00387

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

Facultad de Ciencias

Evolución Temprana de los Genes más Conservados en los Tres Dominios Celulares

T E S I S

QUE PARA OBTENER EL GRADO ACADÉMICO DE

DOCTOR EN CIENCIAS

P R E S E N T A

Biol. Luis José Delaye Arredondo

Director de Tesis: Dr. Antonio Lazcano-Araujo Reyes

POSGRADO EN CIENCIAS BIOLÓGICAS
COORDINACIÓN

MÉXICO, D.F. MAYO, 2005

VNIVERSIDAD NACIONAL
AVTONOMA DE
MEXICO

POSGRADO EN CIENCIAS BIOLÓGICAS
COORDINACIÓN

Ing. Leopoldo Silva Gutiérrez
Director General de Administración Escolar, UNAM
P r e s e n t e

Por medio de la presente me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día 14 de marzo del 2005, se acordó poner a su consideración el siguiente jurado para el examen de DOCTOR EN CIENCIAS del alumno **DELAYE ARREDONDO LUIS JOSÉ** con número de cuenta **93679668** y número de expediente **3992010**, con la tesis titulada: **"Evolución Temprana de los Genes más Conservados en los Tres Dominios Celulares"**, bajo la dirección del **Dr. Antonio Eusebio Lazcano-Araujo Reyes.**

Presidente:	Dr. Germinal Cocho Gil
Vocal:	Dra. Alicia Negrón Mendoza
Vocal:	Dr. Víctor Manuel Valdés López
Vocal:	Dr. Lorenzo Patrick Segovia Forcella
Secretario:	Dr. Antonio Eusebio Lazcano-Araujo Reyes
Suplente:	Dra. Valeria Francisca E.L.M. Souza Saldivar
Suplente:	Dr. Arturo Carlos II Becerra Bracho

Sin otro particular, quedo de usted.

A t e n t a m e n t e
"POR MI RAZA HABLARA EL ESPIRITU"
Cd. Universitaria, D.F., a 28 de abril del 2005.

Dr. Juan José Morrone Lupi
Coordinador del Programa

c.c.p. Expediente del interesado

Este trabajo se llevó a cabo en el Laboratorio de Microbiología del Departamento de Biología Evolutiva de la Facultad de Ciencias de la Universidad Nacional Autónoma de México, bajo la tutoría del Dr. Antonio E. Lazcano-Araujo Reyes.

La investigación fue parcialmente financiada por UNAM-DGAPA proyecto PAPIIT IN 111003-3. Además se contó con el apoyo de una beca para estudios de Doctorado CONACyT (registro 138482) y beca complemento DGEP.

Parte de los análisis presentados en este trabajo fueron realizados gracias al apoyo proporcionado por el Departamento de Súper Cómputo DGSCA de la UNAM. Parte de este trabajo se realizó durante una estancia de investigación en el Laboratorio del Dr. Peter Gogarten (verano del 2000) con una beca NASA-PBI y otra estancia de investigación (Septiembre-Diciembre, 2001) en el Protein Design Group CNB, liderado por el Dr. Alfonso Valencia.

El Comité Tutoral estuvo conformado por los doctores Alicia Negrón Mendoza, Valeria Souza Saldivar y Antonio Lazcano-Araujo Reyes.

Los doctores Arturo Becerra Bracho, Alicia Negrón Mendoza, Valeria Souza Saldivar, Germinal Cocho Gil, Lorenzo Segovia Forcella, Victor Valdés López y Antonio Lazcano Araujo revisaron críticamente este documento y conformaron el Jurado de Examen de Grado.

..Respiré el aire de los tamarindos. Vibraba la noche. llena de hojas he insectos. Los grillos vivaqueaban entre las hierbas altas. Alcé la cara: arriba también habían establecido campamento las estrellas. Pensé que el universo era un vasto sistema de señales, una conversación entre seres inmensos. Mis actos, el serrucho del grillo, el parpadeo de la estrella, no eran sino pausas y sílabas, frases dispersas de aquel diálogo, ¿Cuál sería esa palabra de la cual yo era una sílaba? ¿Quién dice esa palabra y a quién se la dice?..

El Ramo Azul

Octavio Paz

Agradecimientos

La realización de un doctorado es un proyecto que requiere de un esfuerzo considerable, que además involucra de diversas formas a gran cantidad de gente. No hubiera podido realizar este trabajo si no fuera por todas las personas que han creído en mi y me a apoyado a lo largo del camino.

A mis Padres por todo su apoyo y su educación. En particular a mi padre por haberme cultivado la curiosidad científica desde muy pequeño.. no tengo palabras para explicar la sensación que causaba en mi sentarme a platicar con él acerca de la naturaleza de aspectos en principio tan triviales como la condensación del vapor alrededor de un vaso lleno de agua fresca. A mi madre por haberme inculcado la creencia de que vale la pena y es posible luchar por alcanzar los sueños.. creo que para casi cualquier aventura humana existen tantas complicaciones que tal vez la única forma de afrontarlas es creer profundamente que es posible superarlas. Por supuesto también quiero agradecerles, haber sido y seguir siendo un apoyo tan importante en los aspectos pragmáticos de la vida. A ambos, gracias..

A mis hermanos por mostrarme que otros caminos también son posibles. A Irene (Pipis) por su sincera amistad y a Ramón por todo su apoyo logístico (siempre, pero en especial durante las últimas fechas), su filosofía, su jamón serrano y sus deliciosas paellas y bacanales de fin de semana. Gracias también a los dos por estar al lado de mis padres. Y por supuesto a Beto por su amistad.

A Beatriz, por creer en mi y mis locuras y seguir a mi lado, y todo el apoyo que me has dado en la infinidad de detalles diarios.

A mis Tíos Armando y Estela, quienes me alojaron en su casa los primeros días de mi licenciatura en la UNAM, haberme contagiado del espíritu Universitario de nuestra Máxima Casa de Estudios, y su apoyo decidido a mi y mi familia en todo lo que han podido. También por adoptarnos como huéspedes felices en su hermosa finca de Atlihuayan. A mis primos Julia y Diego por compartir estos años de amistad.

A mis Tíos Vicente y Angélica también gracias por acompañarme en estos años del doctorado y por su apoyo en todo lo que he necesitado y los fines de semana en Tepoztlan.

A la UNAM como institución, por ser un espacio en donde se cultiva el conocimiento, y las vidas y sueños de tantas personas se tocan y entrelazan.

Recuerdo bien la impresión que me dio saber que en la UNAM había un investigador que trabajaba en origen de la vida y que además había sido alumno de Alexander I. Oparin. Habiendo nacido Oparin a finales de

1800, supuse que su alumno mexicano sería una persona seria y de avanzada edad (siguiendo el estereotipo más conocido de los científicos naturalistas tales como Darwin), sin embargo mi suposición no pudo estar más equivocada. Quiero agradecer a Herr Professor Antonio Lazcano (*Magister Maximus*), por haber creído en mi desde la licenciatura, por todo el apoyo que me ha brindado y por su amistad a lo largo de los años del doctorado. Me siento feliz y honrado de participar de alguna forma de esta genealogía intelectual (aunque sea una pequeña rama lateral). Por todo, gracias.

A los miembros del laboratorio Sara, Ana María, Ervin y Ulises por su amistad y compañía diaria en el laboratorio, por aguantar mis neurosis y por compartir un proyecto en común, gracias.

A mis amigos Jorge Schondube (el Chon), David Valenzuela (Davison), Alexander De Luna (el Pariente), Arturo Becerra (Don A), Luis Bernardo (LB) y Hugo (Hugol), por tantas aventuras compartidas y por las que están por venir..

A mis amigas, Xitlali Aguirre, Andrea González y Tania Hernández, por su amistad, en la biología en la poesía y en las noches de fiesta por la ciudad.. gracias.

A los talleriues, Ricardo (el cuervo), Chucho (Bubi), German Bonilla, Diego Cortez, y Mariana Benítez por su amistad. A Claudia Sierra y Daniela Sosa por haber confiado en mi como tutor. A Pedro Miramontes (por haberme puesto a pensar sobre la selección natural) y Germinal Cocho. A Peter Gogarten por su amistad y por mostrarme que la heterodoxia puede ser más divertida que la ortodoxia. A mis profesores a todo lo largo de la carrera y el doctorado y a mis héroes intelectuales, algunos de los cuales tengo el placer de conocer en persona(un profesor.. nunca logra saber donde termina su influencia). A mis suegros por su ayuda en tiempos complicados. A Adrián Reyes y Lina Riego por los cursos compartidos y porque se repitan. Al *cenancestro* y al *progenote*, por ser el tema central de mi investigación. A todos ellos, gracias..

24 Abril, 2005
Coyoacán, México DF
Luis José Delaye Arredondo

CONTENIDO

ÍNDICE DE FIGURAS Y TABLAS

ÍNDICE DE TRABAJOS PUBLICADOS O EN PREPARACIÓN

Trabajos publicados

- Delaye, L., and Lazcano, A. (2000) **RNA-binding peptides as molecular fossils** *In* J.Chela-Flores, G. Lemerchand, and J. Oró (eds.) *Astrobiology: Origins from the Big-Bang to Civilisation Proceedings of the First Ibero-American School of Astrobiology* (Kluwer Academic Publishers), pp. 285-288.
- Delaye, L., Vazquez, H., and Lazcano, A.(2001) **The cenancestor and its contemporary biological relics: the case of nucleic acids polymerases** *In* Julian Chela-Flores J., Owen, T. and Raulin, F. (eds.). *First Steps in the Origin of Life in the Universe* (Kluwier Academic Publishers), pp. 223-230.
- Delaye, L., Becerra, A., And Lazcano, A. (2004) **The nature of the last common ancestor**. *In* Lluis Ribas de Pouplana (ed.) *The Genetic Code and the Origin of Life* (Landes Bioscience, and Kluwier Academic), pp. 34-47.
- Luis Delaye and Antonio Lazcano. (2005) **Prebiological evolution and the physics of the origin of life**. *Physics of Life Reviews*, **2**, pp. 47-64.

Trabajos enviados o en prensa

- Lazcano, A., Becerra, A. and Delaye, L. (2004) **On the early evolution of sensory responses: when did life first begin to perceive its surroundings?** *In* Margulis, L. and Asikainen, C. A. (eds) *Human Brain in the Context of Natural History: 3000 million years of evolution of sensory systems* MIT Press, Boston (en prensa).
- Delaye, L., Becerra, A., and Lazcano, A. (2005) **The Last Common Ancestor: what's in a name?** *Origin of Life and Evolution of the Biosphere* (en prensa).

Trabajos en preparación

- Delaye, L., Becerra, A., and Lazcano, A. **On the early evolution of polymerase function and the nature of the genome of the cenancestor** (en preparación).
- Delaye, L., Becerra, A., and Lazcano, A. **On the early evolution of polymerase function and the nature of the genome of the cenancestor** (en preparación).

i. Resumen

El estudio del origen y la evolución temprana de la vida es en esencia un problema histórico. Se requiere de múltiples aproximaciones para tratar de reconstruir los eventos que comenzaron con la sopa prebiótica y que finalmente desembocaron en los primeros seres vivos. En esta tesis se aborda el estudio de tres aspectos fundamentales de la evolución de las primeras células. Primero se pretende una reconstrucción del complemento genético del ultimo ancestro común, mediante el análisis de 20 genomas celulares de organismos no parásitos utilizando una metodología de búsquedas de BLAST de un solo sentido. Los genes así identificados están relacionados principalmente con el metabolismo del RNA. Dicho hallazgo sugiere que previo a los sistemas celulares basados en DNA/RNA/proteínas, las células estaban basadas en RNA/proteínas. En segundo lugar se intenta reconstruir la evolución temprana de la enzima replicativa central (la DNA polimerasa). A pesar de que la replicación del DNA es un proceso central en todos los seres vivos, la enzima que se encarga de replicar el material genético es de origen polifilético entre los linajes Bacterianos y Archaea/Eucariontes. En este trabajo se sugiere que dicho patrón de conservación se debe a una sustitución no ortóloga en la base del árbol universal. Además se sugiere que el dominio catalítico (*palm domain*) de las DNA polimerasas I, II, RNA polimerasas virales y reverso transcriptasas proviene de una etapa previa al mundo del DNA. En tercer lugar, en este trabajo se sugiere que los dominios de unión a RNA son algunas de las proteínas más antiguas que podemos reconocer. De acuerdo a ello, se construyó una base de datos de 68 dominios de unión a RNA, 35 de los cuales están universalmente conservados y presumiblemente provienen de una etapa en donde las células poseían genomas de RNA. El patrón de duplicación de algunas de estos dominios sugiere que la duplicación génica y la evolución por *patchwork* han jugado un papel central en la generación de nuevas proteínas y nuevas funciones desde los albores de la vida en la Tierra.

ii. Abstract

The study of the origin and early evolution of life is an historic problem. In order to understand the events that began in the prebiotic soup and lead to the first living cells it is necessary to use several different approximations. In this thesis we ask three important aspects of the early evolution of ancient cells. First, we make an inference of the gene complement of the Last Common Ancestor (LCA) from 20 extant cellular genomes from non-parasitic organisms using one-way BLAST searches. The set of highly conserved genes identified among these genomes are related mainly with RNA metabolism. This suggest that prior DNA/RNA/protein cells there where RNA/protein cellular systems. On the second place, we study the early evolution of the replicative DNA polymerase. DNA replication is certainly one of the central process in every extant cells, and must have evolved very early. Nevertheless, the main replicative enzyme is not universally conserved. Bacteria and Archaea/Eucarya uses different enzymes. We suggest that this pattern is due to a non-homologous gene displacement in the Bacterial clade. We also suggest that the catalytic *palm domain* found in DNA polymerases class I and II, RNA polymerase and Reverse Transcriptase evolved prior DNA genomes in the RNA/protein world. Finally, in this work we suggest that RNA-binding protein domains are among the oldest polypeptides we can recognize. Accordingly, we analyzed the phylogenetic distribution of 68 different RNA-binding protein domains. Among them, 35 RNA-binding domains are universally conserved and likely evolved in the RNA/protein world. Evidences of *patchwork* evolution among them suggest that gene duplication and fusion played an important role during the early evolution of life.

I. INTRODUCCIÓN

I.I. Sobre el estudio del origen y la evolución temprana de la vida

Un problema histórico

El origen y la evolución temprana de la vida en la Tierra es en esencia un problema histórico. Es decir, que para poder comprender cómo surgió la vida y cual fue la naturaleza de las primeras células que evolucionaron en el planeta, requerimos de una narrativa histórica que explique causalmente los hechos que ocurrieron en el pasado y que han dado forma al presente biológico. Para ello, es necesario basarse por un lado, en la comparación de datos ricos y diversos que registren las consecuencias de los acontecimientos de tiempos ancestrales, a partir de los cuales podemos hacer reconstrucciones del pasado. Y por otro lado, se deben de realizar experimentos que nos permitan evaluar la plausibilidad de que determinados sucesos pudiesen haber ocurrido de acuerdo a hipótesis específicas. Si bien, los eventos particulares que finalmente condujeron a la aparición de la vida, son con seguridad históricamente únicos, todos ellos deben de poderse explicar a partir de principios físicos, químicos y biológicos.

