionary novelties is well stablished. Once a gene duplicates, one of the copies may be free to accumulate non-lethal mutations and acquire new additional properties, which could lead into its specialization or recruitment into new role. Data summarized here supports the idea that primitive biosynthetic pathways were mediated by small, inefficient enzymes of broad substrate specificity (Jensen, 1976). Larger substrate ranges may had not been a disadvantage, since relatively unspecific enzymes may have helped ancestral cells with reduced genomes overcome their limited coding abilities (Ycas, 1974).

The discovery that homologous enzymes catalyzing similar biochemical reactions are part of different anabolic pathways supports the idea that enzyme recruitment took place during the early development of several basic anabolic pathways. Evolutionary tinkering of the products of duplication events apparently had a major role in metabolic evolution. This is supported by the analysis of complete genome sequences, that has shown the large proportion of gene content that is the outcome of duplication events (Tekaia and Dujon, 1999). Such high levels of redundancy represent an illuminating possibility and suggest that the wealth of phylogenetic information older than the cenancestor may be larger than realized, and its analysis may provide fresh insights into a crucial but largely undefined stage of early biological evolution during which major biosynthetic pathways emerged and became fixed.

There is a major exception to the above conclusion. True fungi, euglenids, and chrytridiomycetes synthesize lysine via an eight-step pathway in which α-aminoadipate (AAA) is an intermediate. This route is different from the seven-step diaminopimelate pathway used by bacteria, plants, and most protist (Bhattacharjee, 1985). The phylogenetic distribution of these two pathways suggest that the AAA route is the most recent one. Accordingly, if the patchwork assembly of metabolic pathways (Jensen, 1976) is valid, then it can be predicted that the enzymes catalizing the AAA-route should be homologous to those participating in other major biosynthetic routes.

The recognition that enzyme recruitment may have played a major role in metabolic evolution leads, however, to assume some caution in phylogenetic inferences. Although in some cases metabolic pathways may be sucessfully used to assess the phylogenetic relationship of prokaryotes (DeLey, 1968; Margulis, 1993),

the possibility that some of the enzymes of archaic pathways may have survived in unusual organisms (Keefe et al., 1994), or that important portions of extant metabolic routes may have been assembled by a patchwork process (Jensen, 1976), suggest that considerable prudence should be exerted when attempting to describe the physiology of truly primordial organisms by simple direct back extrapolation of extant metabolism.

4. Molecular cladistics and the origin of life: is there any connection?

"All the organic beings which have ever lived on this Earth", wrote Charles Darwin in the *Origin of Species*, "may be descended from some primordial form". Although the placement of the root of universal trees is a matter of debate, the development of molecular cladistics has shown that despite their overwhelming diversity and tremendous differences, all organisms are ultimately related and descend from Darwin's primordial ancestor. But what was the nature of this progenitor?

The heterotrophic hypothesis suggested by Oparin (1938) not only gave birth to a whole new field devoted to the study of the origin of life, but played a central role in shaping several influential taxonomic schemes and different bacterial phylogenies (Margulis 1993). Although the central role of glycolysis and the wide phylogenetic distribution of at least some of its molecular components are strong indications of its antiquity (Fothergill-Gilmore and Michels, 1993), it is no longer possible to support the *ad hoc* identification of putative primordial traits to assume that the first living system was a *Clostridium*-like anaerobic fermenter or a *Mycoplasma* type of cell (cf. Lazcano et al., 1992). Like vegetation in a mangrove, the roots of universal phylogenetic trees are sumerged in the muddy waters of the prebiotic broth, but how the transition from the non-living to the living took place is still unknown.

Indeed, we are still very far from understanding the origin and attributes of the first living beings, which may have lacked even the most familiar features in extant cells. For instance, protein synthesis is such an essential characteristic of cells, that it is frequently argued that its origin should be considered synonymous with the emergence of life itself. However, the discovery of the catalytic activities of RNA molecules has led considerable support to the possibility that during early stages of biological evolution living systems were endowed with a primitive replicating and catalytic apparatus devoid of both DNA and proteins The scheme may be even more complex, since RNA itself may have been preceded by simpler genetic macromolecules lacking not only the familiar 3',5' phosphodiester backbones of nucleic acids, but perhaps even today's bases (Lazcano and Miller, 1996).

Although molecular cladistics may provide clues to some late steps in the development of the genetic code, it is difficult to see how the applicability of this approach can be extended beyond a threshold that corresponds to a period of cellular

evolution in which protein biosynthesis was already in operation. Older stages are not yet amenable to molecular phylogenetic analysis. Although there have been considerable advances in the understanding of chemical processes that may have taken place before the emergence of the first living systems, life's beginnings are still shrouded in mystery. A cladistic approach to this problem is not feasible, since all possible intermediates that may have once existed have long since vanished. The temptation to do otherwise is best resisted. Given the huge gap existing in current descriptions of the evolutionary transition between the prebiotic synthesis of biochemical compounds and the cenancestor (Lazcano, 1994), it is naive to attempt to describe the origin of life and the nature of the first living systems from the available rooted phylogenetic trees.

Nevertheless, there have been several recent attempts to use macromolecular data to support claims on the hyperthermophily of the first living organisms and the idea of a hot origin of life. The examination of the prokaryotic branches of unrooted rRNA trees had already suggested that the ancestors of both eubacteria and archaebacteria were extreme thermophiles, i.e., organisms that grow optimally at temperatures in the range 90° C and above (Achenbach-Richter et al., 1987). Rooted universal phylogenies appear to confirm this possibility, since heat-loving bacteria occupy short branches in the basal portion of molecular cladograms (Stetter, 1994).

Such correlation between hyperthermophily and primitiveness has led support to the idea that heat-loving lifestyles are relics from early Archean high-temperature regimes that may have resulted from a severe impact regime (Sleep et al., 1989). It has also been interpreted as evidence of a high temperature origin of life, which according to these hypotheses took place in extreme environments such as those found today in deep-sea vents (Holm, 1992) or in other sites in which mineral surfaces may have fueled the appearance of primordial chemoautolithotrophic biological systems (Wächtershäuser, 1990).

Such ideas are not totally without precedent. The possibility that the first heterotrophs may have evolved in a sizzling-hot environment is in fact an old suggestion (Harvey, 1924). Despite their long genealogy, these hypotheses have not been able to bypass the problem of the chemical decomposition faced by amino acids, RNA, and other thermolabile molecules which have very short lifetimes under such extreme conditions (Miller and Bada, 1988). Although no mesophilic organisms older than heat-loving bacteria have been discovered, it is possible that hyperthermophily is a secondary adaptation that evolved in early geological times (Sleep et al., 1989; Confalonieri et al., 1993; Lazcano, 1993). Such possibility is in fact strongly supported by the recent phylogenetic analysis of the G+C content of rRNA genes, which suggest that the last common ancestor was not a hyperthermophilic organism (Galtier et al., 1999).

In fact, hyperthermophiles not only share the same basic features of the molecular machinery of all other forms of life; they also require a number of specific biochemical adaptations. Any theory on the hot origin of life must address the question of how such traits, or their evolutionary precedessors, arose spontaneously in the prebiotic environment. Such adaptations may include histone-like proteins, RNA modificating enzymes, and reverse gyrase, a peculiar ATP-dependent enzyme that twists DNA into a positive supercoiled conformation (Confalonieri et al., 1993). Clues to the origin of hyperthermophily may be hidden in this list, and its evolutionary analysis may contribute to the understanding of the rather surprising phylogenetic distribution of the immediate mesophilic descendants of heat-loving prokaryotes, which shows that at least five independent abandonments events of hyperthermophilic traits took place in widely separated branches of universal trees, one of which corresponds to the eukaryotic nucleocytoplasm (García-Meza et al., 1995).

The antiquity of hyperthermophiles appears to be well established, but there is no evidence that they have a primitive molecular genetic apparatus. Thus, the most basic questions pertaining to the origin of life relate to much simpler replicating entities predating by a long series of evolutionary events the oldest recognizable heat-loving bacteria. Why hyperthermophiles are located at the base of universal trees is still an open question, but the possibility that adaptation to extreme environments is part of the evolutionary innovations that appeared in trunk of the tree cannot be entirely dismissed. The phylogenetic distribution of heat-loving bacteria is no evidence by itself of a hot origin of life, any more than the presence in the hyperthermophile archaeon *Sulfolobus solfataricus* of a gene encoding a thermostable B-type DNA polymerase endowed with 3'-5' exonuclease activity (Pisani et al., 1992) can be interpreted to imply that the first living organism had a DNA genome.

5. Final remarks

Although in the past few years the relationship between molecular biology and microbial phylogenetics has been embittered by frequent clashes and antagonism, the development of rapidly growing sequence databanks has provided a unique view of the evolution of bacterial and eukaryotic microorganisms, and has opened new perspectives in several major fields of life sciences. Molecular evolution was originally the outcome of the wedding of molecular biology with neodarwinian theory, but it has been rapidly transformed into a field of scientific enquiry in its own right. However, its full development requires not only the development of less-expensive, more rapid macromolecular sequencing techniques and more powerful computer algorithms for constructing phylogenetic trees, but also the awareness of its non-stated assumptions and more precise definitions of its conceptual framework.

As summarized by Patterson (1988), the theoretical foundations of molecular cladistics have been based on a number of central concepts, most of which were inherited from older disciplines, such as physiology, anatomy, and neodarwinism. Homology, which is one of the key concepts in evolutionary theory, was originally used by Wolfgang Goethe, Ettiene Geoffroy Saint-Hilaire, Richard Owen, and others, to describe structural resemblance to an archetype (Donoghue, 1992). In recent years it has not only been repeatedly confused with sequence similarity (Reeck et al., 1988), but is also used to describe a wider range of possible evolutionary relationships that include species- or gene-phylogeny. In fact, some classes of homology that describe phenomena at the molecular genetic level may have no exact equivalent in orthodox evolutionary analysis of morphological traits. One such case is paralogy, a term coined by Fitch (1970) to describe the diversification of genes following duplication events.

Since paralogy provides evidence of gene duplication but not of speciation events, it is the basis for infering evolutionary relationships among genes, not among species. Recognition og this distinction has led to repeated recommendations on the avoidance of paralogous sequences in phylogenetic analysis. However, the use of paralogous duplicates in outgroup analyses for determining the evolutionary polarity of character states in universal phylogenies (Gogarten et al., 1989; Iwabe et al., 1989) has rekindled keen theoretical interest in their advantageous properties. Their use, however, does pose some risks. The naive assumption that only one paralogous duplication has taken place in the set of sequences under consideratiuon may lead to incorrect topologies (Forterre et al., 1993). Indeed, the incorporation of genes that are the result of unrecognized multiple paralogous events in a tree may be even more insidious than the problem derived by convergent evolution and lateral gene transfer. The latter phenomena are much more easily identified at the molecular level.

The recognition that paralogous duplicates expand a monophyletic group of sequences raises a number of issues not encountered in classical evolutionary analysis. From a (classical) cladistic point of view, a character that is found only in outgroups is primitive. Nonetheless, in molecular phylogenetic analysis this may not be always the case. Such rule would hold if multiple paralogous duplications have taken place, and if one (or several) of the older sequences is used as an outgroup for an unrooted tree of younger sequences. This would be the case, for instance, if a myoglobin sequence is used to root alpha (or beta) haemoglobin trees However, this rule would not hold if an alpha haemoglobin sequence (or a set of them) is used as an outgroup for the beta haemoglobin tree, or viceversa.

The same is true, of course, with universal phylogenetic trees derived from elongation factors (Iwabe et al, 1989). In this case neither set is older than its homologue. In this case, the reconstruction of ancestral character states from dichotomously varying paralogous genes does not comes from the analysis of the outgroup, but may be inferred from the realization that the root of the tree must have

been preceded by an even older, more primitive condition in which only one copy of the gene existed, prior to the paralogous duplication. Recognition of this fact is likely to play a central role in future understanding of enzyme evolution during the early Archean. Although it is true that the raw material for molecular cladistic analysis is restricted to sequences derived from living organisms (or from fossil samples from which ancient preserved DNA can be retrieved) and cannot be applied to extinct groups of organisms, the construction of trees derived from archaic paralogous sequences may allow us to infer evolution prior to the ealiest detectable nodes.

The flourishing of molecular techniques has led into a proliferation not only of sequences of isolated molecular constituents of living organisms, but also of completely sequenced genomes. This is a storehouse of data that has already provided considerable insights into the phylogeny and the diversity of microbes. But because of its very nature, molecular cladistics separates clusters of adaptative characters into a nested hierarchical set which is expected to reflect the temporal sequence of their evolutionary acquisition. However fruitful, such approach has all the demerits of a reductionist one-trait approach to biological evolution chastised in early literature as "partial phylogeny", and since the birth of molecular phylogeny has rarely been used to attempt a truly integrative analysis of complete character complexes.

Such limitation may be overcomed in several ways, some of which are part of intellectual traditions deeply rooted in comparative biology. As Georges Cuvier contended in his 1805 *Lectures in Comparative Anatomy*, the appearance of the whole skeleton can be deduced up to a certain point by examination of a single bone. The success that Cuvier had in such anatomical reconstructions is legendary, and was based not only in his unsurpassed knowledge and intuition, but also on what he termed the "correlation of parts", i. e., the full recognition of a functional coordination of the parts of the body of a given animal (Young, 1992). Such correlation of parts is not restricted to bones and muscles; at subcellular levels, it underlies the functional coordination among the molecular components of multigenic traits such as metabolic pathways and protein biosynthesis. As shown by the intimate relationship between the biosyntheses of valine and isoleucine, their triplet assignments, and the phylogenetic proximity of their aminoacyl-tRNA synthetases, inquiries on the early evolution of the genetic code and other basic features of living systems should be understood not only by determining the molecular phylogenies of some of their isolated components or by mathetical discussions spiced with a distinct Pythagorean flavor, but with the integrative analysis of character complexes.

But for all its foibles, the relationship between molecular biology and evolutionary theory has opened new, unsuspected avenues of intellectual exploration. Never before has such a wealth of methodological approaches and empirical data been available to the students of life's phenomena. In part because of this prosperity, systematics and evolutionary biology, two of the most broadly oriented fields of life

sciences, are now in a state of intellectual agitation. The symptoms are manifold: it is possible that the traditional species concept may not apply to prokaryotes, time-cherished concepts like that of the existence of kingdoms are under fire, the origin and taxonomic position of genetic mobile elements is unknown. There is an increased awareness that the understanding of the processes underlying the generation of evolutionary novelties and the origin of ontogenic patterns cannot be restricted by classical neodarwinian explanations. We are living in the midst of hectic times in which epoch-making debates are reshaping the future of the life sciences, and the development of a more integrated molecular biology may be a never-ending story. It is said that to wish someone to live in an interesting time is one of the most terrible of all Chinese curses. Whatever the outcome of current discussions and debates, for biology the putative Oriental curse may turn out to be nothing less than an intellectual blessing.

Acknowledgments

We are indebted to Dr. Lynn Margulis for her critical reading of the manuscript and many suggestions. Support from the UNAM-DGAPA Project PAPIIT-IN213598 is gratefully acknowledged.

6. References

Achenbach-Richter, L., Gupta, R., Stetter., K. O., and Woese, C. R. (1987) Were the original eubacteria thermophiles? *System. Appl. Microbiol* 9, 34-39

Bhattacharjee, J. K. (1985) α-aminoadipate pathway for the biosynthesis of lysine in lower eukaryotes, *CRC Crit. Rev. Microbiol.* 12, 131-151

Bowler, P. J. (1990) *Charles Darwin, The man and his influence,* Basil Blackwell, Oxford

Confalonieri, F., Elie, C., Nadal, M., Bouthier de la Tour, C., Forterre, P., and 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

DeLey, J. (1968) Molecular biology and bacterial phylogeny, in T. Dobzhansky,. K. Hecht, and W. C. Steere (eds), *Evolutionary Biology,* Appleton-Century-Crofts, New York, pp. 104-156

Doolittle, W. F. (1999) Phylogenetic classification and the universal tree, *Science* 284, 2124-2128

Doolittle, W. F. and Brown, J. R. (1994) Tempo, mode, the progenote and the universal root, *Proc. Natl. Acad. Sci. USA* 91, 6721-6728

Donoghue, M. J. (1992) Homology, in E. Fox Keller and E. A. Lloyd (eds), *Keywords in Evolutionary Biology,* Harvard University Press, Cambridge, pp. 170-179

Fitch W. M. (1970) Distinguishing homologous from analogous proteins, *Syst. Zool.* 19, 99-113

Fitch, W. M. and 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

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

Galtier., N., Tourasse, N., and Gouy, M. (1999) A nonhyperthermophilic common ancestor to extant life forms, *Science* **283**, 220-221

Garcia-Meza, V., Gonzáiez-Rodriguez, A., and Lazcano, A. (1995) Ancient paralogous duplications and the search for Archean cells, in G. R. Fleischaker, S. Colonna, and P. L. Luisi (eds), *Self-Reproduction of Supramolecular Structures, from synthetic structures to models of minimalliving systems*, Klüwer, Amsterdam, pp. 231-246

Germont, A., Phillipe, H., and Le Guyader, H. (1997) Evidence for the loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in *Nosema locustae*, *Mol. Biochem. Parasitol.* **8**, 159-168

Gogarten-Boekels, M. and Gogarten, J. P. (1994) The effects of heavy meteorite bombardment on the early evolution of life --a new look at the molecular record, *Origins of Life and Evol. Biosph.* **25**, 78-83

Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. L., Poole, J., Date, T., Oshima, Konishi, L., Denda, K., and Yoshida, M. (1989) Evolution of the vacuolar H^+-ATPase, implications for the origin of eukayotes, *Proc. Natl. Acad. Sci. USA* **86**, 6661-6665

Gupta, R. S. and Golding, G. B. (1993) Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes, *J. Mol. Evol.* **37**, 573-582

Harvey, R. B. (1924) Enzymes of thermal algae, *Science* **60**, 481-482

Holm, N. G., ed., (1992) *Marine Hydrothermal Systems and the Origin of Life*, Klüwer Academic Publ., Dordrecht

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T. (1989) Evolutionary relationship of archaebacteria, 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 the evolution of new function, *Ann. Rev. Microbiol.* **30**, 409-425

Kaine, B. P., Mehr, I. J., and Woese, C. R. (1994) The sequence, and its evolutionary implications, of a *Thermococcus celer* protein associated with transcription, *Proc. Natl. Acad. Sci. USA* **91**, 3854-3856

Kandler, O. (1994) The early diversification of life, in S. Bengtson (ed), *Early Life on Earth, Nobel Symposium No. 84*, Columbia University Press, New York, pp. 124-131

Keefe, A. D., Lazcano, A. and Miller, S. L. (1994) Evolution of the biosynthesis of the branched-chain amino acids, *Origins of Life and Evol. Biosph.* **25**, 99-110

Lazcano, A. (1993) Biogenesis, some like it very hot, *Science* **260**, 1154-1155

Lazcano, A. (1994) 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. (1995) Cellular evolution during the early Archaean: what happended between the progenote and the cenancstor? *Microbiologia SEM* **11**, 1-13

Lazcano, A., Fox, G. E., and Oró, J. (1992) Life before DNA, the origin and evolution of early Archean cells, in P. P. Mortlock (ed), *The Evolution of Metabolic Function*, CRC Press, Boca Raton, pp. 237-295

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

Lazcano, A. and Miller, S. L. (1996) The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time, *Cell* 85, 793-798

Margulis, L. (1993) *Symbiosis in Cell Evolution*, W. H. Freeman, New York

Margulis, L. and Guerrero, R. (1991) Kingdoms in turmoil, *New Scientist* 132, 46-50

Mayr, E. (1990) A natural system of organisms, *Nature* 348, 491

Miller, S. L. and Bada, J. L. (1988) Submarine hot springs and the origin of life, *Nature* 334, 609-611

Müller, M. (1988) Energy metabolism of protozoa without mitochondria, *Ann. Rev. Microbiol.* 42, 465-488

Nuttall, G. H. F. (1904) *Blood Immunity and Blood Relationship: a demonstration of certain blood-relationships amongst animals by means of the precipitation test for blood,* Cambridge University Press, Cambridge

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

Ouzonis, C. and Sander, C. (1992) TFIIB, an evolutionary link between the transcription machineries of archaebacteria and eukaryotes.,*Cell* 71, 189-190

Patterson, C. (1988) Homology in classical and molecular biology, *Mol. Biol. Evol.* 5, 603-625

Pisani, F.M., De Martino, C., and Rossi, M. (1992) A DNA polymerase from the archaeon *Sulfolobus solfataricus* shows sequence similarity to family B DNA polymerases. *Nucleic Acid Res.* 20, 2711-2716

Reeck, G. R., de Häen, C., Teller, D. C., Doolittle, R. F., Fitch, W., Dickerson, R. E., Chambon, P., McLachlan, A. D., Margoliash, E., Jukes, T. H., and Zuckerkandl, E. (1987) "Homology" in proteins and nucleic acids, a terminology muddle and a way out of it, *Cell* 50, 667

Rivera, M. C. and Lake, J. A. (1992) Evidence that eukaryotes and eocyte prokaryotes are inmediate relatives, *Science* 257, 74-76

Schwartz, M. and Dayhoff, M. O. (1978) Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts, *Science* 199, 395-403

Sidow, A. and Bowman, B. H. (1991) Molecular phylogeny, *Current Opinion Genet. Develop.* 1, 451-456

Sleep, N. H., Zahnle, K. J., Kastings, J. F., and Morowitz, H. J. (1989) Annihilation of ecosystems by large asteroid impacts on the early Earth, *Nature* 342, 139-142

Snel, B., Bork, P., and Huynen, M. A. (1999) Genome phylogeny based on gene content, *Nature Genetics* 21, 108-110

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

Stark, G. R,. and Wahl, G. M. (1984) Gene amplification, *Ann. Rev. Biochem.* 53, 447-491

Stetter, K. O. (1994) The lesson of archaebacteria, in S. Bengtson (ed), *Early Life on Earth, Nobel Symposium No. 84,* Columbia University Press, New York, pp. 114-122

Tekaia, F. and Dujon, B. (1999) Pervasiveness of gene conservation and persistence of duplicates in cellular genomes, *J. Mol. Evol.* 49, 591-600

Tekaia, F., Lazcano, A., and Dujon, B. (1999) The genomic tree as revealed from whole proteome comparisons, *Genome Research* 9, 550-557

Wächtershäuser, G. (1990) The case for the chemoautotrophic origins of life in an iron-sulfur world, *Origins of Life Evol. Biosph.* 20, 173-182

Wallace, D. C. and Morowitz, N. H. (1973) Genome size and evolution, *Chromosoma* 40, 121-126

Wheelis, M. L., Kandler, O., and Woese, C. R. (1992) On the nature of global classification, *Proc. Natl. Acad. Sci. USA* 89, 2930-2934

Woese, C. R. (1987) Bacterial evolution, *Microbiol. Reviews* 51, 221-271

Woese, C. R. (1993) The archaea, their history and significance, in M. Kates, D. J. Kushner, and A. T. Matheson (eds), *The Biochemistry of the Archaea (Archaebacteria)*, Elsevier Science Publishers, Amsterdam, pp. vii-xxix

Woese, C. R. and Fox, G. E. (1977) The concept of cellular evolution, *Jour. Mol. Evol.* 10, 1-6

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990) Towards a natural system of organisms, proposal for the domains Archaea, Bacteria, and Eucarya, *Proc. Natl. Acad. Sci. USA* 87, 4576-4579

Yeas, M. (1974) On the earlier states of the biochemical system, *J. Theor. Biol.* 44, 145-160

Young, D. (1992) *The Discovery of Evolution*, Natural History Museum Publications, Cambridge

Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history, *J.Theoret. Biol.* 8, 357-366

Phylogenetic distribution of simple sequences: insights from comparative genomics

Arturo Becerra and Antonio Lazcano*

Facultad de Ciencias, UNAM
Apdo. Postal 70-407
Cd. Universitaria
04510 México D.F., MEXICO
* E-mail: alar@hp.fciencias.unam.mx

Abstract

Simple sequences are regions of protein and nucleic acid sequences which are biased in residue composition, and which typically contain repetitive segments. Homopolymeric tracts and tandem arrays of multiple short repeat motifs have their origin in slipped-strand mispairing during DNA replication, and due to their hypermutable character are recognized as a major source of random phenotypic variation, especially among prokaryotic pathogens (Moxon 1999). In order to obtain insights on the phylogenetic distribution of simple sequences and their biological properties, we have used the SEG program (Wootton and Federhen, 1993), with a highly stringent level of granularity, to analyze 23 completely sequenced cellular genomes available in public databases as of March, 2000. Our results indicate that simple sequences (a) have a wide phylogenetic distribution, i.e., their appearance is older than microbial pathogens, but may not be monophyletic; (b) are not restricted to a unique class of enzymes, but are present in catalytic and structural proteins involved in a wide spectrum biological functions; (c) tend to present at both at the carboxy- and amino- ends of proteins; (d) there is a compositional bias in simple sequences, which tend to be rich in alanine, leucine, lysine, serine and glutamine, while histidine, tryptophan, and cisteine are underrepresented; and (e) homopolymeric tracts in hyperthermophyles are enriched in glutamine. The role of simple sequences in protein evolution is discussed.

Keywords: Simple sequences, low complexity composition, proteome, protein evolution.

I. Introduction

Simple sequences are regions of proteins and/or nucleic acids which are biased in residue composition, and which typically contain repetitive segments. Because of the complex nomenclature used to describe such segments can be, in this work simple sequences are defined solely to homopolymeric tracts and tandem arrays of multiple short repeat motifs. The high frequency of simple sequences has been recognized since long ago in eukaryotic genomes and gene databases (Britten and Kohne 1968; Tautz and Renz 1984; Wootton and Federhen 1993), they are known to be also a prokaryotic trait, found both in Bacteria and Archaea (Hancock 1996; Saunders et al. 1998). Homopolymeric tracts and tandem arrays of multiple short repeat motifs have their origin in mutational processes such as slipped-strand mispairing and unequal crossing-over that take place during DNA replication (Bebenek and Kunkel 1990; Richards and Sutherland 1994; Hancock 1995; Epplen and Riess 1997). It has been argued that the occurrence of simple sequences in various types of coding and non-coding sequences plays a role as a source of genetic variation and in the evolution of genome sizes (Tautz et al. 1986; Hancock 1995).

Due to their hypermutable character, simple sequences are recognized as a major source of random phenotypic variation (Hancock 1996), especially among prokaryotic pathogens (Moxon et al. 1994; Moxon 1999). However, their ample phylogenetic distribution opens the posibility that may play a role in gene regulation, gene conversion, dosage compensation, recombination and sex determination (Tautz and Schlötterer 1994). Here we discuss the role that simple sequences may have played together with gene amplification,

horizontal transfer and symbiosis, in the expansion of the encoding abilities of early genomes, and in the emergence of new functions which are dependent of repetitive segments in protein.

The recent availability of complete cellular genomes from the three major cell lineages (Bacteria, Archaea, and Eucarya), opens up the possibility of analyzing the phylogenetic distribution of simple sequences and the role they have play in genome evolution. Here we report the outcome such analysis, and discuss the possible role of simple sequences in the early evolution of coding sequences and their antiquity of the phenomenon.