En este sentido, la teoría químico-física del origen de la vida sugerida independientemente por Alexander I. Oparin (1924) y J.B.S. Haldane (1929) cuya contribución central consiste en proponer que la vida se originó a partir de la evolución química y gradual de compuestos orgánicos que se habían sintetizado y acumulado de forma abiótica en la Tierra primitiva, proporciona la hipótesis sobre la cual descansan los estudios sobre el origen y la evolución temprana de la vida en la Tierra.

Desde la química prebiótica

Bajo este marco conceptual, existen al menos tres formas de aproximar el estudio del origen y la evolución temprana de la vida (modificado de Lazcano & Miller, 1999) (Figura 1). Una primera aproximación consiste en tratar de reconstruir la química prebiótica mediante experimentos que pretenden recrear las condiciones de la Tierra primitiva. Debido a que estos experimentos abordan preguntas específicas acerca de la química de compuestos orgánicos sintetizados de manera abiótica, es posible hacer una narrativa histórica de la evolución química que en determinado momento condujo a la aparición de la vida en la Tierra y a la consecuente evolución biológica. El experimento clásico que abrió esta área de experimentación fue realizado por Stanley Miller en 1953 el cual consistió en poner a reaccionar en un matras con descargas eléctricas como fuente de energía, una mezcla de gases (NH_3, H_2O, CH_4, H_2), que, de acuerdo a un modelo de evolución de la atmósfera primitiva propuesto por Harold C. Urey eran abundantes en la atmósfera primigenia. El resultado espectacular de dicho experimento fue la obtención de una serie de aminoácidos, algunos de los cuales se encuentran presentes en los seres vivos actuales, sugiriendo con ello como se pudieron haber sintetizado algunas de las moléculas básicas a partir de las cuales evolucionaron las primeras células. Como se puede apreciar, el

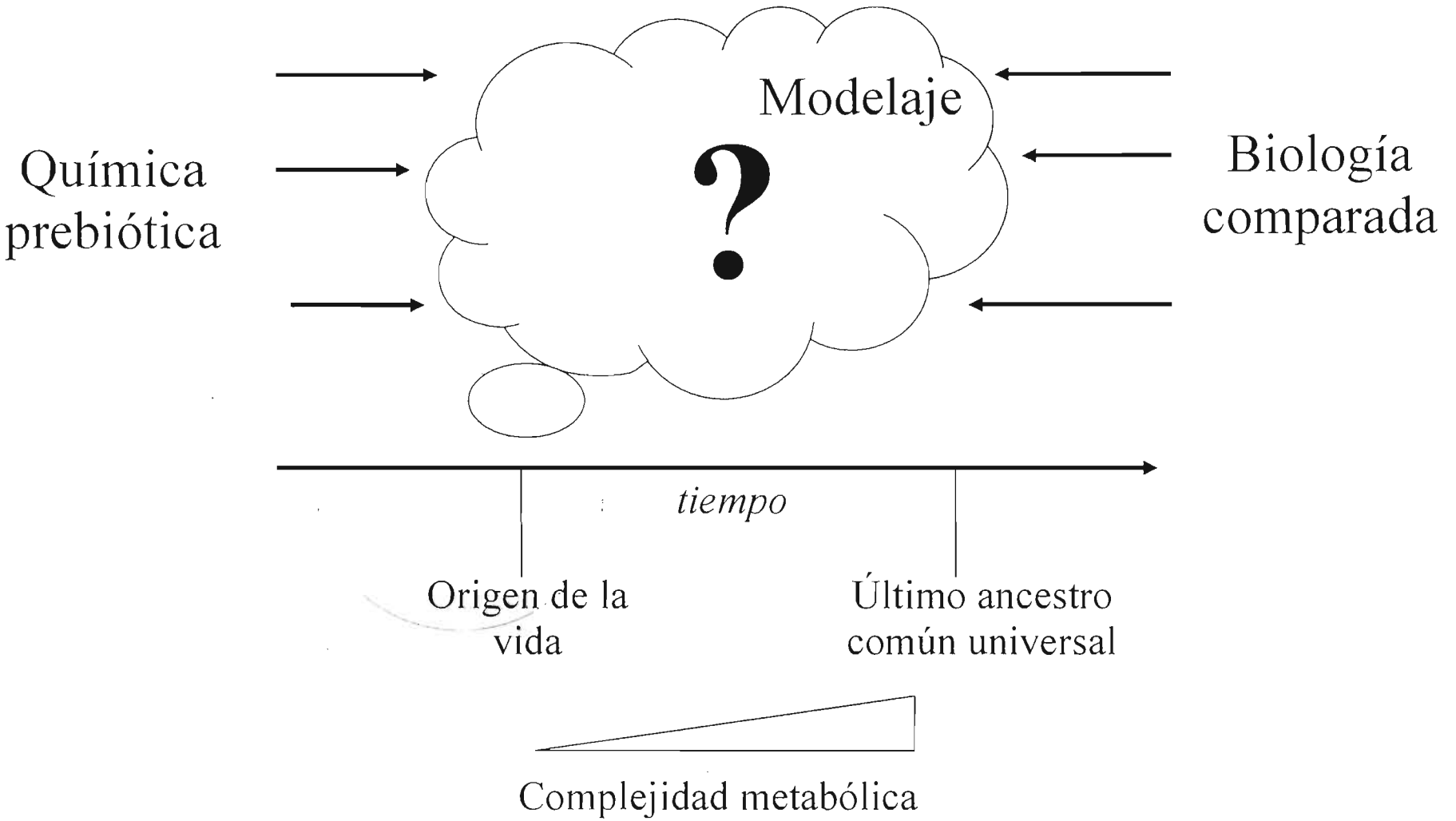


Figura 1. Sobre el estudio del origen y la evolución temprana de la vida. El origen y la evolución temprana de la vida se pueden estudiar a partir de tres aproximaciones distintas, desde la química prebiótica, modelaje *in silico*, *in vitro* o *in papiro* o desde la biología comparada (Modificado de Lazcano & Miller, 1999).

experimento de Miller fue un apoyo empírico sin precedentes a la teoría químico-física Oparin-Haldane del origen de la vida.

Desde la biología actual

La otra aproximación que existe para estudiar el origen, pero sobre todo la evolución temprana de la vida en la Tierra, consiste en tratar de reconstruir la evolución de las primeras células a partir del análisis comparativo de las características de los seres vivos actuales. Esto es, de forma análoga a como un geólogo reconstruye los eventos del pasado a partir del estudio de una columna estratigráfica, es posible utilizar a la biología actual como nuestro registro histórico a partir del cual podemos hacer inferencias del pasado biológico. Los primeros en sugerir que la comparación de secuencias homólogas de proteínas (y por ende de ácidos nucleicos), pueden ser utilizados como documentos históricos a partir de los cuales podemos "leer" los procesos y patrones evolutivos que han precedido y dado forma a los organismos actuales fueron Zuckerkandl & Pauling (1965), iniciando con ello el estudio de la evolución a nivel molecular. Si bien, dichas extrapolaciones reduccionistas hacia el pasado, no nos permitan indagar directamente en el origen de la vida, debido a que es muy probable que utilizando este método no podamos retroceder de una época en donde ya había síntesis de proteínas mediada por ribosomas, como veremos más adelante, esta metodología ha redefinido de forma muy importante el estudio de la evolución temprana de la vida. Esta segunda aproximación, la cual pretende reconstruir el pasado a partir del estudio de las secuencias actuales, es la que se aborda en este trabajo para tratar de entender algunos aspectos de la evolución temprana de la vida en la Tierra.

I.II. La revolución woesiana y la naturaleza del último ancestro común universal

La filogenia universal

A partir de la comparación de secuencias del 16/18 SSU rRNA provenientes de diversos organismos tanto procariontes como eucariontes, Woese y Fox (1977) realizaron probablemente uno de los descubrimientos más importantes que se han hecho en los últimos 50 años sobre la historia filogenética de la vida en la Tierra. Encontraron que de acuerdo a esta molécula, los seres vivos actuales se dividen en tres grandes linajes denominados como Archaeas, Bacterias (ambos procariontes) y el Nucleocitoplasma Eucarionte. Tres líneas principales de ancestría-descendencia para la biota y no dos grandes grupos (procariontes y eucariontes) como había sugerido previamente Chatton (1938) basándose en criterios citológicos. Por muy sutil que nos pueda parecer, dicho descubrimiento revolucionó la manera de entender la biología en varios aspectos. En primer lugar, demostró que la mayor diversidad de los seres vivos se encuentra a nivel microbiano, esto es, dos de los tres grandes linajes (Bacterias y Archaeas) están compuestos exclusivamente por organismos procariontes. Por el otro lado, una buena proporción del linaje eucarionte está compuesto por organismos unicelulares muy divergentes, a las plantas y los animales. En segundo lugar, la filogenia de 16/18 SSU rRNA sirvió como base

a partir de la cual construir una clasificación natural universal (Woese, et al. 1990). Esto es, una clasificación que refleje las relaciones de ancestría y descendencia (parentesco filogenético) entre todas las especies.

El último ancestro común universal

Con la filogenia universal propuesta por Woese y Fox (1977) viene implícita la hipótesis de la monofilia de la biota, y por lo tanto la existencia de una especie ancestral a partir de la cual los seres vivos actuales descendemos. Esta entidad biológica hipotética cuya existencia ya había sugerido Darwin (1859), ha sido definida de diversas formas a lo largo del tiempo. Inicialmente Woese y Fox denominaron *progenote* a la entidad biológica que se encuentra en la base del árbol universal y sugirieron que en dicha entidad el fenotipo aun no se encontraba diferenciado del genotipo (Woese & Fox, 1977). Posteriormente, Fitch y Upper (1987) acuñaron la palabra *cenancestro* que literalmente quiere decir "último ancestro común" y que no necesariamente implica una entidad con características primitivas como aquellas sugeridas para el *progenote*.

En principio y desde el punto de vista metodológico, es posible describir las características biológicas del último ancestro común como el conjunto de aquellas características comunes y homólogas a todos los seres vivos y que se han heredado verticalmente desde el ancestro a sus descendientes, más aquellas características que estuvieron presentes en el ancestro universal pero que se perdieron en uno o más linajes (y que por lo tanto no seremos capaces de reconstruir), menos aquellas características homólogas que son comunes a todos los seres vivos debido a que han sido heredados horizontalmente entre linajes. En este sentido, el esquema filogenético propuesto por Woese y Fox (1977) nos permite definir al genoma del último ancestro común como aquel conjunto de genes homólogos a los tres linajes celulares, que han sido heredados verticalmente desde el ancestro universal (Figura 2). Si bien esta metodología para reconstruir al último ancestro común no es infalible ya que su precisión dependerá de una serie de factores tales como el sesgo en el muestreo de especies con los cuales hagamos la reconstrucción, las pérdidas polifiléticas secundarias tanto de genes individuales como de rutas metabólicas completas, las sustituciones no ortólogas de genes, y de la intensidad del fenómeno de transferencia horizontal entre linajes (Becerra, et al. 1997), es hasta el momento, la mejor metodología de la cual disponemos para poder reconstruir la naturaleza del ancestro universal, aunque sea de forma aproximada. Utilizando esta metodología y basándose en los datos de secuencias disponibles en aquel momento, Lazcano et al. (1992) sugirió que el último ancestro común (LCA, por sus siglas en inglés) era una entidad que en complejidad biológica no diferia significativamente a la de los procariontes actuales (Tabla 1).

La raíz del árbol universal

La localización de la raíz del árbol universal está intimamente ligado a la naturaleza del último ancestro común. Si bien no es posible conocer directamente la raíz del árbol universal de rRNA debido a que por definición no existe organismo que pueda ser usado como grupo externo, es posible conocer la raíz de una filogenia universal de un grupo de moléculas ortólogas si utilizamos a sus parálogos como grupo externo.

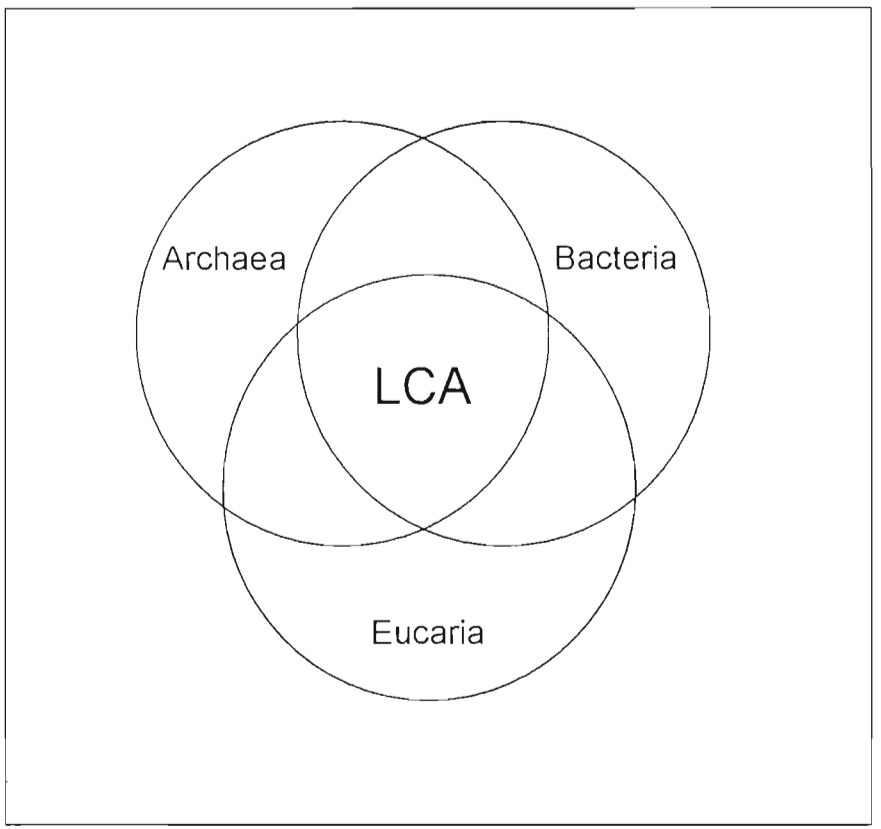


Figura 2. La reconstrucción del último ancestro común. En principio, las características del último ancestro común (LCA por sus siglas en inglés) pueden ser inferidas a partir de las características homólogas entre Bacterias, Archaeas y Eucariontes.