Material and Methods

Proteomes from the following organisms: *Escherichia coli* (Blattner et. al., 1997), *Haemophilus influenzae* (Fleischmann et. al. 1995), *Helicobacter pylori* 26695 (Tomb et. al. 1997), *Helicobacter pylori* J99 (Alm et. al. 1999), *Rickettsia prowazekii* (Anderson 1998), *Bacillus subtilis* (Kunst et al. 1997), *Mycoplasma genitalium* (Fraser, et al. 1996), *M. pneumoniae* (Himmelreich et. al. 1996), *Mycobacterium tuberculosis* (Cole et al.1998), *Chlamydia trachomatis* (Stephens et al. 1998), *Campylobacter jejuni* (Parkhill et. al. 2000), *Borrelia burgdorferi* (Fraser et al. 1997), *Treponema pallidum* (Fraser et al. 1998), *Aquifex aeolicus* (Deckert et al. 1998), *Synechocystis sp* PCC6803. (Kaneko et al. 1996), *Deinococcus radiodurans* (White et. al. 1999), *Thermotoga maritima* (Nelson et. al. 1999), *Methanococcus jannaschii* (Bult et al. 1996), *Methanobacterium thermoautotrophicum* (Smith et al. 1997), *Archaeoglobus fulgidus* (Klenk et al. 1997), *Pyrococcus horikoshii* (Kawarabayasi et al.1998), *Saccharomyces cerevisiae* (Goffeau et. al. 1997), and

Caenorhabditis elegans (in www.sanger.ac.uk/Projects/C_elegans/), were obtained from the KEGG Encyclopedia via the ftp server of the Chemical Institute of Kyoto in the web site http://www.genome.ad.jp/kegg/kegg2.html.

The analysis of complexity of the sequences obtained was performed using the SEG program (*segment sequences by local complexity program;* Wootton and Federhen 1993), with 12 2.0 2.2 granularity level, where the trigger window length was W=12; trigger complexity K1=2.0 (K1= local compositional complexity, Wootton and Federhen 1993; Wootton 1994), and extension complexity K2=2.2. The amino acid composition of the regions with low complexity was calculated using the *aacomp* program from the FASTA package (Pearson and Lipman 1988).

Low complexity sequences detected were grouped using the functional distribution suggested by KEGG (Kanehisa et al. 1996; Goto et al. 1996). The categories proposed by KEGG are: I) carbohydrate metabolism; II) energy metabolism; III) lipid metabolism; IV) nucleotide metabolism; V) amino acid metabolism; VI) metabolism of other amino acids; VII) metabolism of complex carbohydrates; VIII) metabolism of complex lipids; IX) metabolism of cofactors, vitamins, and other substances; X) metabolism of macromolecules; XI) membrane transport; XII) signal transduction; XIII) molecular assembly; and XIV) unassigned.

The distribution of low complexity segments presents in homologous ORFs were estimated by eye. The programs were run on a SUN Ultra 5 computer (Solaris 2.6).

Results and Discussion

3.1 *Presence of simple sequences in proteomes and their amino acid composition*

The percentage of proteomic simple sequences in all organisms studied here is shown in Table 1. Although the level of simple sequences in non-coding regions of eukaryotic genomes is high in this work are observed that the coding regions possesses a significant percentage of simple sequences. However, it is clear that this phenomenon is present in prokaryotic, found both in Bacteria and Archaea , highlighting the genomes of *Mycobacterium tuberculosis*, and *Deinococcus radiodurans*. But in general it could be observed that a important numbers of coding sequences in the three domains (Eukarya, Bacteria and Archaea), it contains sections of low complexity. Is important to remark the high percentage of coding sequences that have been increase by the insertion of low complexity segment, produced by slipped-strand mispairing or unequal crossing-over, during DNA replication process. This results, together with their wide phylogenetic distribution of the phenomenon suggest that the simple sequences have played a role in the evolution of coding sequences at least in amplification process and source of variability. The low complexity segments are found among of the proteins, however this phenomenon is mostly at the C- and N- ends of proteins (Figure 1) and between domains , the sequences that have segments in the middle, are main represented by hypothetical proteins, suggesting that this kind of mutations could produces structural changes when is produced in the middle of the protein and is selected against.

As seen in Figure 2, simple sequences exhibit important compositional biases. They tend to be rich in alanine, serine, isoleucine leucine, lysine, glutamic acid, while aspartic acid, metionine, histidine, arginine, tyrosine, tryptophan, and cysteine are clearly underrepresented. As shown in Figure 3, is obvious that the compositional biases of high mass amino acid (M, H, F, R, Y, W), are selected against, together with cysteine, that produce disulfide bonds, proline tend to adopt *cis* conformation in peptide units and reduce the stability that is produced by *trans*-form (Branden and Tooze 1999), and aspartic acid, that produce less stable peptide bonds. In contrast, the compositional biases present in simple sequences tend to be rich in small, hydrophobic (except K, that is a ionizable residue), and formed of α-helix amino acids (A, L, I, K), that are mainly present in all organism, serine (hydrophilic), in eucariontes studied here (*Saccharomyces cerevisiae* and *Caenorhabditis elegans*), and glutamic acid, that is the high formed of α-helix and tend to produces strong peptide bonds, is mainly present in simple sequences from eucariontes and hyperthermophyles.

3.2 *The functional distribution of simples sequences*

Simple sequences appear to be present in all kinds of functional classes of proteins, suggesting that their role is not restricted to one type of function. However, they are clearly over-represented in membrane- and signal transport proteins (Figure 4).The non globular nature of many simple sequences, can be related with the high tendency to be present in membrane related function, but also because this phenomenon is related in surface-exposed

proteins in phatogenic processes. Important interactions and functions have been attributed to low complexity segments, that includes DNA and RNA binding, interactions in transcriptional regulation, signal transduction, control of protein folding and turnover, cellular and extracellular mechanics, and tumorigenesis (Wootton 1994b). Joined to this activities of simple sequences, their presence in proteins which are recognized as some of the oldest enzymes, such as those involved in translation; RNA polymerase sigma 54, 70 factors, translation initiation factor IF-2, IF-3, ribosomal proteins (L7,L9, L10, L12, L19, L27a), replication; DNA polymerase III, and others like ATP synthase (F0, F1 subunits), and many ABC transporters, shows that they are a very ancient phenomenon, and may have already been present in the last common ancestor, thus their appearance is older than microbial pathogens. Furthermore, the presence of low complexity segments in ribosomal protein L7/12, that their basic nature tend to unite to RNA, in large and small subunit of ribosomal RNAs (Tautz 1989, Hancock 1995), and in development genes like *hunchback*, *mastermind* and *per* (Treier et al.,1989; Newfeld et al., 1994; Peixoto et al., 1992), suggest that the phenomenon of simple sequences could have given origin to functional domains.

Conclusions

The ample phylogenetic distribution and high frequency of simple sequences found in proteomes, reflects that mutational processes such as slipped-strand mispairing, and unequal crossing-over may have played a significant role in coding sequences evolution. Furthermore, the presence of low complexity segment in recently sequences (specific protein to one specie and hypothetical protein), as well as in high conserved protein (suggesting that may have already been present in the last common ancestor thus their

appearance is older than microbial pathogens), suggest that this phenomenon has been operated since early evolution of life, where the simple sequences were a major source of random phenotypic variation, but also a rapid mechanism to obtained.

The mutation that produce the low complexity segment appear at the long of protein, however, are selected zones of the sequences that produce lees structural problem, the C- and N- ends of proteins. The compositional biases of simple sequences tend to be rich in alanine, serine, isoleucine, leucine, lysine, glutamic acid, while aspartic acid, metionine, histidine, arginine, tyrosine, tryptophan, and cysteine are clearly underrepresented. Where the high mass amino acid (M, H, F, R, Y, W), are selected against, together with cysteine, proline, and aspartic acid, that produce several structural changes. Simple sequences appear to be present in all kinds of functional classes of proteins. This suggests that their role is not restricted to one type of function. However, they are clearly over-represented in membrane and signal transport proteins.

References

Bebenek K, Kunkel T A (1990) Frameshift errors initiated by nucleotide misincorporation. Proc Natl Acad Sci USA 87:4946-4950

Britten R J, Kohne D E (1968) Repeated Sequences in DNA. Science 161:529

Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y, (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453-1474

Bult C J, White O, Olsen G J, Zhou L, Fleischmann R D, Sutton G G, Blake J A, FitzGerald L M , Clayton R A, Gocayne J D, Kerlavage A R , Dougherty B A, Tomb, J F, Adams MD, Reich C I, Overbeek R, Kirkness E F, Weinstock K G, Merrick J M, Glodek A, Scott J L, Geoghagen N S M, Venter J C (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschi.* Science 273:1017-1140.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence, Nature 393:537-44

Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M, Huber R, Feldman RA, Short JM, Olsen GJ, Swanson RV. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 392:353-8

Epplen J T, Riess o (1997) Repetitive sequences in DNA. In Bishop M J, Raulings (eds) DNA and protein sequence analysis: a practical approach. IRL Press, Oxford 185-195

Fleischmann 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, Spriggs T, Hedblom 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, Geoghagen N S M, Gnehm C L, McDonald L A, Small K V, Freiser C M, Smith, H O, Venter J C (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496-512.

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, Merrick J M, Tomb J F, Dougherty B A, Bott K F, Hu P C, Lucier T S, Peterson S N, Smith H O, Hutchinson III C A, Venter J C (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270, 397-403.

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, Gwinn M, Dougherty B, Tomb JF, Fleischmann RD, Richardson D, Peterson J, Kerlavage AR, Quackenbush J, Salzberg S, Hanson M, van Vugt R, Palmer N, Adams MD, Gocayne J, Venter JC, et al. (1997), Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390:580-6

Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum KA, Sodergren E, Hardham JM, McLeod MP, Salzberg S, Peterson J, Khalak H, Richardson D, Howell JK, Chidambaram M, Utterback T, McDonald L, Artiach P, Bowman C, Cotton MD, Venter JC, et al. (1998), Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281:375-88

Goffeau et. al., Nature 387 (Suppl.) 5-105 (1997)

Goto S, Bono H, Ogata H, Fujibuchi W, Nishioka T, Sato K, and Kanehisa M (1996) Organizing and computing metabolic pathway data in terms of binary relations. Pacific Symp. Biocomputing 2:175-186

Hancock J M (1995) The contribution of slippage-like processes to the genome evolution. J. Mol. Evol. 41:1038-1047

Hancock J M (1996) Simple sequences in a ´minimal´ genome. Nature 14.14-15

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R (1996), Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res 24(22):4420-49

Kanehisa M, and Goto S (1997) A systematic analysis of gene functions by the metabolic pathway database. In Suhai S (ed) Theoretical and Computational Methods in Genome Research, Plenum Press

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 3:109-36

Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi E, Hosoyama A, Nagai Y, Sakai M, Ogura K, Otsuka R, Nakazawa H, Takamiya M, Ohfuku Y, Funahashi T, Tanaka T, Kudoh Y, Yamazaki J, Kushida N, Oguchi A, Aoki K, Kikuchi H. (1998), Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3 (supplement). DNA Res 5:147-55

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Venter JC, et al. (1997), The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 390:364-70

Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, Borriss R, Boursier L, Brans A, Braun M, Brignell SC, Bron S, Brouillet S, Bruschi CV, Caldwell B, Capuano V, Carter NM, Choi SK, Codani JJ, Connerton IF, Danchin A, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390:249-56

Lazcano A, Fox G E, Oró J F (1992) Life before DNA: The origin and evolution of early Archean cells. In Mortlock R P (ed) Evolution of metabolic function. Boca Raton Ann Arbor 237-295

Moxon ER, Rainey PB, Nowak MA, and Lenski RE (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. Current Biology 4(1):24-33

Moxon ER (1999)Whole-genome analisys of phatogens. In Stephen C. Stearns (ed) Evolution in Health & diseases. Oxford University Press.

Newfeld SJ, Tachida H, and Yedvobnick B (1994) Drive-selection equilibrium: homopolymer evolution in the Drosophila gene *mastermind*. J Mol Evol 38(6):637-41

Pearson W R, and Lipman D J (1988) Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85:2444-2448

Peixoto AA, Campesan S, Costa R, and Kyriacou CP (1993) Molecular evolution of repetitive region within the *per* gene of Drosophila. Mol Bio Evol 10(1):127-39

Richards R I, Sutherland GR (1994) Simlple repeat DNA is not replicated simply. Nature Genet 6:114-116

Saunders N J, Peden J F, Hood D W, Moxon R (1998) Simple sequence repeats in the *Helicobacter pylori* genome. Mol. Microbiol. 27:1091-1098

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, Harrison D, Hoang L, Keagle P, Lumm W, Pothier B, Qiu D, Spadafora R, Vicaire R, Wang Y, Wierzbowski J, Gibson R, Jiwani N, Caruso A, Bush D, Reeve JN, et al. (1997), Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J Bacteriol 179:7135-55

Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, Koonin EV, Davis RW. (1998), Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282:754-9

Sonnhammer E L L, Durbin R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1-10

Tautz D, Renz M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. 12:4127-4138

Tautz D., Trick M, Dover G A (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322: 652-656

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou L, Kirkness EF,

Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzegerald LM, Lee N, Adams MD, Venter JC, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388:539-47

Treier M, Pfeifle C, and Tautz D (1989) Comparison of yhe gap segementation gene *hunchback* between *Drosophila melanogaster* and *Drosophila virilis* reveals novel modes of evolutionary change. EMBO J. 8(5): 1517-25

Wootton J and Federhen S (1993) Statiscs of local complexity in amino acid sequences and sequence databases. Comp Chem. 17:149-163

Wootton J (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. Comp Chem 18:269-285

Figures caption

Table 1
Presence of amino acid simples sequences in complete proteomes; total: number of coding sequences per genome (ORFs), percentage of simple sequences per proteome, and aa that are present in low complexity segment.