(i) Caracteres involucrados en replicación y biosíntesis de proteínas	
DNA polimerasa B	Factor de elongación 1α/Tu
Girasa B	Factor de elongación G/2
DNA topoisomerasa II	Isoleucil-tRNA sintetasa
RNA polimerasa	Ribonucleasa P
Polinucleótido fosforilasa	Proteínas ribosomales S9, S10, S17, S15, L2, L3, L6, L10, L11, L22 y L23
(ii) Caracteres involucrados en la generación de energía y en rutas biosintéticas	
ATPasa Tipo F subunidad α	Arginosuccinato sintetasa
ATPasa Tipo F subunidad β	Aspartato aminotransferasa
Carbamoil fosfato sintetasa	Citrato sintetasa
Glucosa 6 fosfato deshidrogenasa	Enolasa
Glutamoato deshidrogenasa II	Glutamina sintetasa
Malato deshidrogenasa	Fosfoglicerato cinasa
Piruvato: ferredoxin oxidoreductasa	Porfobilinogeno sintetasa
Histidinol fosfato aminotransferasa	Genes de biosíntesis de purinas
Genes de la biosíntesis del triptófano	Genes de biosíntesis de aminoácidos de cadena ramificada
(iii) Caracteres involucrados en respuesta ambiental y señalización celular	
cAMP	
Polipéptidos tipo Insulina	
Proteína "Heat shock" 70	
Mn/Fe superóxido dismutasa	
Fotoliasas	

Tabla 1. Caracteres homólogos a los tres dominios celulares. Basados en los datos disponibles en 1992, Lazcano et al (1992).

Los primeros en utilizar esta metodología fueron Iwabe et al. (1989) usando los factores de elongación (EF-G, EF-Tu) y Gogarten et al. (1989) empleando las subunidades α y β de las ATP sintetasas tipo F. Utilizando distintas técnicas de reconstrucción filogenética, ambos grupos llegaron independientemente a la conclusión de que la raíz universal se encuentra localizada en la rama de las bacterias. Este resultado no solamente implica que la naturaleza del último ancestro común tendrá características similares a las de las bacterias actuales, sino que también sugiere que una porción importante del nucleocitoplasma eucarionte ha sido heredado a partir de las Archaeas.

Genoma mínimo

Con la secuenciación de los primeros genomas fue posible abordar nuevamente la pregunta de la naturaleza del último ancestro común utilizando una base de datos que en principio puede ser cualitativamente diferente. El primer trabajo que utilizó la comparación de genomas para reconstruir estados ancestrales fue realizado por Mushegian y Koonin (1996) al comparar los genomas de *Haemophilus influenzae* y *Micoplasma genitalium*. En primer lugar, sugirieron que el conjunto de genes compartidos entre ambos genomas representa el número mínimo de genes necesarios para mantener una célula viva (el concepto del genoma mínimo) (Figura 3). Esta conclusión es solo parcialmente correcta ya que el conjunto de genes que representan el genoma mínimo es relativo al sistema parásito hospedero que se estudie y por lo tanto a la historia evolutiva de los organismos a partir de los cuales se hizo la comparación genómica. En segundo lugar, sugirieron que este conjunto de genes podía ser representativo del genoma del último ancestro común. Como en el conjunto de genes homólogos entre las dos especies faltaban una serie de proteínas clave en la síntesis del DNA llegaron a la conclusión de que el último ancestro común poseía un genoma de RNA. Esta conclusión fue prematura a todas luces debido principalmente a la naturaleza parásita de los organismos utilizados en la comparación los cuales han perdido independientemente una serie de genes y habilidades metabólicas (Becerra et al. 1997).

La naturaleza química del genoma del LCA y la evolución de las DNA polimerasas

Posteriormente Leipe et al. (1999) basándose en un análisis más extenso que incluía genomas de los tres linajes celulares, sugirió que el último ancestro común poseía un sistema genético similar al de un retrovirus, en donde la información genética codificada en el RNA se retrotranscribía a DNA mediante una reverso transcriptasa. Dicha conclusión estuvo basada por un lado, en el hecho de que la enzima replicativa central (DNA polimerasa) no esta conservada entre Archaeas, Bacterias y Eucariontes (básicamente, las bacterias utilizan una enzima perteneciente a la familia de las DNA polimerasa III, Euriarchaeas utilizan la polimerasa tipo D, y Crenarchaeas y Eucariontes utilizan la DNA polimerasa tipo II como enzima replicativa). Por el otro lado, dicha conclusión también se basa en el hecho de que existen enzimas accesorias relacionadas a la replicación de DNA claramente conservadas, tales como la "*clamp-loader* ATPase" lo que sugiere que el último ancestro común debió de haber tenido DNA de cualquier forma. Bajo este esquema, la replicación del

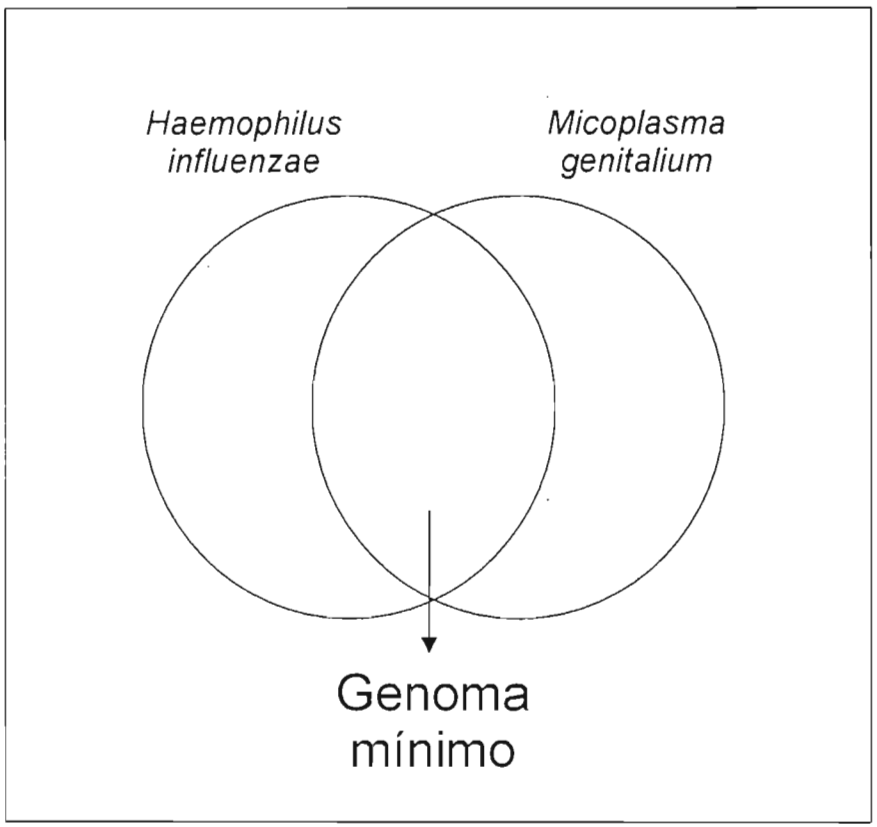


Figura 3. El genoma mínimo. Definido como el menor número de genes necesarios para mantener una célula viva, fue sugerido inicialmente por el grupo de genes homólogos entre los genomas de *H. infuelzae* y *M. Genitalium* (Mushegian et al, 1996).

DNA como la conocemos actualmente, tendría que haber sido inventada al menos dos veces durante la historia evolutiva, una en el linaje bacteriano y otra en el linaje archaea-eucarionte.

En principio, debido a que todas las células actuales poseen un genoma de DNA, el escenario más parsimonioso es un ancestro que igualmente poseía un genoma de DNA. Por ello, no deja de sorprender que la enzima replicativa central no se encuentre universalmente conservada, a diferencia de lo que sucede con la maquinaria de transcripción y traducción.

Transferencia génica horizontal

Como se mencionó anteriormente, la revolución woesiana prometía hacer realidad el sueño iniciado por Darwin en 1859. Esto es, poder representar con un solo árbol evolutivo las relaciones de ancestría-descendencia entre todas las especies actuales. Dicho árbol evolutivo estaría dominado por la herencia vertical y las especies se clasificarían de forma jerárquica de acuerdo al grado de propincuidad entre ellas. La molécula de 16S rRNA, por sus características de universalidad, tasa de evolución lenta y conservación de función, parecía ser el marcador ideal para trazar la evolución a nivel organísmico y construir el árbol universal de la vida. Sin embargo, cuando se obtuvieron las primeras secuencias completas de genomas, se pensó que las filogenias de la mayoría de los genes reflejarían el esquema evolutivo propuesto por Woese y Fox (1977), con la salvedad de un porcentaje moderado de genes que no reflejaría dicho esquema evolutivo debido a los problemas que existen asociados a las reconstrucciones filogenéticas y a unos pocos que hubieran sido transferidos horizontalmente entre los linajes. Sin embargo, los primeros análisis filogenéticos derivados de los proyectos de secuenciación de genomas parecieron contar una historia distinta. Las filogenias provenientes de distintos genes del mismo conjunto de organismos eran incongruentes entre sí, sugiriendo que la transferencia horizontal de genes entre diferentes linajes organísmicos parecía ser mucho más común de lo que hasta entonces se había creído. La situación llegó a ser tal que incluso se sugirió que no era posible describir adecuadamente el árbol universal de la vida bajo un esquema en donde la herencia vertical fuese el eje hereditario que reflejara las relaciones entre las especies, sobre todo entre los microorganismos (los linajes más comunes y antiguos de la biota). Por lo tanto, la idea de la existencia de linajes organísmicos definidos cada uno por los linajes de genes que evolucionan "juntos" en un mismo genoma y que se pueden representar con un solo marcador molecular pareció desvanecerse, y en su lugar se sugirió un esquema en donde la transferencia génica horizontal dominaba la estructura del árbol universal a gran escala (Doolittle, 1999) (Figura 4), implicando que la reconstrucciones del último ancestro común a partir de la comparación de genomas es un ejercicio fútil.

Dicha situación llevó a varios autores a sugerir una naturaleza distinta para el último ancestro común. Tal vez la más elaborada de las propuestas fue presentada por Woese (1998) quien sugirió que el último ancestro común no era una entidad discreta, sino una comunidad de células que evolucionaron como una unidad

biológica. Más específicamente, el ancestro universal estaba constituido por una población de *progenotes* con sistemas de procesamiento de información muy poco precisos en donde la dinámica evolutiva estaba regida principalmente por una alta tasa de mutación y por un elevado nivel de transferencia horizontal. De acuerdo a Woese (1998) conforme evolucionaron estructuras biológicas más complejas y precisas, tanto la tasa de mutación como la cantidad de transferencia horizontal disminuyó y la dinámica evolutiva comenzó a parecerse a la de las células actuales.

Fenogramas de genomas completos

Sin embargo, una serie de fenogramas realizados utilizando la información codificada en el genoma completo, en particular estudiando el número de familias conservadas de genes entre genomas (Tekaia et al. 1999; Snel, et al. 1999; Fitz-Gibbon, et al. 1999) (Figura 5), así como una filogenia universal derivada de concatenar una serie de genes universalmente conservados (Brown, et al. 1999), rescatan el esquema evolutivo de los tres grandes grupos filogenéticos (Archaea, Bacteria y Eucarya). Dichos análisis sugieren la existencia de los tres grandes linajes filogenéticos, y parecen indicar que la transferencia horizontal entre linajes no ha sido tal como para borrar su existencia. Análisis más detallados sugieren que de hecho la cantidad de transferencia horizontal entre procariontes no es tan elevada como se había sugerido (Daubin, et al 2003) y que las incongruencias entre árboles filogenéticos generados a partir de distintos genes del mismo grupo de organismos, se resuelven cuando los genes se concatenan (Rokas et al. 2003). Ello sugiere que el proceso de transferencia horizontal no ha sido tal, como para borrar del todo el pasado biológico. La cuestión no está en absoluto zanjada y son necesarios más estudios cuidadosos acerca del papel que ha tenido la transferencia horizontal de genes en la historia evolutiva.

I.III. La teoría del mundo del RNA

La evolución del genotipo y el fenotipo

Describir la naturaleza del último ancestro común nos ayuda a entender una etapa central de la evolución temprana de la vida en la Tierra. Sin embargo, para poder comprender etapas aun más tempranas de la evolución de la vida, necesitamos de una teoría que explique la evolución de las moléculas responsables de la herencia y el fenotipo. En la actualidad, todas las células utilizan al DNA como molécula hereditaria y a las proteínas como las moléculas que dan cuerpo al fenotipo, las cuales se encuentran codificadas en el DNA y que a través de sus habilidades metabólicas, en última instancia, se encargan de la replicación del DNA. En el centro de esta asociación se encuentran las moléculas de RNA, codificadas también en el DNA, pero que una vez expresadas, tienen un papel central en la síntesis de proteínas. ¿Cómo evolucionó esta asociación molecular entre genotipo y fenotipo? ¿Cuál fue el orden histórico en el que las distintas moléculas aparecieron y cuales pudieron haber sido las causas de esta evolución? La teoría del gen desnudo, propuesta inicialmente por Muller, y elaborada más adelante por Haldane, sugiere que la vida se originó a partir de la aparición de una molécula (gen) capaz de autor replicarse y por lo tanto, capaz de heredar esta propiedad a sus

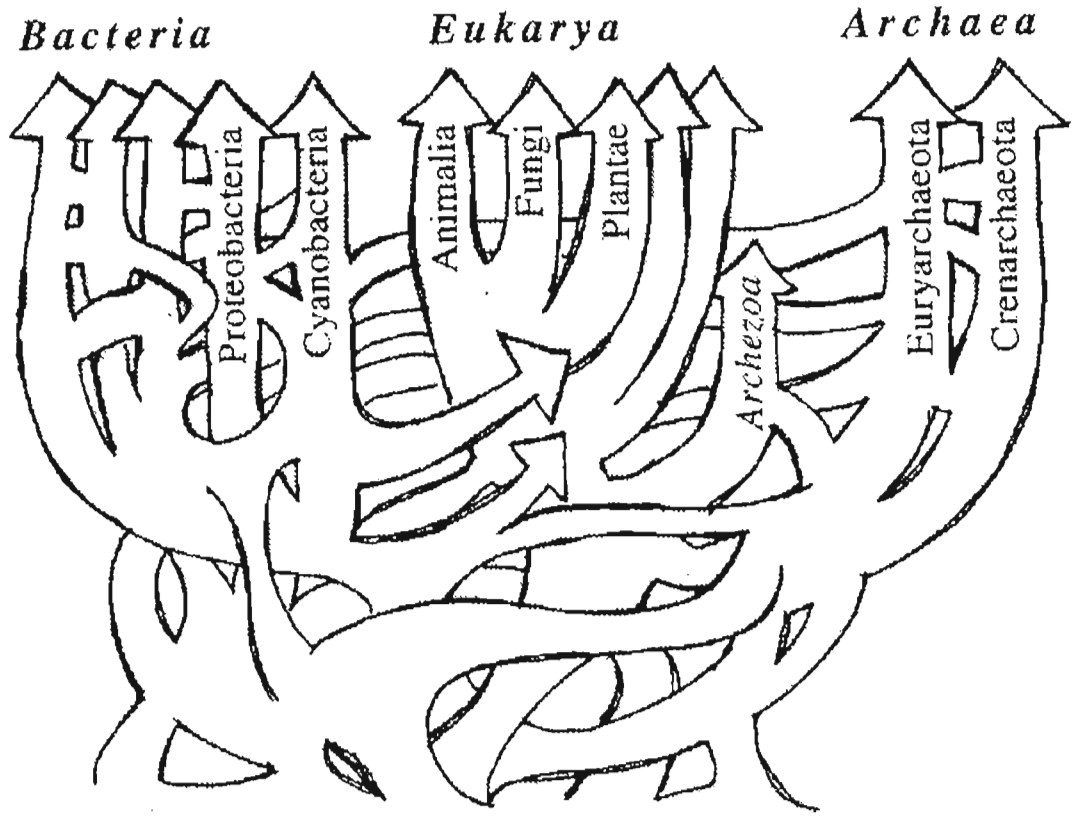


Figura 4. El papel de la transferencia horizontal de genes en el árbol universal. De acuerdo a Doolittle (1999) la transferencia horizontal entre microorganismos ha sido tan intensa que, una red representa mejor el patrón evolutivo que un esquema jerárquico de ancestría-descendencia.