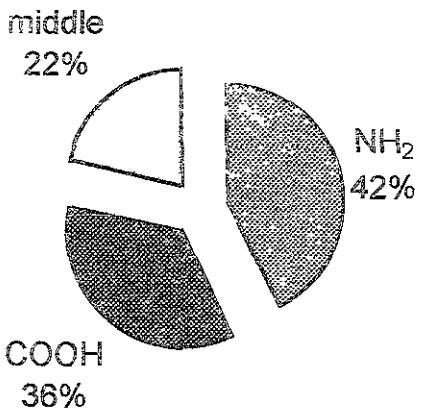
Figure 1
Percents of position of low complexity segment among the ORFs

Figure 2
Percents of amino acid composition from the total of low complexity in all complete proteomes (bars), and percent of occurrences amino acid in data bases (http://chait-sgi.rockefeller.edu/aainfo/struct.htm)

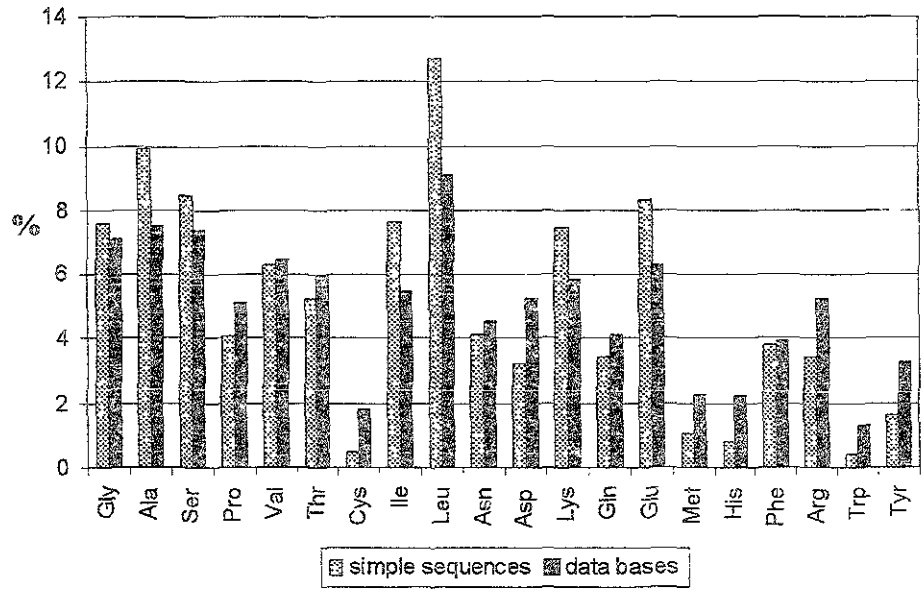
Figure 3
Amino acid composition in low complexity segment from coding sequences of complete genomes (surface graphic). *E. coli* (Ec), *H. influenzae* (Hi), *H. pylori* 26695 (Hp), *H. pylori* J99 (Hp99), *R. prowazekii* (Rp), *B. subtilis* (Bs), *M. genitalium* (Mg), *M. pneumoniae* (Mpn), *D. radiodurans* (Dr) *C. jejuni* (Cj), *M. tuberculosis* (Mt), *C. trachomatis* (Ct), *B. burgdorferi* (Bb), *T. pallidum* (Tp), *A. aeolicus* (Aa), *Synechocystis* sp. (Sy), *T. maritima* (Tm) *M. jannaschii* (Mj), *M. thermoautotrophicum* (Mth), *A. fulgidus* (Af), *P. horikoshii* (Ph), *S. cerevisiae* (Sc), and *C. elegans* (Ce). Showed a) mass, b) hydrophobicity and c) tendency of former α-helix of the amino acid.

Figure 4
Simples sequences percentages among metabolic pathways from all complete genomes. I) Carbohydrate Metabolism, II) Energy Metabolism, III) Lipid Metabolism IV) Nucleotide Metabolism, V) Amino Acid Metabolism, VI) Metabolism of Other Amino Acids, VII) Metabolism of Complex Carbohydrates, VIII) Metabolism of Complex Lipids, IX) Metabolism of Cofactors, Vitamins, and Other Substances, X) Metabolism of Macromolecules, XI) Membrane Transport, XII) Signal Transduction, XIII) Molecular Assembly, and XIV) Unassigned

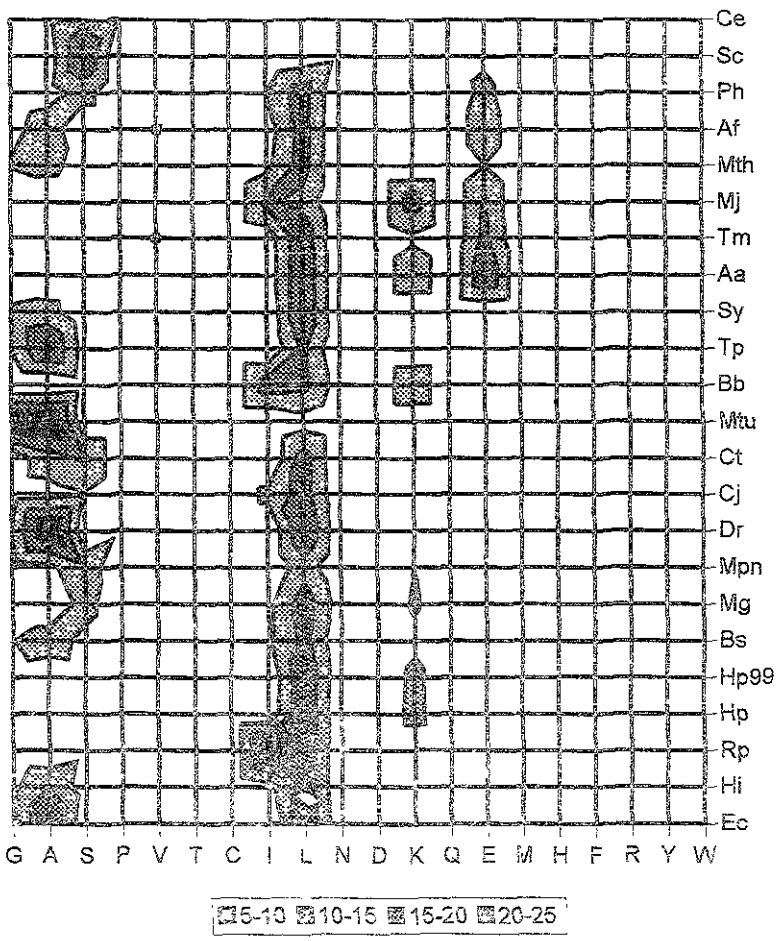
proteome	total ORFs	simple seq. %
Escherichia coli	4289	20.84
Haemophilus influenzae	1717	17.70
Helicobacter pylori 26695	1565	23.06
Helicobacter pylori J99	1491	24.81
Rickettsia prowazekii	834	19.54
Bacillus subtilis	4100	18.82
Mycoplasma genitalium	467	23.34
Mycoplasma pneumoniae	677	26.88
Mycobacterium tuberculosis	3918	43.82
Campilobacter jejuni	608	19.40
Chlamydia trachomatis	894	24.83
Borrelia burgdorferi	1256	28.58
Treponema pallidum	1031	25.21
Deinococcus radiodurans	3103	43.18
Aquifex aeolicus	1522	29.54
Thermotoga maritima	1846	20.74
Synechocystis sp	3166	21.76
Methanococcus jannaschii	1770	27.62
Methanobacterium thermoautotrophicum	1869	18.94
Archaeoglobus fulgidus	2407	21.18
Pyrococcus horikoshii	2061	24.55
Saccharomyces cerevisiae	6126	45.01
Caenorhabditis elegans	19099	41.37



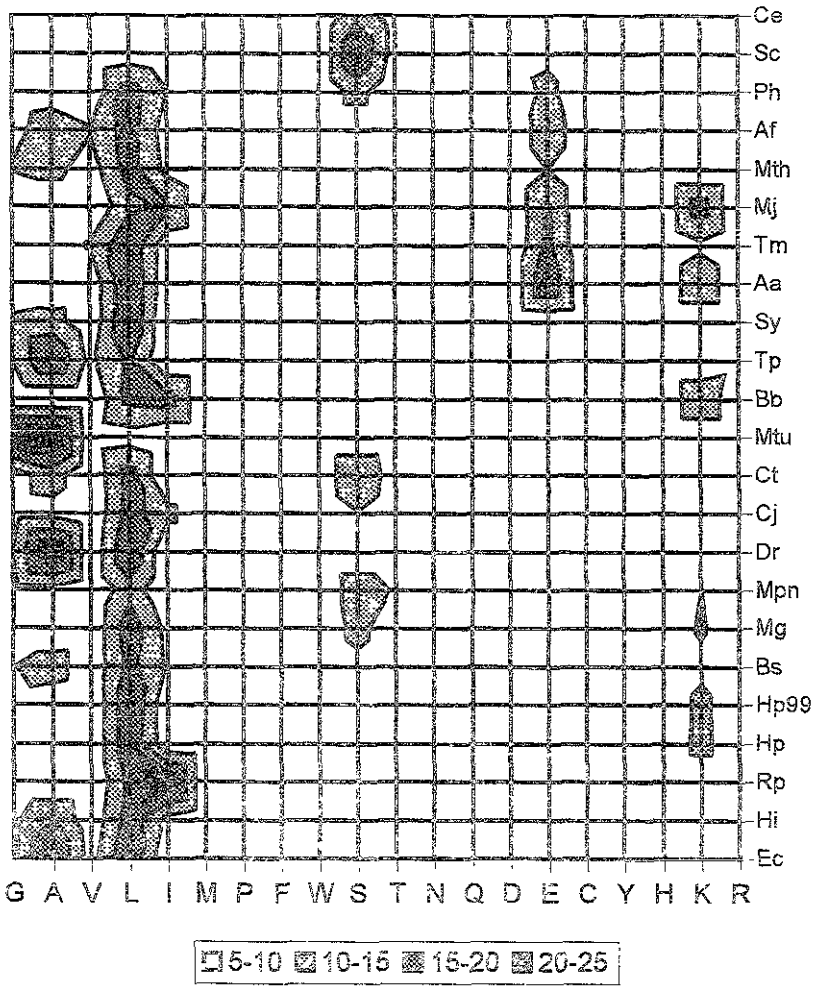
n=15,263


14
12
10
8
6
4
2
0
%
Gly
Ala
Ser
Pro
Val
Thr
Cys
Ile
Leu
Asn
Asp
Lys
Gln
Glu
Met
His
Phe
Arg
Trp
Tyr
simple sequences
data bases

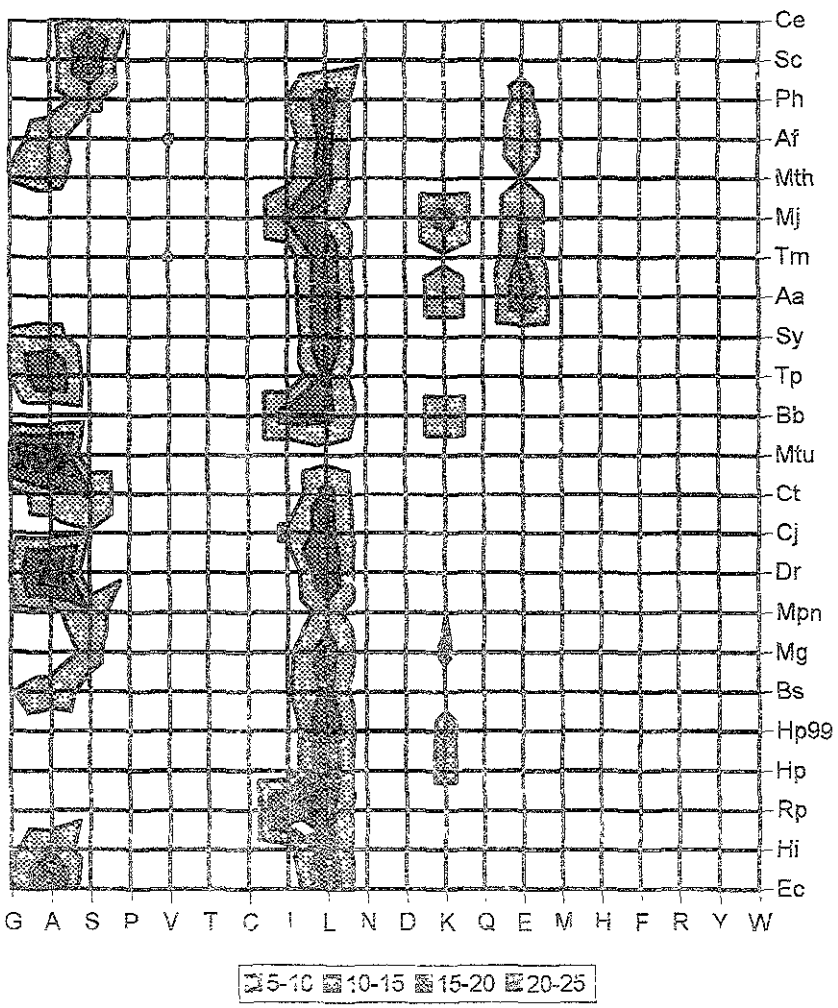
a)