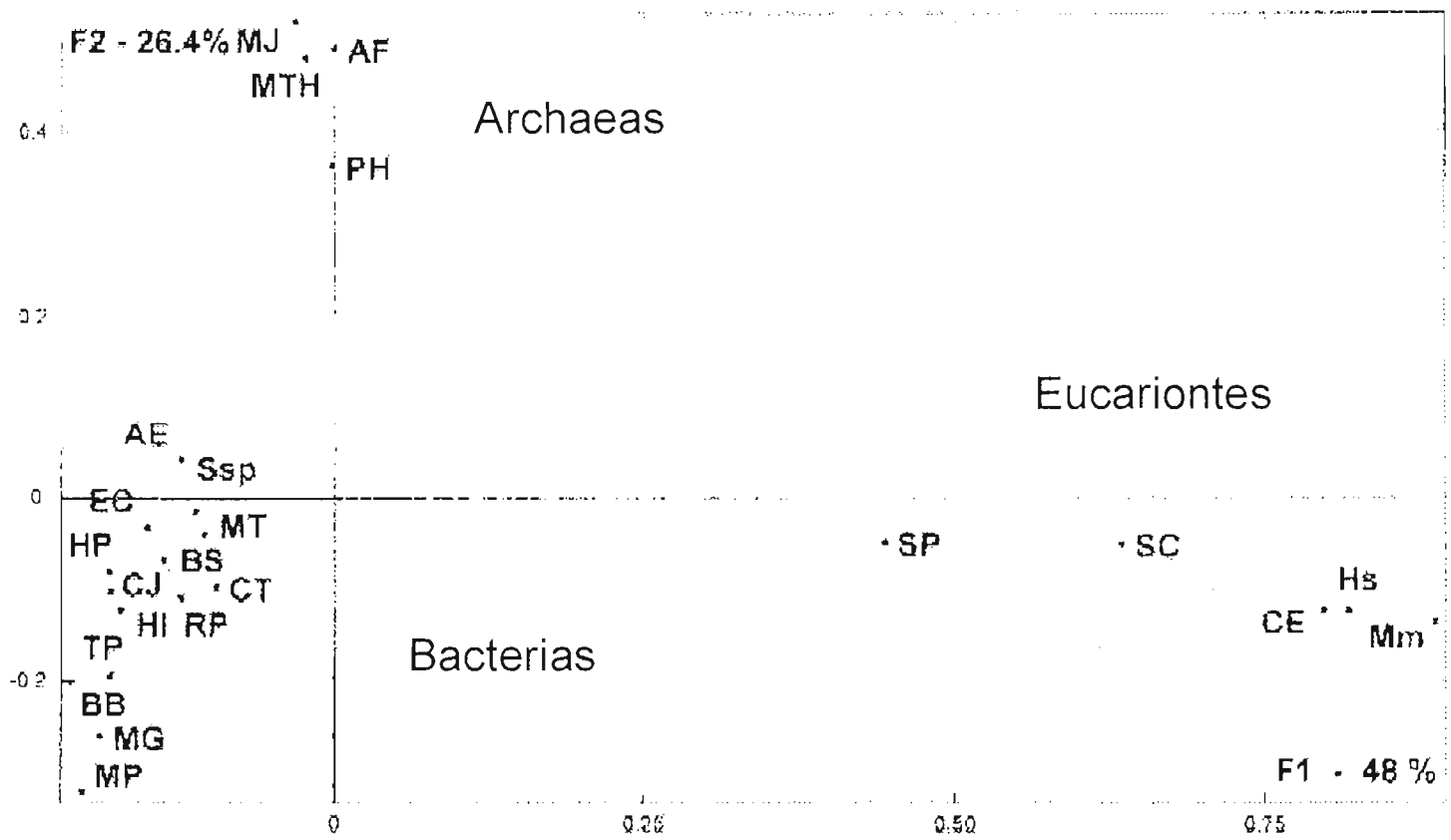


Figura 5. Representación factorial del peso de las duplicaciones ancestrales y ancestría común de cada uno de los genomas, obtenido a partir de correspondencia multidimensional. De acuerdo a Tekaia et al, (1999) fenogramas que reflejan la ausencia o presencia de familias de proteínas sugieren que cada uno de los linajes celulares (Archaea, Bacteria y Eucaria) ha tenido una historia evolutiva independiente.

descendientes. Es decir, que la vida se originó a partir de una molécula que tenía propiedades tanto hereditarias como auto-catalíticas (fenotipo y genotipo en una sola molécula).

El mundo del RNA

La teoría del mundo del RNA, sugerida inicialmente por Gilbert (1986) y elaborada más tarde por Joyce (1989), la cual descansa en las propiedades hereditarias y catalíticas del RNA y en el papel central de este ácido nucleido en la célula, sugiere que la vida comenzó con la aparición de esta molécula a partir de la sopa primitiva. Como veremos, esta teoría ha sido extremadamente exitosa para explicar ciertos aspectos de la evolución temprana de la vida. Sin embargo, aunque la molécula de RNA sea una buena candidata del "gen desnudo" es probable que la vida no haya iniciado con la aparición de éste ácido nucleico debido a que no es claro como se pudieron haber sintetizado algunos de sus componentes a partir de la sopa prebiótica (en especial los azúcares fosfato). Por ello, y a pesar de que no existe evidencia a partir de la biología de los organismos actuales, se ha sugerido que el RNA pudo haber sido precedido por otra molécula, también hereditaria y capaz de auto replicarse (revisado en Joyce, 2002), sin embargo, este sigue siendo un problema sin resolver en campo de origen de la vida.

Un escenario para la evolución temprana de la vida

De cualquier forma, es muy probable que el RNA haya antecedido tanto a! DNA como a las proteínas debido a que esta molécula puede tanto guardar información genética como realizar catálisis biológicas. Sin embargo, si el RNA es la molécula más antigua, ¿cuál fue la siguiente molécula en evolucionar, el DNA o las proteínas?. Si observamos las rutas metabólicas, nos daremos cuenta de que la síntesis de proteínas es básicamente un asunto del RNA (el enlace peptídico es catalizado por el rRNA, el código genético está construido a partir de tRNA aminoacilados que reconocen codones en un mRNA, y la síntesis de proteínas puede tener lugar en ausencia del DNA), mientras que la síntesis y replicación del DNA es sobre todo un asunto de las proteínas (con la excepción del RNA cebador, la síntesis de los dNTP y todo el proceso de replicación está mediado por proteínas (Kronberg & Baker, 1992)). Dicha distribución de funciones nos sugiere el orden histórico en el cual aparecieron estas moléculas (Figura 6). La síntesis de proteínas es a todas luces, un invento del RNA, mientras que los genomas de DNA son una invención de las habilidades metabólicas de las proteínas. Como el DNA no participa en la síntesis de proteínas es poco parsimonioso suponer que existió una etapa en donde las proteínas eran sintetizadas directamente por DNA y que después por alguna razón, el RNA evolucionó para sustituir al DNA en esta función (además de que no se conoce que el DNA tenga alguna función catalítica *in vivo*), y como en principio una célula podría codificar sus propias proteínas en un genoma de RNA (tal como ocurre en algunos sistemas virales) es muy probable que el esquema anterior sea correcto.

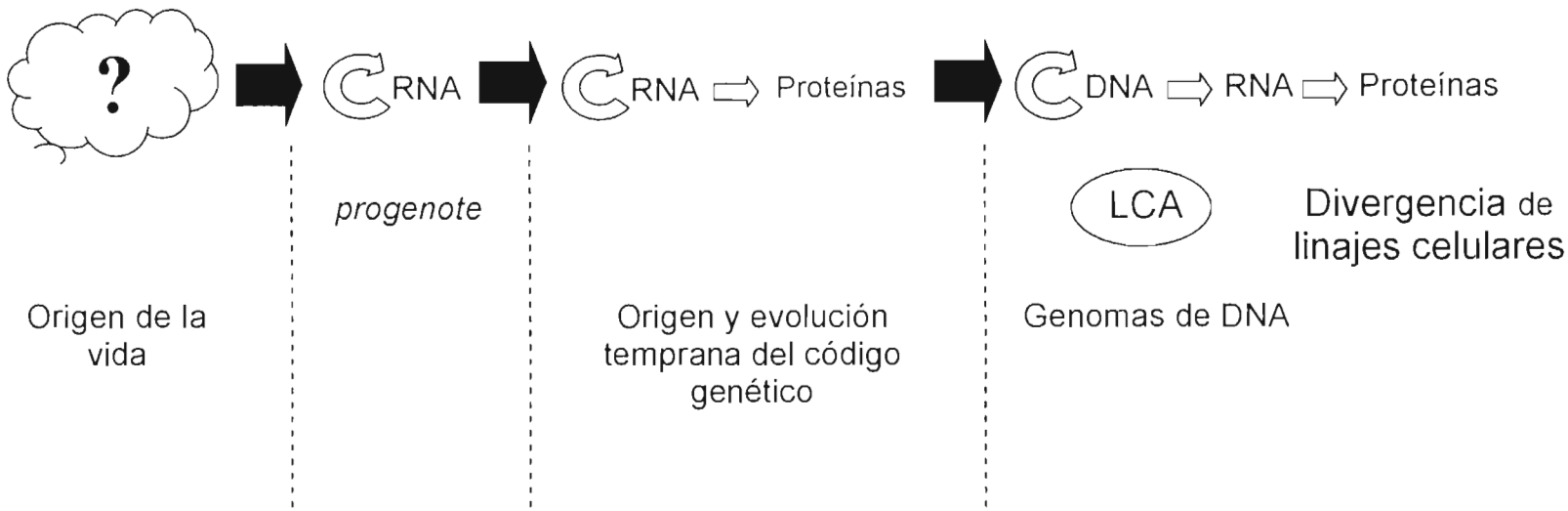


Figura 6. La evolución desde el progenote hasta el LCA. Si el RNA es la molécula semántica más antigua, la explicación más parsimoniosa siguiere que la siguiente molécula en evolucionar fueron las proteínas y finalmente el DNA.

Bajo este esquema, en los sistemas celulares más antiguos que podemos reconocer, el RNA pudo hacer las veces de molécula hereditaria y catalítica. Las evidencias de que el RNA puede realizar una amplia variedad de funciones catalíticas tanto "naturales" como "evolucionadas" en el laboratorio son amplias (revisado en Joyce, 2002) entre las cuales se encuentra la capacidad de ligar templadodependiente un oligonucleótido 3`-hidroxil a un oligonucleótido 5'-trifosfato. Por otro lado, hasta el momento no ha sido posible evolucionar una molécula de RNA capaz de sintetizar dNTP a partir de NTP, habilidad esencial para poder sintetizar DNA (Freeland, 1999). Aunque existen algunas sugerencias de cómo el RNA podría realizar esta síntesis (Joyce, 2002).

Para que las proteínas pudieran entrar al juego evolutivo de herencia, variación y selección natural, el origen de la síntesis de proteínas mediada por RNA, debió de estar acoplado a la codificación de dichas proteínas en el RNA. Sin embargo, comprender el origen y la evolución temprana del código genético sigue siendo la tarea pendiente en los estudios de evolución temprana de la vida. El mundo del RNA/proteína comienza con el origen y la evolución del código genético y la síntesis de las primeras proteínas codificadas y termina con la invención de los genomas de DNA. Con la invención de los genomas de DNA, la información codificada en el RNA tuvo que haber sido copiada al nuevo material genético.

La presión de selección que explica cada transición evolutiva es clara, las proteínas son moléculas mucho más eficientes que el RNA como catalizadores biológicos, y el DNA es una molécula capaz de guardar la información genética de forma más fiel que el RNA, permitiendo codificar una mayor cantidad de habilidades metabólicas, lo que seguramente se pueda traducir en una mayor adecuación en ambos casos (Lazcano, et al 1988).

II. PLANTEAMIENTO DEL PROBLEMA

Reconstrucción del genoma del cenancestro

En principio el genoma del último ancestro común se puede definir como el conjunto de genes homólogos compartidos entre los seres vivos actuales. Sin embargo, fenómenos tales como la transferencia horizontal de genes y las pérdidas secundarias, entre otros, nublan nuestra capacidad para reconstruir de forma correcta la naturaleza genética del *cenancestro*. De cualquier forma, la identificación de los genes universalmente conservados puede arrojar luz acerca de la evolución temprana de la vida. Si bien, es difícil identificar aquellos genes resultados de transferencia horizontal temprana, una reconstrucción del último ancestro común en donde se excluyan genomas de organismos parásitos puede ayudar a disminuir los efectos de las pérdidas secundarias.

Evolución temprana de las DNA polimerasas y la naturaleza del genoma del LCA

La replicación del material genético es uno de los procesos centrales de la vida en la Tierra y con seguridad es uno de los más antiguos. La evolución de los ácidos nucleicos, está íntimamente ligada a la evolución de las enzimas que participan en su metabolismo, siendo las DNA polimerasas, las enzimas que se encargan de su replicación. Como se mencionó anteriormente estas enzimas son de naturaleza distintas entre Bacterias, Archaeas y Eucariontes. En el linaje bacteriano, la enzima replicativa central es la DNA polimerasa III, las euryarchaeas utilizan la polimerasa tipo D, y las crenarchaeas y los eucariontes utilizan la DNA polimerasa tipo II. Siendo la replicación un proceso tan importante y tan antiguo en los seres vivos. Es de extrañarse que la enzima central (la DNA polimerasa) no se encuentre conservada entre los principales linajes celulares, a diferencia de lo que sucede con otros procesos centrales tales como la transcripción, en donde está conservada la subunidad central de la RNA polimerasa DNA dependiente (la subunidad ββ'), y la traducción, en donde están conservadas 34 de las 102 proteínas ribosomales (Lecompte et al, 2002) además de las moléculas de rRNA.

Debido a que todos los sistemas celulares actuales utilizan DNA como molécula de la herencia, el escenario más parsimonioso es un último ancestro común con un genoma de DNA. El objetivo de este trabajo es estudiar la evolución temprana de las DNA polimerasas mediante el estudio de su distribución filogenética así como de su estructura por dominios, con el fin de comprender porque dichas enzimas tienen una naturaleza distinta entre los principales linajes celulares actuales y comprender cual es la relación que existen entre la evolución temprana de estas enzimas y la naturaleza química del genoma del último ancestro común universal (DNA o RNA).

Evolución temprana de dominios de unión a RNA

Si la teoría del mundo del RNA es correcta en indicar que dicha molécula precedió al DNA y las proteínas, y la molécula de DNA fue la última en evolucionar de las tres biomoléculas (Figura 6), las primeras proteínas que se sintetizaron mediante el ribosoma, debieron haber sido seleccionadas ya sea por sus propiedades catalíticas superiores a las del RNA, o por su capacidad de unirse al RNA actuando como chaperonas moleculares. Entonces, entre las proteínas actuales de unión a RNA se encuentran algunos de los dominios de proteínas más antiguos que podemos reconocer. El estudio de la distribución filogenética y las relaciones de ancestría-descendencia entre los dominios de unión a RNA presentes en los genomas completos puede arrojar luz sobre la evolución temprana de algunas de las proteínas más antiguas que podemos reconocer.

III. ESTRATEGIA EXPERIMENTAL

Reconstrucción del genoma del cenancestro

Identificación de genes altamente conservados. En este trabajo hemos analizado veinte genomas celulares completos provenientes de organismos no parásitos, utilizando dos búsquedas de BLAST de un solo sentido para identificar aquellos genes universalmente conservados. Los genes así identificados provienen en principio de etapas tempranas de la evolución de la vida en la Tierra. Las dos búsquedas con los mismos genomas pero en distinto orden permite identificar los falsos negativos.

Identificación de dominios conservados. Debido a que los dominios son las unidades estructurales y evolutivas de las proteínas, se identificaron los dominios conservados entre los grupos de genes homólogos altamente conservados para las especies *Escherichia coli*, *Metanococcus jannashii* y *Sacharomyces cerevisiae*. Este análisis por dominios permite identificar los falsos positivos.

Análisis funcional de dominios altamente conservados. Se realizó un análisis funcional de aquellas proteínas conservadas con la misma estructura de dominios en las tres especies antes mencionadas. Las funciones fueron asignadas de acuerdo a la clasificación funcional de la base de datos del KEGG.

Evolución temprana de las DNA polimerasas y la naturaleza del genoma del LCA

Identificación de los dominios de las polimerasas. De acuerdo a información proveniente tanto de las estructuras terciarias así como de las secuencias de aminoácidos, las polimerasas se pueden clasificar en al menos 10 familias distintas. Con el fin de identificar la estructura por dominios de dichas enzimas, se obtuvieron las estructuras terciarias disponibles para cada una de las familias del Protein Data Bank, y los dominios se identificaron de acuerdo a las publicaciones originales y a la base de datos CATH. Para las familias de polimerasas que no contaban con una estructura terciaria determinada, la estructura por dominios se asignó de acuerdo con la base de datos Pfam.

Análisis de la distribución filogenética de las polimerasas. A continuación se estudio la distribución filogenética de las polimerasas en las bases de datos de proteomas completos utilizando el algoritmo PSI-BLAST.

Análisis evolutivo de las DNA polimerasas. Bajo la premisa de que dos proteínas que comparten la misma estructura terciaria son homólogas, se identificaron aquellos dominios estructurales que fuesen compartidos por dos o más familias de polimerasas. El patrón de conservación de dominios, permite inferir parte de la evolución temprana de dichas enzimas.

Identificación de dominios de unión a RNA. Debido a que los dominios de las proteínas se identifican mejor a nivel de estructura terciaria, se realizó una búsqueda exhaustiva en la literatura, de las proteínas que se ha reportado se unen al RNA y para las cuales se conoce su estructura tridimensional. Una vez localizadas estas proteínas, se identificaron los dominios de unión a RNA presentes en estas proteínas utilizando tres bases de datos distintas: SCOP, CHAT y Pfam.

Análisis de la distribución filogenética de los dominios de unión a RNA. A continuación se buscaron las secuencias homólogas para cada uno de los dominios de unión a RNA identificados previamente, en la bases de datos de proteínas SWISS-PROT y en los proteomas completos disponibles, utilizando el algoritmo PSI-BLAST. Este análisis permitió identificar las proteínas universalmente conservadas (definidas por un lado como los dominios conservados en todos los proteomas analizados, y por otro lado, con homólogos en los tres linajes celulares de acuerdo a SWISS-PROT). Presumiblemente dichas proteínas provienen de una época anterior al origen de los genomas de DNA.

Análisis evolutivo de los dominios de unión a RNA. La duplicación de genes seguida de divergencia es uno de los principales mecanismos que existen para generar nuevos genes, es por ello que para comprender la evolución temprana de los dominios de unión a RNA se identificaron de entre los dominios universalmente conservados, aquellos dominios de unión a RNA que comparten una misma topología y que presumiblemente son el resultado de una duplicación ancestral. Ello permite estudiar el patrón de duplicaciones que dio origen a dichas proteínas.

IV. RESULTADOS

Los resultados se muestran en dos partes. En la primera, muestra un resumen de la principal contribución de cada uno de los trabajos (publicados, en prensa o en preparación) desarrollados durante el doctorado. En la segunda parte se presentan cada uno de los trabajos correspondientes.

IV.I. Resumen de resultados

- En el capítulo de libro: Delaye, L., and Lazcano, A. (2000) **RNA-binding peptides as molecular fossils** *In* J.Chela-Flores, G. Lemerchand, and J. Oró (eds.) *Astrobiology: Origins from the Big-Bang to Civilisation Proceedings of the First Ibero-American School of Astrobiology* (Kluwer Academic Publishers), pp. 285-288, se presenta por primera vez la hipótesis de los dominios de unión a RNA como algunas de las proteínas más antiguas que podemos reconocer. Se presentan algunos ejemplos. También se propone que el mecanismo de regulación genética de la proteína ribosomal S8 representa uno de los mecanismos de regulación de la expresión genética más antiguos.

- En el capítulo de libro: Delaye, L., Vazquez, H., and Lazcano, A.(2001) **The cenancestor and its contemporary biological relics: the case of nucleic acids polymerases** *In* Julian Chela-Flores, J., Owen, T. and Raulin, F. (eds.). *First Steps in the Origin of Life in the Universe* (Kluwier Academic Publishers), pp. 223-230, se muestra evidencia que indica que las DNA polimerasas pertenecientes a las familias I y II comparten el dominio catalítico (el dominio *palm*).

- En el capítulo de libro: Delaye, L., Becerra, A., And Lazcano, A. (2004) **The nature of the last common ancestor**. *In* Lluis Ribas de Pouplana (ed.) *The Genetic Code and the Origin of Life* (Landes Bioscience, and Kluwier Academic), pp. 34-47, se discuten las distintas propuestas que se han hecho para describir la naturaleza del *cenancestro* y se analiza la posibilidad de que haya tenido un genoma de DNA con base en la conservación del dominio *palm* de las polimerasas.

- En el artículo: Luis Delaye and Antonio Lazcano. (2005) **Prebiological evolution and the physics of the origin of life.** *Physics of Life Reviews*, **2**, pp. 47-64, se discuten diversas teorías del origen de la vida y su relación con la naturaleza del *cenancestro*.

- En el artículo: Delaye, L., Becerra, A., and Lazcano, A. (2005) **The Last Common Ancestor: what's in a name?** *Origin of Life and Evolution of the Biosphere* (en prensa), se muestran los

resultados del análisis de 20 genomas celulares utilizando dos análisis de BLAST de "un solo sentido". Los genes universalmente conservados están relacionados principalmente al metabolismo del RNA, lo que sugiere que previo a los sistemas celulares basados en DNA/RNA y proteínas, existió una etapa en donde las células estaban basadas en RNA y proteínas.

- En el capítulo de libro: Lazcano, A., Becerra, A. and Delaye, L. (2004) **On the early evolution of sensory responses: when did life first begin to perceive its surroundings?** *In* Margulis, L. and Asikainen, C. A. (eds) *Human Brain in the Context of Natural History: 3000 million years of evolution of sensory systems* MIT Press, Boston (en prensa), se discuten los posibles sistemas de regulación genética de los primeros sistemas celulares.

- En el manuscrito: Delaye, L., Abascal, F., Fernández, JM., Valencia, A. and Lazcano, A. **Ancient RNA-binding domains: relics from early protein evolution** (en preparación), se estudia la distribución filogenética de 68 dominios de unión a RNA. Se identifican 35 dominios universalmente conservados en 122 proteomas provenientes de los tres linajes celulares (Archaea, Bacteria y Eucaria). Debido a la amplia distribución filogenética de estos dominios, se sugiere que se encontraban presentes en el *cenancestro* y que probablemente representan algunas de las proteínas más antiguas que podemos identificar. Se propone que su origen se remonta al mundo del RNA/proteína. De acuerdo a la similitud en estructuras terciarias, ocho de estos dominios universalmente conservados son el resultado de duplicaciones genéticas. Esto sugiere que la duplicación genética ha jugado un papel importante en la generación de nuevas proteínas desde etapas muy tempranas de la evolución biológica.

- En el manuscrito: Delaye, L., Becerra, A., and Lazcano, A. **On the early evolution of polymerase function and the nature of the genome of the cenancestor** (en preparación), se muestran los resultados del estudio sobre la evolución temprana de las DNA polimerasas y la naturaleza del genoma del LCA. En este manuscrito se clasifican a las polimerasas con base en secuencia y en estructura en 10 grupos distintos y se demuestra que en la evolución de estas enzimas no ha estado ausente el fenómeno de sustitución no ortóloga. De acuerdo a ello, se sugiere que dicho evento pudo haber ocurrido en la base del árbol universal, teniendo como resultado una DNA polimerasa bacteriana de distinta naturaleza a las polimerasas Archaeo/Eucariontes. También se sugiere que el dominio catalítico (el dominio *palm*) de las polimerasas pertenecientes a las familias I, II, RNA polimerasas virales y reverso transcriptazas es homólogo entre sí. Dicho dominio pudo provenir de la polimerasa encargada de replicar los genomas de RNA antes de la aparición de los genomas de DNA.

IV.II Trabajos publicados, en prensa o en preparación durante el doctorado

RNA-BINDING PEPTIDES AS EARLY MOLECULAR FOSSILS

LUIS DELAYE and ANTONIO LAZCANO
Facultad de Ciencias, UNAM
Apdo. Postal 70-407
Cd. Universitaria, 04510
México D.F., MÉXICO

Abstract

Comparisons of complete cellular genomes indicate that a set of genes whose products synthesize, degrade, or interact with RNA molecules are among the most highly conserved sequences common to all living beings, and therefore may have been present in their last common ancestor, i.e., the cenancestor. In order to obtain insights on the evolution of sequences which may date from an early evolutionary period during which RNA played a genetic role prior to the emergence of DNA genomes, we have analyzed the conserved RNA-binding sites of these highly conserved molecules, since these may be some of the recognizable peptides in our dataset. The characteristics of some of these highly conserved amino acid stretches which are essential in RNA metabolism are discussed.

1. Introduction

The early steps of the evolution of life on Earth can be investigated by analyzing the extant molecular fossil record. A *molecular fossil* as defined by Maizels and Weiner (1994), is *any molecule whose contemporary structure, function (or its phylogenetic distribution) provides a clue to its evolutionary history.* Thus, phylogenetic markers such a rRNA (Woese 1987) are good molecular fossils, but the same is true for other biological molecules.

The discovery of ribozymes has given considerable credibility to prior suggestions on the existence of the RNA world, a hypothetical stage before the evolutionary development of DNA genomes protein-based metabolism. Whether RNA was the first genetic macromolecule or not is a matter of debate, but acceptance of the RNA world can help define the evolutionary *polarity* of a number of molecular traits, i.e., to recognize which character states are ancestral and which are derived. Thus, if RNA preceded DNA as the reservoir of cellular genetic information, and proteins predate DNA, it is likely that proteins which interact with RNA are older than those that do so with DNA.

J. Chela-Flores et al. (eds.), Astrobiology, 285–288.

It is highly unlikely that the first proteins were complex enzymes with exquisitely finely tuned catalytic activity. Although the first peptides that were synthesized biologically (i.e., ribosome-mediated translation), could be positively selected by two properties that are not mutually exclusive, i.e. chaperone-like properties or catalytic activity, in these paper we explore the possibility that the first peptides could have enhanced the catalytic properties or biological functions of ribozymes simply by stabilizing their structures. This chaperone-like property would be the primitive equivalent to the stabilizing effect that the protein subunit of RNAse P plays *in vivo* (Guerrier-Takada *et al.*, 1983), and can in principle be explored by a detailed analysis of extant RNA-binding sites.

Comparisons of completely sequenced cellular genomes (Table 1) from the three primary domains (i.e., Bacteria, Archaea and Eucarya) suggest that a set of genes whose products synthesize, degrade, or interact with RNA molecules are among the most highly conserved sequences common to all living beings (Tekaia *et al.*, 1999). It is thus possible that their corresponding RNA-binding sites, which may be among the oldest motifs in current sequence databases, can provide important insights on the early evolution of ribosome-mediate polypeptide synthesis. Here we report the preliminary outcome of such analysis, and discuss the evolutionary significance of our findings.

2. Material and methods

A list of RNA-binding domains reported in the literature was compiled, which include the those of highly conserved proteins as defined by Tekaia *et al.* (1999). This dataset is now being completed with searches on the following databases: SWISS-PROT (http://www.expasy.ch/sprot/), SRS (http://srs5.hgmp.mrc.ac.uk/), and PROSITE (http://www.expasy.ch/prosite/). The phylogenetic distribution of these RNA-binding sites was analyzed following the three domain taxonomic scheme.

3. Results and discussion

In this first phase of our study we have restricted ourselves to the RNA-binding motifs reported in the literature. Thus, it is likely that the results reported represent a lower limit of the different functional kinds of RNA-binding sequences, i.e., that there are many other as yet undetected cases of such polypeptides.

The distribution of such RNA-binding domains among different highly conserved proteins with different functions (Table 1), suggests that domain recruitment and fusion have taken place in the early evolution of such polypeptides. It also implies that some of these domains are probably older than the proteins in which they are found today, i.e., they are molecular fossils in the sense described above.