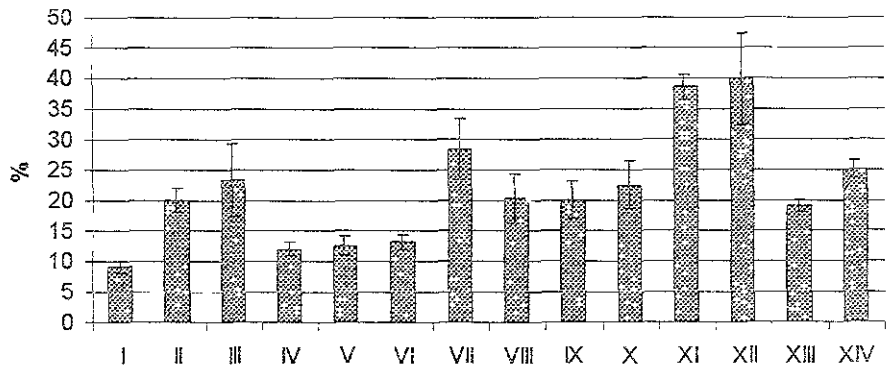

Ce
Sc
Ph
Af
Mth
Mj
Tm
Aa
Sy
Tp
Bb
Mtu
Ct
Cj
Dr
Mpn
Mg
Bs
Hp99
Hp
Rp
Hi
Ec
G A S P V T C I L N D K Q E M H F R Y W
5-10 10-15 15-20 20-25

b)


Ce
Sc
Ph
Af
Mth
Mj
Tm
Aa
Sy
Tp
Bb
Mtu
Ct
Cj
Dr
Mpn
Mg
Bs
Hp99
Hp
Rp
Hi
Ec
G A V L I M P F W S T N Q D E C Y H K R
5-10
10-15
15-20
20-25

c)


Ce
Sc
Ph
Af
Mth
Mj
Tm
Aa
Sy
Tp
Bb
Mtu
Ct
Cj
Dr
Mpn
Mg
Bs
Hp99
Hp
Rp
Hi
Ec
G A S P V T C I L N D K Q E M H F R Y W
5-10 10-15 15-20 20-25


50
45
40
35
30
25
20
15
10
5
0
%
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XIV

Internal Microbiol (1999) 2:3–10

REVIEW ARTICLE

Juli G. Pereto[1]
Ana María Velasco[2]
Arturo Becerra[2]
Antonio Lazcano[2]

[1]Department of Biochemistry and Molecular Biology, University of Valencia, Spain
[2]Faculty of Sciences, UNAM, México

Comparative biochemistry of CO_2 fixation and the evolution of autotrophy

Received 15 October 1998
Accepted 25 November 1998

Summary Carbon dioxide fixation is a polyphyletic trait that has evolved in widely separated prokaryotic branches. The three principal CO_2-assimilation pathways are (i) the reductive pentose-phosphate cycle, i. e. the Calvin-Benson cycle; (ii) the reductive citric acid (or Arnon) cycle; and (iii) the net synthesis of acetyl-CoA from CO/CO_2, or Wood pathway. Sequence analysis and the comparative biochemistry of these routes suggest that all of them were shaped to a considerable extent by the evolutionary recruitment of enzymes. Molecular phylogenetic trees show that the Calvin-Benson cycle was a relatively late development in the (eu)bacterial branch, suggesting that some form(s) of carbon assimilation may have been operative before chlorophyll-based photosynthesis. On the other hand, the ample phylogenetic distribution of both the Arnon and the Wood pathways does not allow us to infer which one of them is older. However, different lines of evidence, including experimental reports on the NiS/FeS-mediated C–C bond formation from CO and CH_3SH are used here to argue that the first CO_2-fixation route may have been a semi-enzymatic Wood-like pathway.

Key words Calvin-Benson cycle · Arnon cycle · Wood pathway · Semi-enzymatic synthesis · Carbon dioxide assimilation

Correspondence to:
Juli G. Pereto, Department of Biochemistry and Molecular Biology, University of Valencia, Dr. Moliner, 50, 46100 Burjassot, Spain.
E-mail: juli.pereto@uv.es

Introduction

Several different mechanisms for biological CO_2 fixation account for the diversity and evolutionary success of autotrophic life. According to the classical formulation of the heterotrophic theory of the origin of life [30], once the supply of abiotic organic compounds had become a limiting factor, primitive cells evolved other ways of obtaining carbon and energy. This led first to the development of photoautrophy, and afterwards to oxygen-releasing photosynthesis [30]. Several lines of evidence support the antiquity of the reductive pentose-phosphate pathway, or Calvin-Benson cycle. These include (i) the cyanobacteria-like microfossils in the $3.5 \cdot 10^9$ year old Australian Apex sediments, which suggest that the Calvin-Benson cycle appeared during early Archean times [35]; and (ii) the isotopic fractionation profiles of the early Archaean carbon cycle, which are consistent with the ribulose bisphosphate carboxylase/oxygenase (rubisco)-catalyzed carbon fixation process [15]. However, 16/18S rRNA-based universal phylogenies indicate that chlorophyll-based photosynthesis was a relatively late development in the (eu)bacterial branch [31, 45]. Thus, it is likely that the first autotrophs used chemical energy rather than light, and that the reductive pentose-phosphate pathway was preceded by alternative, older modes of CO_2-assimilation autotrophy.

Carbon dioxide assimilation is a widespread biological trait, but the biochemical dissimilarities between different pathways by which it takes place suggest that this ability evolved convergently in widely separated prokaryotes. In addition to the Calvin-Benson cycle, there are other CO_2-assimilation mechanisms, including (i) the reductive citric acid pathway, or Arnon cycle; (ii) the reductive acetyl-CoA cycle, i.e., the Wood pathway (sometimes also referred to as the Ljundahl-Wood pathway); and (iii) other less common mechanisms, such as the hydroxypropionate pathway first found in *Chloroflexus*, a green non-sulfur photosynthetic bacteria.

Twenty years ago it was argued that the ribulose-diphosphate cycle evolved via the ribulose monophosphate cycle in an ancestral heterotrophic population [32]. More recently, Wächtershäuser [41] has proposed a chemoautotrophic scheme of the origin of life in which pyrite formation is linked with early CO_2 fixation. In this paper we propose that none of these two alternatives is correct, and that the phylogenetic distributions of the Arnon and Wood pathways do not indicate by themselves which of these two cycles is the oldest. We also argue that energetic considerations and the experimental evidence on the NiS/FeS-mediated formation of C–C bonds from CO and CH_3SH [17] can be interpreted as supporting the hypothesis that a Wood-like semi-enzymatic pathway was the earliest biological carbon fixation route.

Biological carbon fixation can take place by different mechanisms

The reductive pentose-phosphate pathway, or Calvin-Benson cycle During the 1950s Calvin and his associates established, in a series of elegant experiments, the pathway by which CO_2 is assimilated by photoautotrophic eukaryotes [5]. The Calvin-Benson cycle, which originated in the cyanobacterial ancestors of chloroplasts, is the outcome of a process in which enzyme recruitment had a major role. As summarized in Table 1, in biochemical terms there are only two enzymes unique to the Calvin-Benson cycle, namely phosphoribulokinase, or ribulose 5-phosphate kinase (PRK), and ribulose bisphosphate (RuBP) carboxylase/oxygenase (rubisco). The other eleven chemical reactions are catalyzed by eight different enzymes (nine in plastids) that have additional roles in several heterotrophic pathways, such as glycolysis/gluconeogenesis and the pentose-phosphate oxidative route. Thus, a major portion of the Calvin cycle may be explained as the result of a patchwork assembly of a route [20] from pathways already extant in previously evolved heterotrophic anaerobes, such as the ability to synthesize pentoses from C_3- or C_6-compounds [29].

Analysis of the completely-sequenced genomes of *Methanococcus jannaschii* [4] and the closely related euryarchaeota *Archaeglobus fulgidus* [23] has led to the identification of ORFs which exhibit considerable levels of similarity with the rubisco large subunit (Table 1). Identification of rubisco-homologues in these euryarchaeotal genomes confirms previous reports of the presence of this enzyme in some non-retinal-containing species of halobacteria [1]. However, there is no report of Calvin-Benson cycle-based autotrophic growth in archaebacteria, and it has been suggested that the presence of rubisco-like sequences in archaeal genomes is due to horizontal transfer phenomena [22].

Primary structure comparisons do not indicate any obvious evolutionary relationships between rubisco and all the other

Table1 Enzymatic steps in the Calvin-Benson cycle

Enzyme	EC	Reaction	Other pathways	Distribution
phosphoribulokinase	2.7.1.19	ATP + D-ribulose 5-phosphate = ADP + D-ribulose 1,5 bisphosphate		B, E
ribulose bisphosphate carboxylase	4.1.1.39	D-ribulose 1,5-bisphosphate + CO_2 = 2,3-phospho-D-glycerate	glyoxylate & dicarboxylate metabolism	B, E (Small subunit) B, A, E (Large subunit)
phosphoglycerate kinase	2.7.2.3	ATP + 3-phospho-D-glycerate = ADP + 3-phospho-D-glyceroyl phosphate	glycolysis/gluconeogenesis	B, A, E
glyceraldehyde-3-phosphate dehydrogenase	1.2.1.13	D-glyceraldehyde 3-phosphate + phosphate + $NADP^+$ = 3- phospho-D-glyceroyl phosphate + NADPH		B, E
triosephosphate isomerase	5.3.1.1	D-glyceraldehyde 3-phosphate = glycerone phosphate	glycolysis/gluconeogenesis: fructose & mannose metabolism glycerolipid metabolism	B, A, E
fructose-bisphosphate aldolase	4.1.2.13	D-fructose 1,6-bisphosphate = glycerone phosphate + D-glyceraldehyde 3-phosphate	glycolysis/gluconeogenesis pentose phosphate cycle: fructose & mannose metabolism	B, E
fructose-bisphosphatase	3.1.3.11	D-fructose 1,6-bisphosphate + H_2O = D-fructose 6-phosphate + phosphate	glycolysis/gluconeogenesis pentose phosphate cycle: fructose & mannose metabolism	B, E
transketolase	2.2.1.1	Sedoheptulose 7-phosphate + D-glyceraldehyde 3-phosphate = D-ribose 5-phosphate + D-xylulose 5-phosphate	pentose phosphate cycle	B, A, E
ribulose-phosphate 3-epimerase	5.1.3.1	D-ribulose 5-phosphate = D-xylulose 5-phosphate	pentose phosphate cycle: pentose & glucuronate interconversions	B, A, E
ribose 5-phosphate epimerase	5.3.1.6	D-ribose 5-phosphate = D-ribulose 5-phosphate	pentose phosphate cycle	B, A, E

B= Bacteria; A= Archaea; E= Eukarya

Fig. 1 Multiple sequence alignment of a conserved phosphoribulokinase (ribulose 5-phosphate kinase) motif with its bacterial and eukaryotic uridine kinase/cytidine kinase homologues

Abbreviations: prk, phosphoribulokinase; udk, uridine kinase/cytidine kinase; At, *Arabidopsis thaliana*; Mc, *Mesembryanthemum crystallinum*; Ta, *Triticum aestivum*; So, *Spinacia oleracea*; Cr, *Chlamydomonas reinhardtii*; Ssp, *Synechocystis* sp. PCC6803; Bs, *Bacillus subtilis*; Ec, *Escherichia coli*; Mg, *Mycoplasma genitalium*; Sc, *Saccharomyces cerevisiae*

sequences found in the available databases. On the other hand, phosphoribulosekinase (*prk*) is probably derived through gene duplication and divergence events from an ancestral, less specific kinase. This possibility is supported by a conserved 200-odd amino acid segment that phosphoribulokinase sequences share with uridine kinase/cytidine kinase (*udk*), a pyrimidine ribonucleotide biosynthetic enzyme that catalyzes the U(C)+ GTP = U(C)MP + GDP reaction (Fig. 1).

In every examined case, the carboxylating enzyme rubisco has been shown to perform the oxygenolytic cleavage of RuBP. This bifunctional character is the result of the competition between O_2 and CO_2 for the same catalytic site. From an autotrophic viewpoint, the inhibitory effect of oxygen on carbon assimilation (or photorespiration) appears to be a wasteful process. However, the ancestral carboxylase probably evolved in a CO_2-rich environment in which very little free oxygen was available. Throughout Earth's geological history this situation has been reverted, and rubisco appears as the outcome of an adaptative process that led to an increase in the ratio of substrate specificities CO_2/O_2 that compensated for the O_2 inhibitory effect. Analysis of the different strategies followed by the different photosynthetic groups that contain rubisco to reduce the impact of increasing oxygen pressures, such as the CO_2-concentrating mechanisms in cyanobacteria and algae, and the Hatch-Slack pathway in C4 plants, make this trend rather obvious [7].

Although it is possible that subterranean lithotrophs contribute significantly to biological carbon fixation [31], today the Calvin-Benson cycle appears to be responsible for the bulk of biological CO_2-fixation. It is present in all photosynthetic microorganisms, including cyanobacteria and purple bacteria, and also in some Gram-positive chemolithotrophs. As reviewed by Margulis [28], this appears to be the only autotrophic route acquired by eukaryotes through symbiotic events, involving either photoautotrophic or chemolithotrophic prokaryotes.

The Arnon cycle: the reductive citric acid pathway In 1966 Arnon and his coworkers proposed that carbon assimilation in the bacterium *Chlorobium limicola* (which photochemically disproportionates elementary sulphur to sulphide and sulphate), proceeds not by the standard Calvin-Benson cycle, but via a reductive citric acid cycle, or reverse Krebs cycle [3]. As summarized in Table 2, this pathway requires two additional enzymes from those involved in the cyclic oxidation of acetyl-CoA, namely ferredoxin-dependent 2-oxoglutarate, and ATP-citrate lyase*. Recent structural comparisons have shown that

* Our search for homologues of ATP -citrate lyase indicated a considerable level of sequence similarity with NCBI entry 482640 (pir A60956), which was originally deposited as a 300 amino acid fragment of the sea urchin embryonic ciliary dynein beta heavy chain [11]. However, detailed analyses of both dynein and ATP- citrate lyase sequences demonstrated that NCBI 482640 was misidentified, and that it is not a member of the dyneins but a ATP citrate lyase.