RNA-binding domain	Highly conserved proteins	References
OB fold	RpS17, AspRS, LysRS, PheRS, AsnRS	Arnez and Cavarelli, 1997; Eriani *et al.*, 1990
RNP	EF-G, PheRS	Liljas and Garber, 1995; Mosyak *et al.*, 1995
Left handed βαβ cross over	RpS5, EF-G	Stams *et al.*, 1998
HU DNA-binding like	RpS7, RpL14	Draper and Reynaldo, 1999
Novel α/β fold	HisRS, ProRS, ThrRS, GlyRS*	Arnez *et al.*, 1995; Cusack *et al.*, 1998; Sankaranarayanan *et al.*, 1999, Logah *et al.*, 1995

Table 1. RNA-binding domains found in some of the highly conserved proteins.
*GlyRS from *Thermus thermphilus*; Rp (ribosomal protein); RS (aminoacyl-tRNA synthetase).

One example of such molecular fossil could be the ribosomal protein S8 (RpS8). This polypeptide is one of the core ribosomal proteins. It binds to the 16/18S rRNA with high affinity, and plays a central role in the assembly of the 30S subunit of the ribosome. The *E. coli* RpS8 also regulates its gene expression by binding to its own mRNA, thereby acting as translational repressor of the *spc* (spectinomycin-resistance) operon. The *spc* operon includes the genes of the ribosomal proteins L14, L24, L5, S14, S8, L6, L18, S5, L30, and L15 (Mattheakis & Nomura, 1988). The RpS8 target site on the *spc* mRNA is similar to the 16S rRNA S8 binding site in both at primary and secondary structure levels (Cerreti *et al.*, 1988; Gregory *et al.*, 1988).

The ribosomal protein S8 resulted from the fusion of two domains. Its N-terminal domain is similar to portions of the DNase I and *Hae*III methyltransferase that bind to DNA (Davies, *et al.*, 1996). The C-terminal domain seems to be an RNA-binding domain found only in this protein. Mutants with only eight residues missing from the C-terminus exhibit significantly lowered RNA-binding properties (Uma *et al.*, 1995). The level of conservation of the primary structure of the RpS8 across the three cellular domains and of the *spc* operon in the two prokaryotic domains (Siefert *et al.*, 1997); together with the fact that almost the same RNA structure is recognized by the protein for its function on the rRNA and for its own regulation in the mRNA, strongly suggest that this mechanism of auto-regulation of gene expression may be among the oldest ones that evolved during the evolution of the protein biosynthesis.

Perhaps not surprisingly, some of the RNA-binding domains resemble DNA-binding domains, such as the ETS DNA-binding motif-like of RpS4, or the helix-hairpin-helix motif of RpS13. It is possible that such domains would correspond to polypeptides with functions already in the RNA-protein world, adapted to the DNA-RNA-protein world.

We are currently working on the construction of a catalog of RNA-binding domains. Analysis of this dataset, is expected to provide insight into the characteristics of which may be some of the oldest polypeptides still recognizable today.

Support from the DGAPA-UNAM/ PAPIIT IN213598 project is gratefully acknowledged.

4. References

Arnez, J.G. and Cavarelli, J. (1997) Structure of RNA-binding proteins. *Quart. Rev. Bioph.* **30**, 195-240

Arnez, J.G., Harris, D.C., Mitschler, A., Rees, B., Francklyn, C.S., and Moras, D. (1995) Crystal structure of histidyl-tRNA synthetase from *Escherichia coli* complexed with histidyl-adenylate. *EMBO J.* **14**, 4143-4155

Cerretti, D.P., Mattheakis, L.C., Kearney, K.R., Vu, L., and Nomura, M. (1988) Translational regulation of the *spc* operon in *Escherichia coli*. Identification and structural analysis of the target site for S8 repressor protein. *J. Mol. Biol.* **204**, 309

Cusack, S., Yaremchuk, A., Krikliviy, I., and Tukalo, M. (1998) $tRNA^{Pro}$ anticodon recognition by *Thermus thermophilus* prolyl-tRNA synthetase. *Structure*, **6**, 101-108

Davies, C., Ramakrishnan, V., and White, S.W. (1996) Structural evidence for specific S8-RNA and S8-protein interactions within the 30S ribosomal subunit: ribosomal protein S8 from *Bacillus stearothermophilus* at 1.9 Å resolution. *Structure*, **4**, 1093-104

Draper, D.E., and Reynaldo, L.P. (1999) Survey and summary RNA binding strategies of ribosomal proteins. *Nucl. Acids Res.* **27**, 381-388

Eriani, G., Dirheimer, G. and Gangloff, J. (1990) Aspartyl-tRNA synthetase from *Escherichia coli*; cloning and characterization of the gene, homologies of its translated amino acid sequence with asparaginyl- and lysyl-tRNA syhtetase. *Nucl. Acids Res.* **18**, 7109-7118

Gregory, R.J., Cahill, P.B., Thurlow, D.L., and Zimmermann, R.A. (1988) Interaction of *Escherichia coli* ribosomal protein S8 with its binding sites in ribosomal RNA and messenger RNA. *J. Mol. Biol.* **204**, 295-307

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849-57

Liljas, A. and Garber, M. (1995) Ribosomal proteins and elongation factors. *Curr. Opin. Struct. Biol.* **5**, 721-727

Logah, D.t., Mazauric, M.-H., Kern, D., and Moras, D. (1995) Crystal structure of glycyl-tRNA synthetase from *Thermus thermophilus*. *EMBO J.* **14**, 4156-4167

Maizels, N. and Weiner, A.M. (1994) Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation, *Pro. Natl. Acad. Sci. USA* **91**, 6729-34

Mattheakis, L.C. and Nomura, M. (1988) Feedback regulation of the *spc* operon in *Escherichia coli*: Translational coupling and mRNA processing, *J. Bacteriol.* **170**, 4484-92

Mosyak, L., Reshetnikova, L., Goldgur, Y., Delarue, M., and Safro, M.G. (1995) Structure of phenylalanyl-tRNA synthetase from *Thermus thermophilus*, *Nature Struct. Biol.* **2**, 537-547

Sankaranarayanan, R., Dock-Bregeon A.C., Romby, P., Caillet, J., Springer, M., Rees, B., Ehresmann, C., Ehresmann, B., and Moras, D. (1999) The structure of threonyl-tRNA synthetase (Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site, *Cell*, **97**, 371-381

Siefert, J.L., Martin, K.A., Abdi, F., Widger, W.R., and Fox, G.E. (1997) Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *J. Mol. Evol.* **45**, 467-72

Stams, T., Niranjanakumari, S., Fierke, C.A., and Christianson, D.W. (1998) Ribonuclease P protein structure: evolutionary origins in the translational apparatus, *Science*, **280**, 752-755

Tekaia, F., Dujon, B., and Lazcano, A. (1999) Comparative genomics: products of the most conserved protein-encoded genes synthesize, degrade, or interact with RNA. Abstract of the 12th International Conference on the Origin of Life & 9th ISSOL meeting, ISSOL'99 (San Diego, California U.S.A. July / 11-16 / 1999) pp. 53

Uma, K., Nikonowicz, E.P., Kaluarachchi, K., Wu, H., Wower, I.K., and Zimmermann, R.A. (1995) Structural characterization of *Escherichia coli* ribosomal protein S8 and its binding site in 16S ribosomal RNA. *Nucleic Acids Symp. Series* **33**, 8-10

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

THE CENANCESTOR AND ITS CONTEMPORARY BIOLOGICAL RELICS: THE CASE OF NUCLEIC ACID POLYMERASES

Luis Delaye, Héctor Vázquez, and Antonio Lazcano
Facultad de Ciencias, UNAM
Apdo. Postal 70-407
Cd. Universitaria, 04510 México D.F.
MÉXICO
E-mail: alar@hp.fciencias.unam.mx

1. Introduction

The recognition that different macromolecules may be uniquely suited as molecular chronometers in the construction of nearly universal phylogenies has widened the range of phylogenetic studies to previously unsuspected heights. In particular, the use of small subunit ribosomal RNAs (16/18S rRNA) as molecular markers led to the construction of a trifurcated, unrooted tree in which all known organisms can be grouped in one of three major monophyletic cell lineages: the eubacteria, the archaebacteria, and the eukaryotic nucleocytoplasm, now referred to as the domains *Bacteria*, *Archaea*, and *Eucarya*, respectively (Woese et al., 1990). The construction of the 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 branches 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 suggests that the conclusion that the last common ancestor was a progenote may have been premature. Pending the issue of horizontal gene transport (Figure 1), a partial description of the last common ancestor (LCA) of eubacteria, archaebacteria, and eukaryotes may be inferred from the distribution of homologous traits among its descendants. Ten years ago, the set of such genes that had been sequenced and compared was still small, but the sketchy picture that had emerged suggested that the most recent common ancestor of all extant organisms, or *cenancestor*, as defined by Fitch and Upper (1987), was a rather sophisticated cell (Lazcano, Fox and Oró, 1992) with at least (a) DNA polymerases endowed with proof-reading activity; (b) ribosome-

J. Chela-Flores et al. (eds.). First Steps in the Origin of Life in the Universe, 223–230.

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).

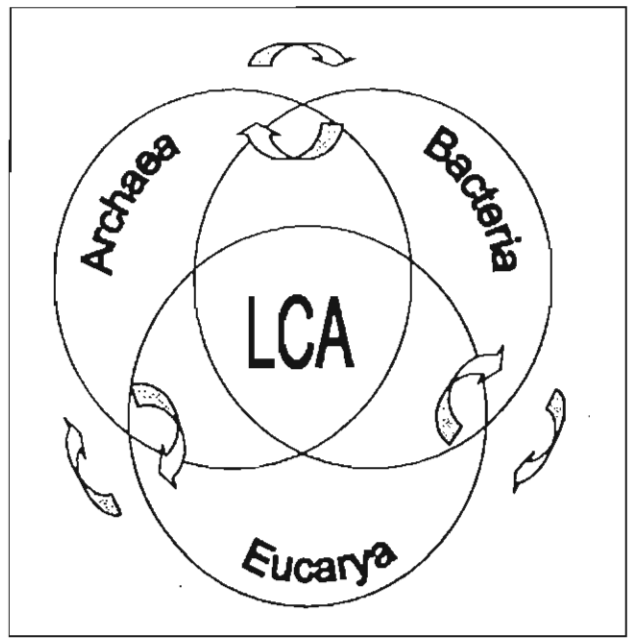


Figure 1. The gene complement of the LCA is defined by the intersection of the complete genomes of the three domains. The arrows represent the horizontal gene transfer between cellular domains.

Recent results have confirmed the above conclusions. These traits 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 cenancestral population was not a direct, immediate descendant of the RNA world, a protocell or any other pre-life progenitor system (Lazcano, 1995). Very likely, the LCA 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. Moreover, the universal distribution of the same essential features of genome replication, gene expression, basic anabolic reactions, and membrane-associated ATPase-mediated energy production in all known organisms 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.

While trees based on whole genome information have confirmed at a broad level the 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. 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), but there is evidence of extensive horizontal transfer events that have taken place in the past (Doolittle, 1999). In fact, in addition to lateral gene transfer (Figure 1), insights into cenancestral states can be strongly hindered by inadequate biodiversity sampling, polyphyletic gene losses, unequal rates of molecular evolution,

convergence, polyphyly, and secondary loss of organelles. These factors clearly limit our ability to recognize the extant molecular relics of the cenancestor.

1.1 THE SEARCH FOR THE ANCESTRAL NUCLEIC ACID POLYMERASE

Replication of genetic material must have been one of the oldest functions to evolve (Figure 2). Ideally, abiotic laboratory polymerization of nucleotides should provide insights into the transition from the prebiotic broth to the extant enzyme-mediated replication of nucleic acids.

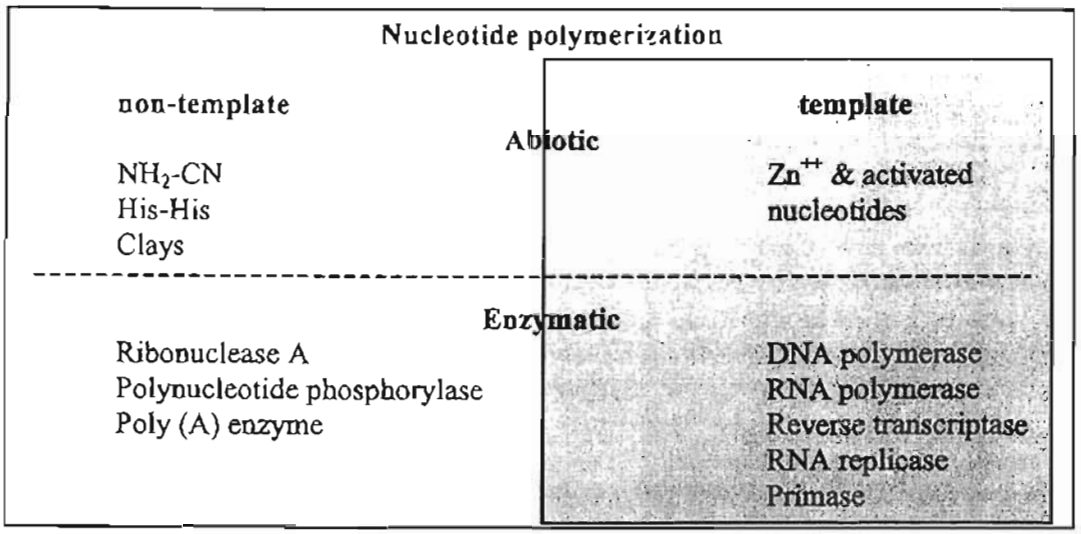


Figure 2. The abiotic and enzymatic polymerization of nucleotides.

In principle, this could also explain the evolutionary development of polymerases, an issue directly related to the chemical composition of the cenancestral genome. Since all extant cells are endowed with DNA genomes, the most parsimonious conclusion is that such genomes were already present in the cenancestral population. However, this hypothesis has been contested by suggestions of an RNA- (Mushegian and Koonin, 1996) or even a mixed DNA-RNA genome for the LCA (Leipe et al., 1999). These proposals are based, at least in part, on the low level of conservation of the primary structure of DNA polymerases (Olsen and Woese, 1996; Edgell and Doolittle, 1997), as well as on the striking differences in their phylogenetic distribution compared with rRNAs, aminoacyl-tRNA-synthetases, and other molecules involved in transcription and translation. This has led to suggestions that DNA genomes, together with the corresponding polymerases, may have been invented independently in the different cell domains (Mushegian and Koonin, 1996; Leipe, *et al.* 1999).