Table 2 Enzymatic steps in the reductive citric acid cycle (Arnon pathway)

Enzyme	EC	Catalyzed reaction	Other pathways	Distribution
2-oxoglutarate synthase	1.2.7.3	2-oxoglutarate + CoA + oxidized ferredoxin = succinyl CoA + CO_2 + reduced ferredoxin	citric acid cycle	B, A
isocitrate dehydrogenase ($NADP^+$)	1. 1.1.42	isocitrate + $NADP^+$= 2- oxoglutarate + CO_2 + NADPH	citric acid cycle; glutathione metabolism	B, A, E
aconitate hydratase	4. 2. 1. 3	citrate= cis-aconitate + H_2O	glyoxylate & dicarboxylate metabolism; citric acid cycle	B, E
ATP-citrate lyase	4. 1. 3. 6	citrate = acetate + oxacetate	citric acid cycle	B
malate dehydrogenase	1.1.1.37	(s)-malate + NAD^+ = oxaloacetate + NADH	citric acid cycle; pimvate, glyoxylate & dicarboxylate metabolism	B, A, E
fumarate hydratase	4.2.1.2	(s)-malate = fumarate + H_2O	citric acid cycle	B, A, E
succinate dehydrogenase	1.3.99.1	succinate + acceptor = fumarate + reduced acceptor	citric acid cycle; oxidative phosphorilation; butanoate metabolism	B, E
succinate-CoA ligase (ADP-forming)	6.2.1.5	ATP + succinate + CnA = ADP + succinyl CoA + phosphate	citric acid cycle; propanoate metabolism; C5-branched dibasic acid metabolism	B, A, E

B= Bacteria; A= Archaea; E= Eukarya

ATP citrate lyase belongs to an enzyme superfamily characterized by an unusual nucleotide-binding fold, i.e., the palmate- or ATP-grasp fold. This superfamily includes other ATP-dependent carboxylate-thiol ligases (succinate- and malate-CoA ligases), as well as enzymes endowed with carboxylate-amine ligase activity (glutathione synthetase, biotin carboxylase, and carbamoyl-phosphate synthetase) [13].

The reductive citric acid cycle is found in both bacterial and archaeal prokaryotes. It was first reported in the moderately thermophilic hydrogen-oxidizing *Hydrogenobacter thermophilus*, the aerobic *Aquifex pyrophilus*, and the sulphate reducer proteobacteria *Desulfurobacter hydrogenophilus*, as well as in archeal species including members of the aerobically grown *Sulfolobus* genus, and *Thermoproteus neutrophilus* (when grown with H_2 and elemental sulphur) [12, 34]. The wide distribution of this anabolic pathway and its modifications (such as the reductive acetyl-CoA or the reductive malonyl-CoA pathways) among anaerobic archaea and the most deeply rooted eubacteria strongly suggest that it evolved prior to the Calvin-Benson cycle [22]. This cycle is currently favored as the primordial metabolic pathway by the adherents of the pyrite-based chemoautotrophic theory of the origin of life [41].

The reductive acetyl CoA cycle (Wood pathway) Although A. F. Lebedeff had suggested in 1908 that direct assimilation of CO_2 is a widespread biological trait, it was not until sixty years later that Wood and his co-workers [26] demonstrated the fixation of atmospheric carbon dioxide into reduced organic compounds by the heterotrophic propionic acid bacteria [47, 48]. Further studies demonstrated that assimilation of CO_2 by the net synthesis of acetyl CoA could sustain autotrophic growth in *Clostridium thermoaceticum* [50]. Fixation by the autotrophic reductive acetate pathway is a simple process that involves the combination of two CO_2 (or CO) molecules to form a two-carbon compound, from which reduced organic components are formed by non-autotrophic or anaplerotic carboxylation processes and other typical heterotrophic reactions [48]. The first reaction in this pathway is the reduction of carbon dioxide to CH_3, which reacts as methyltetrahydrofolate with CoASH to form acetyl-CoA via a carboxyl donor such as CO or CO_2. Instead of the typical biotin carboxyl carrier protein found in *E. coli* and animals, the intermediates are carrier-bound organometallic complexes involving Ni, Fe, and S. In this cycle the breakage and formation of the thioester bond between CoASH and the C=O group of the acetyl moiety are both catalyzed by CO dehydrogenase/acetyl CoA synthase. This bifunctional enzyme is the central catalyst in this pathway, and mediates both the oxidation of CO to CO_2 and the synthesis of acetyl-CoA [33, 49].

The Wood pathway, which is the major CO_2 fixation mechanism under anaerobic conditions [49], also has an ample phylogenetic distribution, and is known to be used by acetogenic bacteria, methanogens, and sulphate-reducers for both anabolic and catabolic purposes [44]. The list includes *Acetobacterium woodii* and *Sporomusa* sp., as well as the sulphate reducers *Desulfobacterium autotrophicum* and *Desulfovibrio baarsii*. It is also widely distributed among methanogens, including *Methanobacterium thermoautotrophicum*, *Methanosarcina barkeri*, and in the early diverging hyperthermophilic genera *Methanopyrus*, *Methanococcus*, and *Methanothermus* [12, 34]. In spite of

the differences in formate utilization and the peculiar cofactors employed by methanogens, the first steps of acetyl-CoA synthesis are similar among these autotrophic archaea and the eubacteria [21, 43, 46], confirming the monophyletic origin of this pathway. Although no carbon monoxide dehydrogenase activity has been found in *Archaeoglobus profundis*, it has been reported in *A. lithotrophicus* [38]. This finding is consistent with the presence of the Wood pathway genes in the *A. fulgidus* genomes [23], where they are probably involved in the anaerobic oxidation of acetate to CO_2.

The hydroxypropionate pathway Several studies on the photoautotrophic growth of the thermophilic, green non-sulphur bacterium *Chloroflexus aurantiacus* indicated that, in some strains, CO_2 was assimilated into reduced organic compounds via a pathway different from those described above. Labelling experiments with cells grown in the presence of the aconitase-blocking fluoroacetate, demonstrated that acetyl-CoA could be both an intermediate and a product of this CO_2-fixation pathway [16, 36]. This cyclic route involves the carboxylation of acetyl-CoA, which is then reductively converted to form 3-hydroxy-propionate [16]. This unusual intermediate is first reduced and carboxylated to form propionyl-CoA, which is converted through a second carboxylation into succinate and then to malyl-CoA [10, 36, 37]. It is possible that the starting point of succinate biosynthesis via this pathway, which is present also in the archaea *Thermoproteus neutrophilus* [36] and in a somewhat modified form in *Acidianus brierleyi* [19], involves glyoxylate derived from the reordination and escission of 2-methyl malonyl CoA through the glyoxylate shunt (Stanley L. Miller, personal communication).

Is autotrophic CO_2-fixation a primordial process?

Several autotrophic theories on the origin of life have been proposed which do not require preformed organic compounds of abiotic origin. Two of these theories tie the origin of CO_2-assimilation pathways to the appearance of life, i.e., they assume the first living systems were already endowed with the ability of fixing atmospheric carbon dioxide. One such theory involving non-enzymatic reactions was patterned after extant biochemical pathways of intermediate metabolism assumes that the citric acid cycle started with acetyl-CoA by two CO_2-fixations [14]. The development of such a system is envisioned to require UV light, clays, and transition state metals, all of which are likely components of the primitive environment. However, such cyclic pathways need to be very efficient, or they will stop working. One such example is precisely the Krebs cycle, which comes to a complete standstill unless the oxaloacetate lost by non-enzymatic decarboxylation is replaced. In any case, this theory has not been examined in detail, and there is no experimental evidence supporting its basic assumptions [25].

Currently the most popular and elaborate autotrophic theory on the origin of life is that of Wächtershäuser [39–42]. According to this hypothesis, life began with the appearance of an autocatalytic two-dimensional chemolithotrophic metabolic system based on the conversion of iron sulphide into the stable crystalline mineral pyrite (FeS_2). Synthesis and polymerization of organic compounds are assumed to have taken place on the surface of FeS and FeS_2 in environments that resemble those of deep-sea hydrothermal vents. Replication followed the appearance of non-organismical iron sulfide two-dimensional life, in which chemoautotrophic carbon fixation is assumed to have taken place by a reductive tricarboxylic (TCA) or citric acid cycle of the type originally described for the photosynthetic green sulphur bacterium *Chlorobium limicola*. The ample phylogenetic distribution of this anabolic cycle and its modifications (such as the reductive acetyl-CoA and the reductive malonyl-CoA pathways) have been interpreted as evidence of its primordial character, which is assumed to have been primed by a carbon dioxide fixation process akin to the reductive acetyl-CoA pathway [22, 27, 41].

The reaction $FeS + H_2S = FeS_2 + H_2$ is highly exergonic ($\Delta G^\circ = -9.23$ kcal/mol, $E^\circ = -620$ mV), and it has been demonstrated to take place under anaerobic conditions at neutral pH and 100°C [9]. The FeS/H_2S combination is a strong reducing agent, and it provides an efficient source of electrons for the reduction of organic compounds under mild conditions. The FeS/H_2S combination can produce molecular hydrogen, and reduce nitrate to ammonia, acetylene to ethylene, thioacetic acid to acetic acid, as well as more complex synthesis [27], including peptide-bond formation by activation with carbon monoxide on (Ni, Fe)S surfaces [18].

However, these experimental results are also compatible with a more general, modified model of the classical heterotrophic theory in which pyrite formation is recognized as an important source of electrons for the reduction of organic compounds [24]. It is possible, for instance, that under certain conditions atmospheric CO_2 would have been photoreduced by ferrous iron in solution, and pyrite formation on submerged rocks might have reduced nitrogen to ammonia [2] and organic compounds. The essential question in deciding between this chemoautotrophic theory and the heterotrophic hypothesis on the origin of life is clearly not pyrite-mediated organic synthesis, but whether direct CO_2 reduction and synthesis of organic compounds can take place by a hypothetical two-dimensional living system that lacks genetic information [24].

A semi-enzymatic model for the first CO_2-fixation pathway

Given the strong dissimilarities in carbon dioxide assimilation routes that have developed in widely separated autotrophs, it is important to understand which of them is oldest and

how it evolved from previous heterotrophic modes of nutrition. Answers to these questions should provide a coherent, non-teleological historical narrative consistent with (i) the phylogenetic distribution of autotrophic metabolisms; (ii) the geological conditions of the early Earth (such as its anoxic environment) and the availability of inorganic precursors and catalysts; (iii) recognition of the limited catalytic abilities of the first autotrophs, i.e., the earliest biological CO_2 assimilation pathways must have been relatively simple, and may have depended on spontaneous or semi-enzymatic carboxylation reactions leading to C–C bond formation.

Quayle and Ferenci [32] suggested a slow, stepwise development of the highly endergonic ribulose bisphosphate pathway in which the simpler ribulose-monophosphate cycle was assumed to be an intermediate stage in the evolution of rubisco-mediated CO_2 assimilation. However, this model is not consistent with the phylogenetic distribution of the ribulose biphosphate pathway, which appears to be a relatively late development in the (eu)bacterial branch. Thus, it is likely that an older form of chemoautotrophic carbon assimilation evolved before the appearance of chlorophyll-based photoautotrophy.

Adherents of the pyrite-based chemoautotrophic theory of the origin of life, on the other hand, have argued that the reductive TCA cycle, which has an autocatalytic nature and provides both archaeal and bacterial metabolisms with the starting material for practically all biosynthetic routes, was originally driven by pyrite formation and the very first anabolic pathway [22, 27, 41].

The ubiquity of Fe-S active centers in many ancient enzymes, including CO dehydrogenase, has been explained as the evolutionary outcome from this FeS/H_2S-mediated reduction of organic compounds [8, 41]. The archaeal acetyl-CoA synthase (CH_3CO-SCoA), like the *Clostridium thermoaceticum* enzyme [51], also includes Ni in its Fe-S reaction center [6, 8], and uses CO_2 or CO as precursors for acetyl-CoA [47, 48]. Based on these observations and on the key role of acetyl-CoA in manifold biosynthetic pathways (Eq. 1), Huber and Wächtershäuser [17] developed a non-enzymatic synthesis of CH_3–CO–SCH_3 from a mixture of CO and CH_3SH in a high-temperature reaction catalyzed by a mixture of co-precipitated NiS/FeS (Eq. 2):

$$CH_3\text{-corrin} + CO + HSCoA \xrightarrow{\text{acetyl-CoA synthase}} CH_3CO(SCoA) \quad (1)$$

HSCoA = coenzyme A

$$2CH_3SH + CO \xrightarrow{NiS/FeS} CH_3CO(SCH_3) + H_2S \quad (2)$$

Addition of selenium to the catalytic mixture NiS/FeS led to the synthesis of acetic acid and CH_3SH (Eq. 3).

$$CO + H_2S \xrightarrow{Se/NiS/Fe_3S_4} CH_3COOH + CH_3SH \quad (3)$$

The abiotic C–C bond formation from CH_3SH and CO (Eq. 2), which is analogous to the metal-catalyzed industrial synthesis of CH_3COOH from CH_3OH and CO via the migration of a methyl group to a coordinated CO [8], demonstrates the feasibility of carbon monoxide fixation in a Wood-like reaction catalyzed by transition metal ions [17]. Although this reaction does not take place in a two-dimensional system as postulated by Wächtershäuser [39–42], the metal sulphide-catalyzed C–C bond-forming process has been interpreted as evidence of a CO-assimilation process that would feed an archaic autocatalytic chemoautotrophic carbon-fixation cycle. As required by Wächsterhäuser's theory, such a cycle must have been a primitive variant of the reverse citric acid cycle, which once sparked by a C–C forming process akin to the Wood cycle, would become the starting point for all anabolic pathways. The appearance of the ancestral reductive TCA cycle would then be followed by the development of the reductive acetyl-CoA pathway [41].