Evolution of enzymes in biological systems often involves the acquisition of new catalytic or binding properties by an existing protein scaffold. However, identification of several non-homologous classes of nucleic acid polymerases (primase, reverse transcriptase (RT), RNA polymerase and DNA polymerase) shows that this is not the

case for these enzymes, and demonstrates the polyphyletic origin of template-dependent enzyme-mediated synthesis of phosphodiester bonds (Steitz, 1999).

Based on sequence similarity and crystal structure analysis (Steitz, 1999) DNA polymerases have been classified into five families (Table 1). Three dimensional structures are available for the DNA polymerase families defined by the DNA pol I, DNA pol α, RT, and rat DNA pol β prototypes.

Family	Representatives
DNA polymerase I family (A polymerase family)	- Klenow fragment of *Escherichia coli* DNA polymerase I - Klenow fragment of *Bacillus* DNA polymerase I - *Thermus aquaticus* DNA polymerase - T7 RNA and DNA polymerases
DNA polymerase α (B family DNA polymerase or family II)	- All eukaryotic replicating DNA polymerases (α,δ,ε) - Phage T4 DNA polymerase - RB69 Phage polymerase
Reverse transcriptase family	- HIV reverse transcriptase - RNA-dependent RNA polymerase - Telomerase
Rat DNA polymerase β	- DNA polymerase β (rat)
Bacterial DNA polymerase III	- Bacterial DNA polymerase III, on the basis of amino acid sequence comparisons.

Table 1. Classification of DNA polymerases into five families according to sequence similarity and tertiary structure criteria (cf. Steitz, 1999).

All DNA polymerases whose tertiary structure has been determinated appear to share a common overall architectural feature comparable to a right hand shape. This structure is not so evident, however, in the case of rat DNA pol β and its homologues. The structure of the other polymerases has been described as consisting of "thumb", "palm", and "finger" domains (Kohlstaedt, et al, 1992). Detailed analysis of the three dimensional structure of DNA polymerases from the pol I, pol α, and RT families suggest that their palm sub-domain has a single origin, i.e., it is homologous in all of them, while the fingers and the thumb sub-domains are different in all four of the families for which structures are known (Brautigam and Steitz, 1998). The complex evolutionary history of nucleic acid polymerases, combined with the wide sequence space explored by these enzymes during biological evolution, strongly hinders the identification of the ancestral polymerase.

As argued here, the three-dimensional homology between the palm domains of DNA polymerase I and DNA polymerases B, which includes all eukaryotic replicating DNA polymerases (Steitz, 1999), can be extended to suggest that such domain, which catalyses the phosphodiester bond, was already present in the cenancestor. As shown here, the structural multiple alignment of the palm sub-domain of DNA polymerases belonging to the pol I and pol α families from the tree cellular domains of life strongly

suggests that this sub-domain is the most ancient protein segment found within these enzymes and could have been present in the LCA.

2. Material and Methods

The crystal structures from the following DNA polymerases sequences were downloaded from Protein Data Bank (www.rcsb.org/pdb/): DNA polymerases A family: 1KLN, *Escherichia coli*; 1TAQ, *Thermus aquaticus*; 1XWL *Bacillus stearothermophilus*; and from DNA polymerases B family: 1TGO, *Thermococcus gorgonarius*; 1D5A, and *Desulfurococcus sp.* Tok;.

The palm sub-domains of all of them, following the classification of CATH database (www.biochem.ucl.ac.uk/bsm/cath_new/index.html), were aligned manually using the program SPDBV (Guex, and Peitsch, 1997) (www.expasy.ch/spdbv/text/refs.htm) to construct a structural multiple alignment.

The sequence of the palm domain from *T. gorgonarius* (1TGO) was used as a query against the SwissProt database in the NCBI server (www.ncbi.nlm.nih.gov/BLAST/), using Blast (Altschul, et al, 1997). Sequences from eukaryoticDNA polymerases thus identified using this metod, were added to the structural multiple alignment using the program ClustalX v1.81 (Thompson, et al, 1997).

The multiple structural alignment was performed by first aligning the two archaeal and the three bacterial palm sub-domains separately, in order to identify the conserved residues in each of the families. This was followed by the manual alignment of all structures looking for the 3-dimensional conserved residues identified before.

3. Results

The multiple structural alignment of the primary structure of the different palm sub-domains in, is shown in Figure 3.

The Blast search found eight eucaryotic DNA polymerases: DPOD_HUMAN DNA polymerase delta catalytic subunit (Expect = 4e-04); DPOD_BOVIN DNA polymerase delta catalytic subunit (Expect = 5e-04); DPOD_MESAU DNA polymerase delta catalytic subunit (Expect = 7e-04); DPOD_RAT DNA polymerase delta catalytic subunit (Expect = 8e-04); DPOZ_HUMAN DNA polymerase zeta catalytic subunit (Expect = 9e-04); DPOD_MOUSE DNA polymerase delta catalytic subunit (Expect = 0.001); DPOZ_MOUSE DNA polymerase zeta catalytic subunit (Expect = 0.001); DPOD_SOYBN DNA polymerase delta catalytic subunit (Expect = 001).

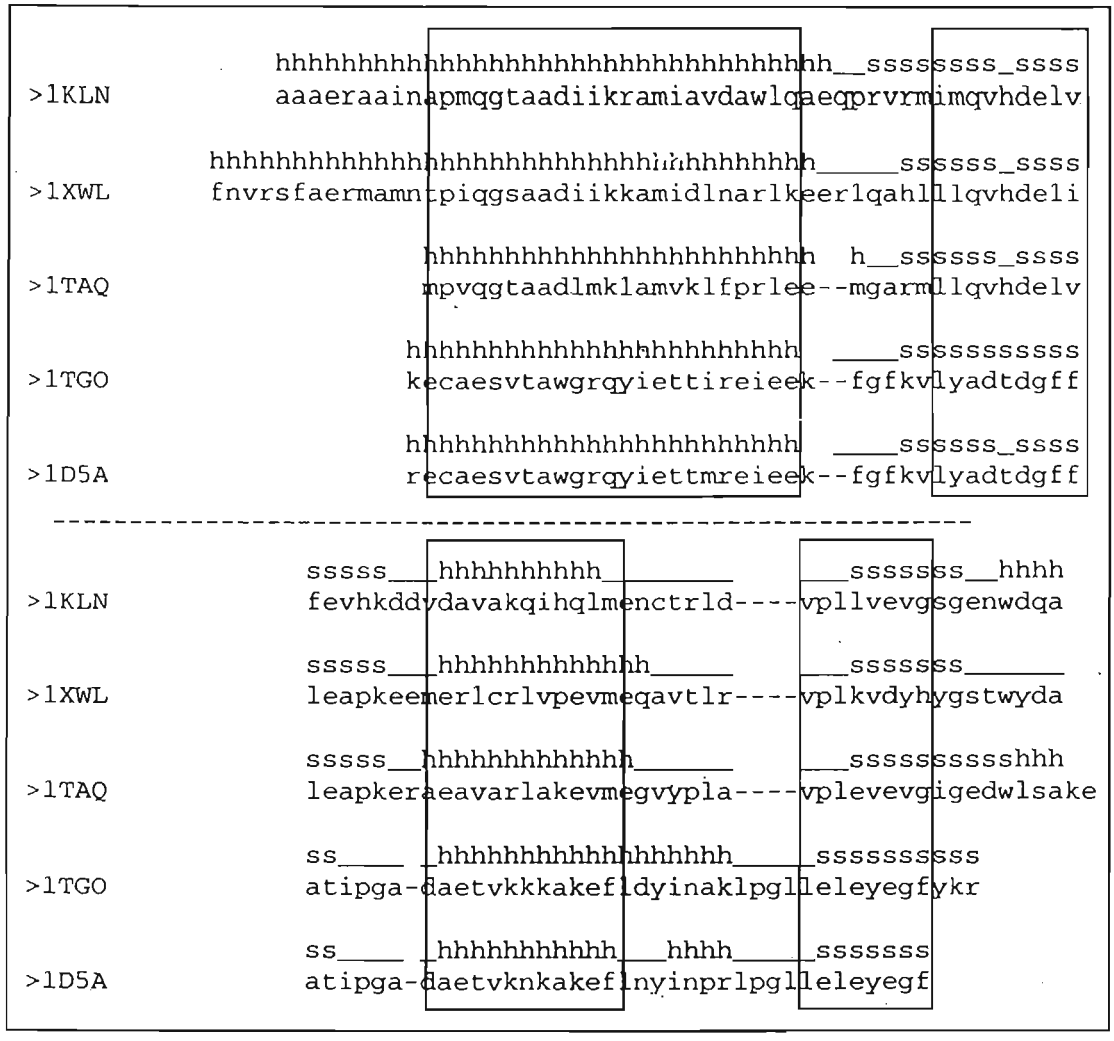


Figure 3. Multiple structural alignment of the palm subdomains. Enclosed in boxes are the residues which are structurally homologous in all the domains studied here. The sequences are as follows: DNA polymerases A family: 1KLN, *Escherichia coli*; 1XWL *Bacillus stearothermophilus;* 1TAQ, *Thermus aquaticus*; and from DNA polymerases B family: 1TGO, *Thermococcus gorgonarius*; 1D5A, and *Desulfurococcus sp*. Tok. Above each sequence the secondary structure is shown: h, α-helix, and s, for β-strand.

4. Discusion and conclusions

Although the availability of completely sequenced cellular genomes has enhanced the likehood of more accurate reconstructions of ancestral states, horizontal gene transfer can strongly hinder our ability to understand the characteristics of the last common ancestor. As shown by the current discussions on the chemical nature of the LCA genome (Mushegian and Koonin, 1996; Leipe, *et al.*, 1999), our attempts to understand the distant past can also be limited by the polyphyletic origin of DNA polymerases, whose classification into different families reflects a case of convergence. Nevertheless, the evidence presented here clearly shows that the palm subdomains of the I and II

DNA polymerase families, which are found in all three cell lineages, have a common origin that has conserved the same tertiary structure and is thus an indication of the monophyletic origin of these enzymes. The lack of a crystalized eukaryotic replicative DNA polymerase has not allowed the recognition of the common origin of these polymerases. As shown here however, their monophyletic origin is recognizable even at the primary structure level (Blast search). The evolutionary conservation of this subdomain, which is involved in the catalysis of the phosphoribosyl transfer reaction (Steitz, 1999), is probably due to the central role it plays in the synthesis of polynucleotides.

On the other hand, the lack of homology between the other subdomains (i.e., the thumb and finger) indicates the easyness by which nucleotide-binding motifs can evolve. A possible evolutionary sequence of nucleotide polymerization agents, starting from the prebiotic synthesis of phosphodiester bonds (and omitting the existence of possible preRNA worlds) is shown in Figure 4. This scheme is based on Steitz (1999) suggestion of a stepwise-emergence of functional peptides in an ribozymic replicase, and on the evolution of polymerization agents discussed elsewhere (Lazcano et al., 1988).

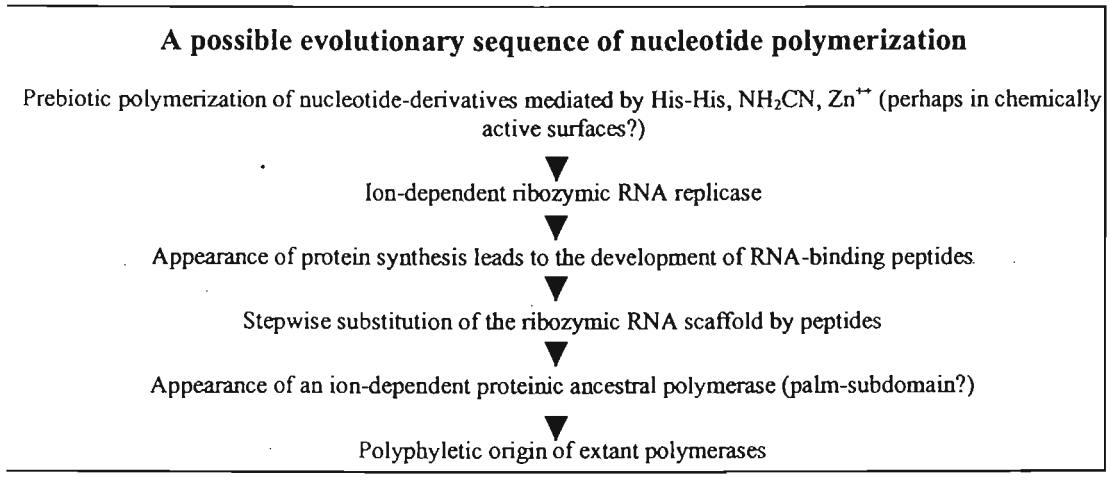


Figure 4. Possible evolutionary sequence of nucleotide polymerization agents, starting from the prebiotic synthesis of phosphodiester bonds (and omitting the existence of possible preRNA worlds).

Given the lack of absolute chemical specificity that polymerases exhibit for both template and substrate (Lazcano et al., 1988), it is quite possible that the conserved ion-dependent palm-subdomain discussed here was part of an ancestral replicase and transcriptase during the RNA/protein world stage (Figure 4). This possibility is supported by the homology between the viral T7 RNA and DNA polymerase. However, the highly conserved sequences of the β and β' subunits of the DNA-dependent RNA polymerase which are found in all three cellular domains, indicate that by the time the LCA had evolved, a modern type of oligomeric RNA polymerase had already evolved. Why polymerases have originated independently several times and why the level of divergence within each family of DNA polymerases is so high, are still open questions that deserve further attention.

Acknowledgements

Support from DGAPA-UNAM/PAPIIT IN213598 project is gratefully acknowledged.

5. References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* **25**, 3389-3402

Brautigam, C.A., and Steitz, T.A. (1998) Structural and functional insights provided by crystal structures of DNA polymerases and their substrate complexes, *Curr. Opin. Struct. Biol.*, **8**, 54-63

Doolittle, W.F. (1999) Phylogeny 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

Edgell, R.D. and Doolittle, W.F. (1997) Archaea and the origin(s) of DNA replication proteins, *Cell*, **89**, 995-998

Fitch, W.M., Upper, K. (1987) The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code, *Cold Spring Harb Symp Quant Biol.* **52**, 759-767

Guex, N. and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling, *Electrophoresis.* **18**, 2714-2723

Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice,, P.A., and Steitz, T.A. (1992) Crystal structure at 3.5 Å of HIV-1 reverse transcriptase complexed with an inhibitor, *Science*, **264**, 1781-1790

Lazcano, A., Fastag, J., Gariglio, P., Ramírez, C. and Oró, J. (1988) On the early evolution of RNA polymerase, *J. Mol. Evol.* 27, 365-376

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

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

Leipe, D.D., Aravind, L., Koonin, E.V. (1999) Did DNA replication evolve twice independently? *Nucleic. Acids. Res.* **27**, 3389-3401

Mushegian, A.R., and Koonin, E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes, *Proc. Natl. Acad. Sci. U S A*, **93**, 10268-10273

Olsen, G.J. and Woese, C.R. (1996) Lessons from an Archaeal genome: what are we learning from *Methanococcus jannaschii*? *Trends. Genet.* **12**, 377-379

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

Steitz, T.A. (1999) DNA polymerases: structural diversity and common mechanisms, *J. Biol. Chem.* **274**, 17395-17398

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

Thompson,J.D., Gibson, T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res.* **24**, 4876-4882

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

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

MOLECULAR BIOLOGY
INTELLIGENCE
UNIT

The Genetic Code and the Origin of Life

Lluís Ribas de Pouplana, Ph.D.
The Scripps Research Institute
La Jolla, California, U.S.A.
and
ICREA and Barcelona Institute for Biomedical Research
Barcelona Science Park, Barcelona, Spain

LANDES BIOSCIENCE / EUREKAH.COM
GEORGETOWN, TEXAS
U.S.A.