However, there is an alternative intepretation to the results reported by Huber and Wächtershäuser [17], i.e., the reaction summarized in Eq. 2 suggests the possibility that the Wood pathway had preceeded the reductive citric acid cycle. It is possible, for instance, that a semi-enzymatic Wood-like cycle evolved in an ancestral heterotrophic population of limited catalytic abilities. According to this alternative interpretation, the utilization of metal sulphides as reducing agents also corresponds to an early step in biochemical evolution, i.e., acetyl-CoA synthase is the evolutionary outcome of a simple Ni-Fe-S catalyst with carbon monoxide dehydrogenase activity. The C_2-units generated by a reaction equivalent to that shown in Eq. 2 could be incorporated into cell material following a ferredoxin-dependent (or pyrite-dependent) reductive carboxylation. Corrin-skeletons used first in the reduction of ribonucleotides in the RNA → DNA transition would be then selected as methyl-transfer molecules (CH_3-corrin), in a process originally mediated by broad-substrate primitive enzymes. This view is consistent with (i) the widespread distribution of the Wood pathway (Table 3); and (ii) the hypothesis that catalytic iron-sulphur clusters found in electron-transfer proteins have an ancient origin. However, it does not require a hot origin of life or an autotrophic emergence of living systems.

Conclusions

In this paper we have reviewed some of the biochemical characteristics of the basic CO_2-assimilation pathways. Sequence comparisons demonstrate that the patchwork assembly of catalysts has played a central role in the evolution of these different modes of carbon fixation. As underlined by Pace [31], phylogenetic distribution of the different types of energy metabolism and carbon dioxide fixation in universal molecular phylogenies does not follow a simple development

Table 3 Enzymatic steps in the reductive acetyl-CoA cycle (Wood pathway)

Enzyme	E C	Catalyzed reaction	Other pathways	Distribution
formate dehydrogenase	1.2.1.2	formate + $NAD^+ = CO_2$ + NADH	related to many other dehydrogenases	B, A, E
formyl tetrahydrofolate synthetase	6.3.4.3	ATP + formate + tetrahydrofolate = ADP + phosphate + 10- formyl tetrahydrofolate	glyoxylate & dicarboxylate metabolism	B, E
methenyl tetrahydrofolate	3.5.4.9	5,10-methylenetetrahydrofolate + H_2O = 10-formyl tetrahydrofolate	glyoxylate & dicarboxylate metabolism	B, E
methylene tetrahydrofolate dehydrogenase ($NADP^+$)	1.5.1.15	5,10-methylenetetrahydrofolate + $NADP^+$ = 5,10-methenyl tetrahydrofolate + NADPH	glyoxylate & dicarboxylate metabolism	B, E
5,10 methylene tetrahydrofolate reductase (FADH)	1.7.99.5 1.5.1.20	5-methyltetrahydrofolate + acceptor = 5,10-methylenetetrahydrofolate + reduced acceptor	dicarboxylate metabolism	B, A
carbon monoxide dehydrogenase	1.2.99.2	CO + H_2O + acceptor = CO_2 + reduced acceptor	methane metabolism	B, A, E

B= Bacteria; A= Archaea; E= Eukarya

and may indicate lateral gene transfer. However, it is obvious that some form of carbon assimilation, which is found even in strict heterotrophs, was operative before chlorophyll-based photosynthesis.

In contrast with the chemoautotrophic pyrite-based theory of the origin of life that assumes that the autotrophic reductive citric acid cycle was the first anabolic pathway [41, 42], we have proposed here that the experiments on metal sulphide-mediated C–C bond formation [17] can also be interpreted as an evidence that the Wood pathway is the most primitive carbon dioxide assimilation pathway. This alternative interpretation is consistent with (i) the ample phylogenetic distribution of the Wood cycle; (ii) the relative simplicity and energetically more favourable process of CO_2-assimilation mediated by the Wood pathway compared to other autotrophic routes [12]; and (iii) the possibility that the CO_2 assimilation was originally a semi-enzymatic pathway, in which the net synthesis of acetyl-CoA from CO/CO_2 was mediated by a simple NiS/FeS catalytic mixture ancestral to acetyl-CoA synthase [17].

Acknowledgments This work was initiated during a leave of absence of A. Lazcano as Visiting Professor at the University of Valencia, during which the hospitality of Juli Peretó and his associates was enjoyed. Financial support of CICyT (grant BIO96-0895) is gratefully acknowledged by J. P. Support for this work has been provided by the Proyecto PAPIIT (DGAPA-UNAM, Mexico) IN210598 to A. L.

References

1. Altekar W, Rajagopalan R (1990) Ribulose bisphosphate carboxylase activity in halophilic Archeabacteria. Arch Microbiol 153:169–174
2. Brandes JA, Boctor NZ, Cody GD, Cooper BA, Hazen RM, Yoder HS Jr (1998) Abiotic nitrogen reduction on the early Earth. Nature 395:365–367
3. Buchanan BB, Arnon DI (1990) A reverse Krebs cycle in photosynthesis: consensus at last. Photosynt Res 24:47–53
4. Bult CJ, 38 other, and Venter JC (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073
5. Calvin M (1962) The path of carbon in photosynthesis. Science 135:879–889
6. Cammack R (1996) Iron and sulfur in the origin and evolution of biological energy conversion system. In: Baltscheffsky H (ed) Origin and Evolution of Biological Energy Conversion. New York: VCH Publ, pp 43–69
7. Codd GA, Vakeria D (1987) Enzymes and genes of microbial autotrophy. Microbiol Sci 4:154–159
8. Crabtree RH (1997) Where smokers rule. Science 276:222
9. Drobner E, Huber C, Wächtershäuser G, Rose D, Stetter KO (1990) Pyrite formation linked with hydrogen evolution under anaerobic conditions. Nature 346:742–744
10. Eisenreich W, Strauss G, Werz U, Fuchs G, and Bacher A (1993) Retrobiosynthetic analysis of carbon fixation in the phototrophic eubacterium *Chloroflexus aurantiacus*. Eur J Biochem 215:619–632
11. Foltz KR, Asai DJ (1990) Molecular cloning and expression of sea urchin embryonic ciliary dynein beta heavy chain. Cell Motil Cytoskeleton 16:33–46
12. Fuchs G (1989) Alternative pathways of autotrophic CO_2 fixation. In: Schlegel HG, Bowien B(eds) Autotrophic Bacteria. Madison: Science Tech Publishers, pp 365–382
13. Galperin MY, Koonin EV (1997) A diverse superfamily of enzymes with ATP-dependent carboxylate-amine/thiol ligase activity. Protein Sci 6:2639–2643
14. Hartmann H (1975) Speculations on the origin and evolution of metabolism. *J Mol Evol* 4:359–370
15. Hayes JM (1994) Global methanotrophy at the Archean-Proterozoic transition. In: Bengtson S (ed) Early Life on Earth. Nobel Symposium No 84. New York: Columbia University Press, pp 220–236
16. Holo H (1989) *Chloroflexus aurantiacus* secretes 3-hydroxypropionate, a possible intermediate in the assimilation of CO_2 and acetate. Arch Microbiol 151:252–256

17. Huber C, Wächtershäuser G (1997) *Activated acetic acid by carbon fixation on (Fe,Ni)S under primordial conditions.* Science 276:245–247
18. Huber C, Wächtershäuser G (1998) Peptides by activation of amino acids with CO on (Ni, Fe)S surfaces and implications for the origin of life. Science 281:670–672
19. Ishii M, Miyake T, Satoh T, Sugiyama H, Oshima Y, Kodama T, Igarashi Y (1997) Autotrophic carbon dioxide fixation in *Acidianus brierleyi*. Arch Microbiol 166:368–371
20. Jensen RA (1976) Enzyme recruitment in the evolution of new function. Annu Rev Microbiol 30:409–425
21. Jetten MSM, Stams AJM, Zehnder AJB (1992) Methanogenesis from acetate: a comparison of the acetate metabolism in *Methanothrix soehngenii* and *Methanosarcina* spp. FEMS Microbiol Rev 88:181–198
22. Kandler O (1994) The early diversification of life. In: Bengtson S (ed) Early Life on Earth: Nobel Symposium No 84. New York: Columbia University Press, pp 152–160
23. Klenk HP, 49 other, and Venter JC (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 390:364–370
24. Lazcano A (1998) Origin of life. In: Briggs DEG, Crowther PR (eds) Palaeobiology II. Oxford: Blackwell Science.
25. Lazcano A, Miller SL (1996) The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. Cell 85:793–798
26. Ljungdahl L, Irion E, Wood HG (1965) Role of corrinoids in the total synthesis of acetate from CO_2 by *Clostridium thermoaceticum*. Biochemistry 4:2771–2780
27. Maden BEH (1995) No soup for starters? Autotrophy and the origin of metabolism. Trends Biochem Sci 20:337–341
28. Margulis L (1993) Symbiosis in Cell Evolution. New York: Freeman Co
29. McFadden BA, Tabita FR (1974) D-ribulose-1,5-diphosphate carboxylase and the evolution of autotrophy. BioSystems 6:93–112
30. Oparin AI (1938) The Origin of Life. New York: MacMillan
31. Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276:734–740
32. Quayle JR, Ferenci T (1978) Evolutionary aspects of autotrophy. Microbiol Rev 42:251–273
33. Ragsdale SW (1991) Enzymology of the acetyl-CoA pathway of CO_2 fixation. Crit Rev Biochem Mol Biol 26:261–300
34. Schönheit P, Schäfer T (1995) Metabolism of hyperthermophiles. World J Microbiol Biotech 11:26–57
35. Schopf JW (1993) Microfossils of the early Archean Apex chert: new evidence of the antiquity of life. Science 260:640–646
36. Strauss G, Eisenreich W, Bacher A, Fuchs G (1992) ^{13}C-NMR study of autotrophic CO_2 fixation pathways in the sulfur reducing Archaebacterium *Thermoproteus neutrophilus* and in the phototrophic Eubacterium *Chloroflexus aurantiacus*. Eur J Biochem 205:853–866
37. Strauss G, Fuchs G (1993) Enzymes of a novel autotrophic CO_2 fixation pathway in the phototrophic bacterium *Chloroflexus aurantiacus*, the 3-hydroxypropionate cycle. Eur J Biochem 215:633–643
38. Verhoit J, Kunow J, Stetter KO, Thauer RK (1995) Enzymes and coenzymes of the carbon monoxide dehydrogenase pathway for autotrophic CO_2 fixation in *Archaeoglobus lithotrophicus*, and the lack of carbon monoxide dehydrogenase in the heterotrophic *A. profundis*. Arch Microbiol 163:112–118
39. Wächtershäuser G (1988) Pyrite formation, the first energy source for life: a hypothesis. Syst Appl Microbiol 10:207–210
40. Wächtershäuser G (1988) Before enzymes and templates: theory of surface metabolism. Microbiol Rev 52:452–484
41. Wächtershäuser G (1990) Evolution of the first metabolic cycles. Proc Natl Acad Sci USA 87:200–204
42. Wächtershäuser G (1992) Groundworks for an evolutionary biochemistry: the iron-sulphur world. Prog Biophys Mol Biol 58:85–201
43. Weiss DS, Thauer RK (1993) Methanogenesis and the unity of biochemistry. Cell 72:819–822
44. White D (1995) The Physiology and Biochemistry of Prokaryotes. New York: Oxford University Press
45. Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271
46. Wolfe RS (1990) Novel coenzymes of archaebacteria. In: Hauska G, Thauer R (eds) The Molecular Basis of Bacterial Metabolism. Berlin, Springer-Verlag, pp 1–12
47. Wood HG (1972) My life and carbon dioxide. In: Woessner JF Jr, Huijing F (eds) The Molecular Basis of Biological Transport. New York: Academic Press, pp 1–54
48. Wood HG (1976) Trailing the propionic acid bacteria. In: Kornberg A, Horecker BL, Cornudella L, Oró J (eds) Reflections on Biochemistry, in honour of Severo Ochoa. Oxford: Pergamon Press, pp 103–115
49. Wood HG (1991) Life with CO or CO_2 and H_2 as a source of carbon and energy. FASEB J 5:156–163
50. Wood HG, Ljungdahl LG (1991) Autotrophic characters of the acetogenic bacteria. In: Shively JM, Barton LL (eds) Variations in Autotrophic Life. London: Academic Press, pp 201–250
51. Xia J, Sinclair JF, Baldwin TO, Lindahl PA (1996) Carbon monoxide dehydrogenase from *Clostridium thermoaceticum*: quaternary structure, stoichiometry of its SDS-induced dissociation, and characterization of the faster migrating form. Biochemistry 35:1965–1971

CONCLUSIONES

Si bien los caracteres moleculares como el 16/18S rRNA han demostrado el origen monofilético de todos los organismos y, por ende, la existencia de un ancestro común a todos ellos (Woese 1983), es importante reconocer los limites de este enfoque tanto en la reconstrucción de las filogenias y las características del *cenancestro*, como respecto a la poca o nula información que nos proporcionan sobre el origen de la vida mismo. No hacerlo puede llevar a conclusiones prematuras, como ocurrió con la definición original del *progenote* sugerida por Woese y Fox (1977), o con la caracterización del *cenancestro* como una célula con genoma de RNA y que supuestamente dependía heterotróficamente de nucleótidos de origen abiótico, tal como lo sugirieron Mushegian y Koonin (1996).

El descubrimiento de enzimas homólogas que catalizan reacciones bioquímicas similares en diferentes rutas anabólicas apoya la idea de que existió un reclutamiento masivo de enzimas durante el desarrollo primario de rutas metabólicas. Como ya se mencionó arriba, cuarenta años antes de que se reconociera el papel del RNA en la evolución biológica temprana Horowitz (1945) sugirió la llamada hipótesis retrógrada para explicar el origen de las rutas biosintéticas. Los resultados presentados aquí sugieren que la aplicabilidad de esta idea es bastante limitada. El análisis cladístico de las enzimas que participan en las rutas de salvamento de nucleótidos de purinas demuestra que es difícil sostener la idea de Horowitz (1945, 1965), que supone que los genes homólogos codifican enzimas que catalizan pasos sucesivos en una misma ruta biosintética. El análisis comparativo de genomas celulares completos ha demostrado que el orden de los genes no se encuentra conservado en la evolución de los procariontes (Mushegian y Koonin, 1996b; St. Jean y Charlebois, 1996; Becerra y Lazcano 1998; Delaye 1998; Islas et al., 1998).

Por otra parte, el descubrimiento de que una porción importante de los genomas bacterianos ha resultado de duplicaciones parálogas ancestrales (Fleischmann et al., 1995; Fraser et al., 1,995; Koonin et al., 1995; Labedan y Riley, 1995) es consistente con la hipótesis de que las rutas metabólicas fueron ensambladas via patchwork. Así, en muchos casos se puede afirmar que las rutas ancestrales fueron mediadas por enzimas que poseían una baja especificidad al substrato (Waley, 1969; Ycas, 1974; Jensen, 1976), que pudieron participar en rutas metabólicas que actualmente no se encuentran directamente conectadas, por ejemplo, la biosíntesis de histidina y pirimidinas. Por otro lado, la demostración del papel que la duplicación génica jugó en el ensamblaje de las rutas de salvamento de nucleótidos de purinas no solamente es consistente con la hipótesis de patchwork, sino que también permite explicar la aparente rapidez con la que pudieron haber evolucionado las rutas anabólicas durante el Arqueano temprano (Lazcano y Miller, 1994).

El reconocimiento del papel de la duplicación génica en la evolución de las rutas metabólicas no resuelve el problema del origen las enzimas que no surgieron de esta manera. En algunos casos, las enzimas ancestrales pudieron haber tenido su origen en reacciones no enzimáticas, y fueron proteínas que aceleraron un proceso espontáneo pero lento (Lazcano y Miller, 1996), como es el caso de la descarboxilación fotoquímica del ácido orótico descrita por Ferris y Joshi (1979). Todo sugiere que las rutas primitivas pudieron haber existido con un número reducido de pasos mediados por enzimas poco específicas.

La comparación de secuencias demuestra que el ensamblaje tipo *patchwork* ha jugado un papel importante en la evolución de las diferentes rutas metabólicas de fijación de carbón. Como ya lo hizo notar Pace (1997), la distribución filogenética de los diferentes tipos de metabolismos de energía y los de fijación de dióxido de carbono en la filogenia universal no obedecen a un esquema simple, lo cual sugiere que varios

eventos de transporte horizontal pudieron haber ocurrido durante su evolución. Sin embargo, es evidente que alguna forma de asimilación de carbón, que se encuentra de igual manera en los heterotrofos estrictos, estaba operando antes de la fotosíntesis clorofila-dependiente.

En contraste con la teoría quimioatotrófica basado en la pirita para el origen de la vida, que supone que el ciclo autotrófico de reducción del ácido cítrico fue la primer ruta anabólica (Wächtershäuser 1990, 1992), nosotros proponemos que los experimentos realizados por Huber y Wächtershäuser en 1997 sobre la formación de enlaces C-C en metales azufrados, pueden ser interpretados como evidencias de que la ruta Wood es la forma mas antigua de fijación de carbón. Esta interpretación alternativa es consistente con : (a) la amplia distribución filogenética del ciclo de Wood, (b) energéticamente esta ruta es mas favorable que las otras formas de fijación (Fuchs 1989); y (c) la posibilidad de que la asimilación del CO_2 fuera originalmente una ruta semi-enzimática, en donde la síntesis de acetil-CoA a partir de CO/CO_2 fue mediada por NiS/FeS junto con acetil-CoA sintasa (Huber y Wächtershäuser 1997).

Finalmente, quisiera agregar que la amplia distribución filogenética y la alta frecuencia de secuencias simples en los proteomas analizados, refleja que los procesos mutacionales (slipped-strand) que producen los segmentos de baja complejidad, pueden haber jugado un papel importante en la evolución de secuencias codificantes. Además, la presencia de segmentos de baja complejidad en proteínas altamente conservadas y muy antiguas como RNA polimerasa sigma 54, los factores de iniciación IF-2 y IF-3, las proteínas ribosomales L7, L9, L10, L12, L19, L27a, DNA polimerasa III, y en las subunidades hidrofílicas F_0 y F_1 de ATP sintetasas, sugiere que este fenómeno ha operado desde la evolución temprana de la vida. Es decir, las secuencias simples han funcionado como una fuente de variación genética (Moxon

1994), pero sobre todo como un importante mecanismo de amplificación génica que al menos en algunos casos han dado origen a regiones funcionales.

Las mutaciones que producen los segmentos de baja complejidad se presentan a todo lo largo de las proteínas. Sin embargo, su distribución (78%) hacia los extremos NH_2- y -COOH en las proteínas sugiere que son seleccionadas en contra aquellas que modifican radicalmente la estructura al perturbar dominios estructurales. El sesgo en la composición de la secuencias simples tiende a presentar altas concentraciones de aminoácidos que tienden a formar α-hélices (alanina, serina, isoleucina, leucina, lisina, ácido glutámico), mientras que el ácido aspartico, la metionina, la histidina, la arginina, la tirosina, el triptofano, y la cisteína están subrepresentados. Los aminoácidos de alto peso molecular (M,H,F,R,Y,W), son seleccionados en contra mientras que la ausencia de cisteína, la prolina y el ácido aspártico, se debe probablemente a que inducen importantes cambios estructurales. Las secuencias simples están presentes en todos los tipos de funciones metabólicas, aunque muchos de ellos se concentran en proteínas de membrana.

Referencias

Becerra A. Silva E. Velasco A. M. and Lazcano A. (1997), Polyphyletic gene losses Can Bias Backtrack Characterizations of the cenancestor, J Mol Evol 45, 115-118

Becerra A. and Lazcano A. (1998) The role of gene duplication in the evolution of purine nucleotide salvage pathways, Origin of Life and the evolution of the biosphere, 28, 539-553

Britten R J, Kohne D E (1968) Repeated Sequences in DNA. Science 161:529

Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y, (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453-1474

Bult C J, White O, Olsen G J, Zhou L, Fleischmann R D, Sutton G G, Blake J A, FitzGerald L M, Clayton R A, Gocayne J D, Kerlavage A R , Dougherty B A, Tomb, J F, Adams MD, Reich C I, Overbeek R, Kirkness E F, Weinstock K G, Merrick J M, Glodek A, Scott J L, Geoghagen N S M, Venter J C (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschi.* Science 273:1017-1140.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence, Nature 393:537-44

Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M, Huber R, Feldman RA, Short JM, Olsen GJ, Swanson RV. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus.* Nature 392:353-8

Doolittle R.F. (2000) Searching for the last common ancestor, Res, Microbiol 151(2) 85-89

Fitch W. M. (1987) The phylogeny of tRNA sequences provides evidences of ambiguity reduction in the origin of the genetic codeQuant. Biol. 52:759-767

Fleischmann 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, Spriggs T, Hedblom 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, Geoghagen N S M, Gnehm C L, McDonald L A, Small K V, Freiser C M, Smith, H O, 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 Ch., Labedan B. (1993) The nature of the last universal ancestor and the root of thetree of life, still open question. BioSystem 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, Merrick J M, Tomb J F, Dougherty B A, Bott K F, Hu P C, Lucier T S, Peterson S N, Smith H O, Hutchinson III C A, Venter J C (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270, 397-403.

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, Gwinn M, Dougherty B, Tomb JF, Fleischmann RD, Richardson D, Peterson J, Kerlavage AR, Quackenbush J, Salzberg S, Hanson M, van Vugt R, Palmer N, Adams MD, Gocayne J, Venter JC, et al. (1997), Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390:580-6

Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum KA, Sodergren E, Hardham JM, McLeod MP, Salzberg S, Peterson J, Khalak H, Richardson D, Howell JK, Chidambaram M, Utterback T, McDonald L, Artiach P, Bowman C, Cotton MD, Venter JC, et al. (1998), Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281:375-88

Goffeau et. al., Nature 387 (Suppl.) 5-105 (1997)

Gogarten, J.P. Kibak, H., Dittrich P., Taiz L., Bowman , E.J. Manolson M., Poole, J., Date T., Oshima T., Konishi L., Denda, K., and Yoshida M. (1989) Proc. Nstl. Acad. Sci. USA 86:6661-6665

Godarten J.P. (1994) J. Mol. Evol. 39:541-543

Goto S, Bono H, Ogata H, Fujibuchi W, Nishioka T, Sato K, and Kanehisa M (1996) Organizing and computing metabolic pathway data in terms of binary relations. Pacific Symp. Biocomputing 2:175-186

Hancock J M (1995) The contribution of slippage-like processes to the genome evolution. J. Mol. Evol. 41:1038-1047

Hancock J M (1996) Simple sequences in a ´minimal´ genome. Nature 14:14-15

Hayes J.M. (1994) Global methanotrophy at the Archaean-Proterozoic transition, Nobel Symposium No. 84 New York, Columbia University Press, 220-236

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R (1996), Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res 24(22):4420-49

Horowitz, N. H. (1945) Proc. Nstl. Acad. Sci. USA 31, 153-157

Horowitz N. H. (1965) In Bryson V. and Vogel H. J. (eds), Evolving genes and proteins, Academic Press, Ney York, 15-23

Islas E., Becerra A., Leguina J.I., and Lazcano A. (1998), In J. Chela Flores and Raulin (eds), Exobiology, 167-174

Iwabe N., Kuma K., Hasegawa M., Osawa S., and Miyata T. (1989) Proc. Nstl. Acad. Sci. USA, 86: 9355-9359

Jensen R.A. (1976) Annu Rev Microbiol. 30,409-425

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 3:109-36

Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A, Nagai Y, Sakai M, Ogura K, Otsuka R, Nakazawa H, Takamiya M, Ohfuku Y, Funahashi T, Tanaka T, Kudoh Y, Yamazaki J, Kushida N, Oguchi A, Aoki K, Kikuchi H. (1998), Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3 (supplement). DNA Res 5:147-55

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Venter JC, et al. (1997), The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 390:364-70

Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, Borriss R, Boursier L, Brans A, Braun M, Brignell SC, Bron S, Brouillet S, Bruschi CV, Caldwell B, Capuano V, Carter NM, Choi SK, Codani JJ, Connerton IF, Danchin A, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390:249-56

Lazcano A, Fox G E, Oró J F (1992) Life before DNA: The origin and evolution of early Archean cells. In Mortlock R P (ed) Evolution of metabolic function. Boca Raton Ann Arbor 237-295

Moxon ER, Rainey PB, Nowak MA, and Lenski RE (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. Current Biology 4(1):24-33

Moxon ER (1999) Whole-genome analisys of phatogens. In Stephen C. Stearns (ed) Evolution in Health & diseases. Oxford University Press.

Oparin A. I. (1938) The Origin of Life, Macmillan, Ney York

Pace N.R., (1997) A molecular view of microbial diversity and the biosphere. Science 276, 734-740

Pearson W R, and Lipman D I (1988) Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85:2444-2448

Quayle J.R. Ferenci T. (1978) Evolutionary aspects of autotrophy, Microbiol Rev. 42, 251-273

Saunders N J, Peden J F, Hood D W, Moxon R (1998) Simple sequence repeats in the *Helicobacter pylori* genome. Mol. Microbiol. 27:1091-1098

Schopf J. W., (1993) Microfossils of early Archaean apex chert: new evidence of the antiquity of life. Science 260, 640-646

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, Harrison D, Hoang L, Keagle P, Lumm W, Pothier B, Qiu D, Spadafora R, Vicaire R, Wang Y, Wierzbowski J, Gibson R, Jiwani N, Caruso A, Bush D, Reeve JN, et al. (1997), Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J Bacteriol 179:7135-55

Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, Koonin EV, Davis RW. (1998), Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282:754-9

Sonnhammer E L L, Durbin R, (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1-10

Tautz D, Renz M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. 12:4127-4138

Tautz D., Trick M, Dover G A (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322: 652-656

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou L, Kirkness EF, Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzegerald LM, Lee N, Adams MD, Venter JC, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388:539-47

Wächtershäuser G (1990) Evolution of the first metabolic cycles, Proc. Nstl. Acad. Sci. USA 87, 200-204

Waley S. G. (1969), Comp. Biochem. Physiol. 30, 1-7

Wootton J and Federhen S (1993) Statiscs of local complexity in amino acid sequences and sequence databases. Comp Chem. 17:149-163

Wootton J (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. Comp Chem 18:269-285
Woese C.R. and Fox G. (1977) The concept of cellular evolution. Jour. Mol. Evol. 10: 1-6

Woese C.R. (1987) Bacterial Evolution. Microbiol. Review 51:221-271

Woese C.R. (1983) In Bendall D.S. K. (de), Evolution from Molecules to Man, Cambridge University Press

Woese C.R. , Kandler O, and Wheelis M. L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA 87:4576-4579

Ycas M. (1974) J. Theoret. Biol. 44, 145-160