KLUWER ACADEMIC / PLENUM PUBLISHERS
NEW YORK, NEW YORK
U.S.A.

CHAPTER 3

The Nature of the Last Common Ancestor

Luis Delaye, Arturo Becerra and Antonio Lazcano

Introduction

Until the late 1970s cellular evolution was assumed to be a continuous, unbroken chain of progressive transformations that begun with the emergence of life itself and continued until the endosymbiotic origin of eukaryotes marked the major biological discontinuity. This scheme was challenged when the comparison of small subunit ribosomal RNA (16/18S rRNA) sequences led to the construction of a trifurcated, unrooted tree in which all known organisms can be grouped in one of three major monophyletic cell lineages, i.e., the domains Bacteria (eubacteria), Archaea (archaeabacteria), and Eucarya (eukaryotes).[1] Information from one single molecular marker does not necessarily yield a precise reconstruction of evolutionary processes, but as shown by numerous phylogenies constructed from other genes such as those encoding polymerases, elongation factors, F-type ATPase subunits, heat-shock and ribosomal proteins, the identification of the three major lineages is not an artifact based solely upon the reductionist extrapolation of information derived from the rRNA tree, but a true reflection of an ancient trifurcation.

Cladistic analysis of rRNA sequences is acknowledged as a prime force in systematics, and from its very inception had a major impact in our understanding of cellular evolution. As shown by the unrooted rRNA trees, no single domain predates the other two and all three derive from a common ancestor. Recognition of the significant differences that exist between the transcriptional and translational machineries of the Bacteria, Archaea and Eucarya, which were assumed to be the result of independent evolutionary refinements, led to the conclusion that the primary branches were the descendants of a progenote, a hypothetical biological entity in which phenotype and genotype still had an imprecise, rudimentary linkage relationship.[2]

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 However, the conclusion that the last common ancestor (LCA) was a progenote was disputed over ten years ago when the analysis of homologous traits found among some of its descendants suggested that it was not a direct, immediate descendant of the RNA world, a protocell or any other prelife progenitor system. Under the assumption that horizontal gene transfer (HGT) had not been a major driving force in the distribution of homologous traits in the three domains, it was concluded that the LCA was a complex organism, much alike extant bacteria.[3,4] A decade ago the inventory of such shared features was small, but it was surmised that the sketchy picture developed with the limited databases would be confirmed when completely sequenced cell genomes from the three primary domains. This has not been the case: the availability of an increasingly large number of completely sequenced cellular genomes has sparked new debates, rekindling the discussion on the nature of the ancestral entity.[5] This is shown, for instance, in the diversity of names that have been coined to describe it: progenote,[2] cenancestor,[6] LUCA, last universal cellular ancestor,[7] and LCC, last common community,[8] among others. These terms are not truly synonymous, and they reflect the current controversies on the nature

The Genetic Code and the Origin of Life, edited by Lluís Ribas de Pouplana.

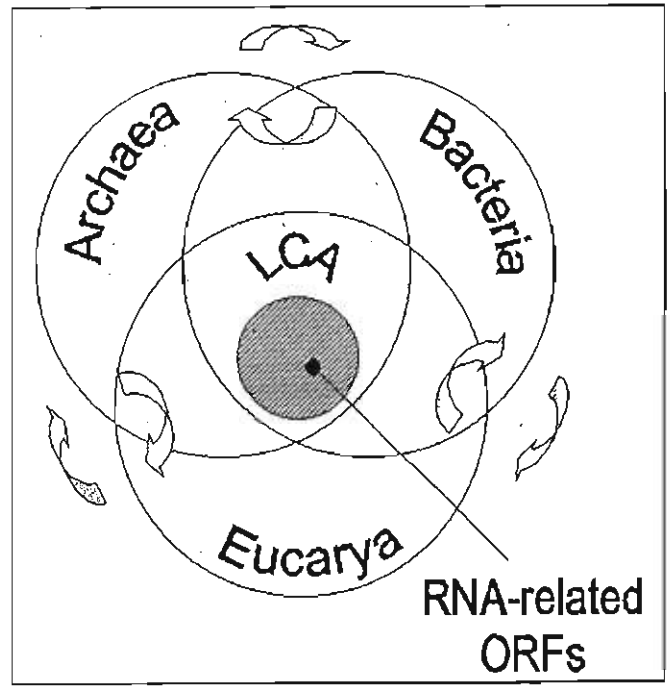


Figure 1. Venn diagram scheme indicating the most parsimonious characterization of the gene complement of the last common ancestor (LCA). The inner subset corresponds to highly conserved RNA metabolism-related sequences (see text), and the arrows indicate horizontal gene transfer (HGT) events, which is some cases involved endosymbiotic events.

of the universal ancestor and the evolutionary processes that shaped it. In this chapter we survey some the difficulties encountered in the description of the last common ancestor, and summarize ongoing discussions on its nature, reviewing briefly how this information can be used to infer earlier steps in biological evolution.

Universal Phylogenies and the Search for the Cenancestor

The traits shared by all known living beings are far to numerous and complex to assume that they evolved independently. Minor differences in the basic molecular processes of the three main cell lines can be distinguished, but all known organisms share the same genetic code and 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 strongly selected against.

The variations that are observed in extant species can be easily explained as the outcome of divergent processes from an ancestral life form, *fons et origo* of all contemporary organisms. Of course, no geological remains will bear testimony of its existence, as the search for a fossil of the universal ancestor is bound to prove fruitless;[9] from a cladistic viewpoint, the LCA is merely an inferred inventory of features shared among extant organisms (Fig. 1), all of which are located at the tip of the branches of molecular phylogenies. However, if the term "universal distribution" is restricted to its most obvious sense, i.e., that of traits found in all completely sequenced genomes, then quite surprisingly the resulting repertoire is formed by relatively few features

and by incompletely represented biochemical processes.[10-12] Analysis of some of the most likely *a priori* candidates for strict universality, such as those sequences involved in DNA replication, have turned out to be not only poorly preserved but also, in some cases, of polyphyletic origin.[13,14]

In principle, determination of the evolutionary polarity of character states in universal phylogenies should lead to the recognition of the oldest phenotype. Accordingly, the most parsimonious characterization of the LCA can be achieved by proceeding backwards and summarizing the features of the oldest recognizable node of the universal cladogram, i.e., rooting of the universal tree would provide direct information on the nature of the LCA. However, the plesiomorphic traits found in the space defined by rRNA sequences allow the construction of topologies that specify branching relationships but not the position of the ancestral phenotype. This phylogenetic *cul-de-sac* was overcomed by Iwabe[15] et al and Gogarten[16] et al, who analyzed paralogous genes encoding (a) the two elongation factors (EF-G and EF-Tu) that assist in protein biosynthesis; and (b) the α and β 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 the archaea and the eukaryotic nucleocytoplasm on the other. By rooting deep phylogenies, ancient paralogous duplications provide the means to place the LCA in the universal tree. The conclusion that Bacteria are the oldest recognizable cellular phenotype, and the Archaea and Eucarya sister groups, is consistent with sequence analyses that have shown that the eukaryotic genes involved in the transcription/transcriptional molecular machineries are closer to their archaeal counterparts than to the eubacterial ones.[17-20]

However, the issue is far from solved, and has in fact been further complicated with the advent of genomics. For instance, Philippe and Forterre7 have argued that the bacterial root is a long-branch attraction artifact due to the mutational saturation of the more than 3.5 x 109 years-old marker sequences used in the construction of deep phylogenies. As part of an attempt to overcome this limitation, they have used a covarion-based phylogeny-building methodology that allows for rate variation of conserved sites under varying constraints, which led to cladograms with an eukaryotic root.

This conclusion has been enthusiastically embraced by Penny and Poole,[21] who in a number of publications have argued that the eucaryal fragmented genome (as indicated by the existence of separate chromosomes) and intranuclear RNA processing are evidence of the primitiveness of nucleated cell genomes, i.e., that the LUCA was a eukaryote. This hypothesis has been presented, albeit with somewhat different emphasis, by others.[5,22] However, there are several reasons that lead us to disagree with the proposal made by Penny and Poole (1999).[21] These include not only the presence of a widely distributed set of conserved set of DNA repair enzymes that are present in the three domains,[23] which may be interpreted as evidence of a cenancestral DNA genome, but also the following:

a. Although it is likely that the segmented genomes found among certain RNA viruses represent an evolutionary strategy to overcome the Eigen error threshold,[24] the average length of eukaryotic chromosomes is in general well above the size of each viral RNA genomic segment. Moreover, multiple chromosomes and other traits of eukaryotic genome architecture are not by themselves indicative of the antiquity of the eucaryal nucleocytoplasm; as summarized by Bendich and Drlica,[25] yeast telomerase-defficient cells are endowed with circular chromosomes, and other architectural features typical of eukaryotic genomes, such as polyploidy, linear chromosomes, and very large amounts of DNA have also been described in different prokaryotic species;
b. Intranuclear RNA processing is characterized self-splicing reactions of the immature RNA phosphodiester backbone. However, there is no conclusive evidence that intron self-splicing and ribozyme-mediated RNA processing are truly primordial activities: ribozymes with ligase activity and self-cleaving RNAs ribozymes are extremely abundant, and distinct mechanisms by which editing can occur have been described.[26] These observations demonstrate the polyphyly of ribozyme-mediated processes, and imply that not all of them are truly vestigial activities, i.e., not all eukaryotic RNA processing is a relic of a preDNA/protein world but may be in fact a later development; and
c. Cholesterol and related sterols are hallmarks of nucleated cells. This is true even of anaerobic, amitochondrial ancient species such as Giardia lamblia, where cholesterol is furnished by its host. Although eucaryal genome architecture and sterol biosynthesis are independent features, the highly flexible eukaryotic internal membrane system which underlies the endoplasmic reticulum and the nuclear membrane, which defines the environment where RNA processing takes place, would not be possible in the absence of cholesterol. Since the anaerobic biosynthesis of cholesterol is not feasible, this suggests that, in contrast to prokaryotes, eukaryotes could have not appeared until free oxygen accumulated in the Precambrian environment. This strongly diminishes the likelihood of a eucaryal-like LCA.

Progenote Swarms or Prokaryote-Like Cenancestors

Analysis of an increasingly large number of genes and genomes has revealed major discrepancies with the topology of rRNA trees. As summarized by Brown (this volume) very often these differences have been interpreted as evidence of horizontal gene transfer (HGT) events between different species, questioning the feasibility of the reconstruction and proper understanding of early biological history.[27] Depending on their different advocates, a wide spectrum of mix-and-match recombination processes have been described, ranging from the lateral transfer of few genes, to cell fusion events involving organisms from different domains. There is clear evidence that genomes have a mosaic-like nature whose components come from a wide variety of sources.[28] However, not all sequences have the same likelihood of undergoing horizontal transfer events. Proteomic analysis of functional groups of sequences suggest that while housekeeping genes are more prone to HGT, genes involved in transcription, translation, and related process are less likely to be transferred.[29] On the one hand, these observations help to understand the peculiarities of metabolic gene phylogenies[30] and, on the other, the fact that even rRNA can undergo HGT events[31,32] supports contentions of a web-like pattern of early biological history.[27]

Reticulate phylogenies greatly complicate the inference of cenancestral traits. Driven in part by the impact of lateral gene acquisition, as revealed by the discrepancies of different gene phylogenies with the rRNA tree, and in part by the surprising complexity of the universal ancestor as suggested by direct backtrack characterizations of the oldest node of universal cladograms, Woese[33] proposed that the LCA was not a single organism, but rather a highly diverse population of metabolically complementary, cellular progenotes endowed with multiple, small linear chromosome-like genomes that benefited from massive multidirectional horizontal transfer events. According to this model, the essential features of translation and the development of metabolic pathways took place before the earliest branching event, but what led to the three domains was not a single ancestral lineage, but a rapidly differentiating community of genetic entities. This communal ancestor occupied as a whole the node located at the bottom of the universal tree, in which the decrease of sequence exchange and increasing genetic isolation would eventually lead to the observed tripartite division of the biosphere.

We have an alternative opinion. The genetic entities that formed the communal ancestor proposed by Woese[33] may have been extremely diverse, but an indication of their ultimate monophyletic origin from a sole progenitor is provided by universally distributed features such as the genetic code and the gene expression machinery. Did this hypothetical communal progenote ancestor diverged sharply into the three domains soon after the appearance of the code and the establishment of translation? Not necessarily. The origin of the mutant sequences ancestral to those found in all extant species, and the divergence of the Bacteria, Archaea, and Eucarya were not synchronous events, i.e., the separation of the primary domains took place later, perhaps even much later, than the appearance of the genetic components of their last common ancestor. Moreover, 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. What we have been calling the root of the universal tree

is in fact the tip of its trunk: inventories of LCA genes include sequences that originated in different precenancestral epochs.[11,34-36]

Universal gene-based phylogenies ultimately reach a single universal entity, but the bacterial-like LCA, which we favor, was not alone. Company must have been kept by its siblings, a population of entities similar to it that existed throughout the same period. They may have not survived, but some of their genes did if they became integrated via lateral transfer into the LCA genome. The cenancestor is one of the last evolutionary outcomes of a tree trunk of unknown length, during which the history of a long but not necessarily slow[37] series of ancestral events including lateral gene transfer, gene losses, and paralogous duplications probably played a significant role in the accretion of complex genomes.[3,38,39]

It is currently difficult to propose a unifying hypothesis. However, the scheme outlined here is supported by gene content trees, which exhibit an excellent broad-level agreement with rRNA-based phylogenies.[40-42] Such trees are not cladograms but phenograms, i.e., they are merely hierarchical representations of similarities and differences in gene content, where the presence or absence of a sequence is counted as a character. Since different lineages evolve at different rates, such overall similarity may be an equivocal indicator of genealogical relationships. Nevertheless, these trees are consistent with rRNA phylogenies, and do not support the hypothesis of massive HGT between distant species. Comparisons of combined orthologous protein data sets that exclude sequences that may have undergone lateral transfer are equally consistent with rRNA-based trees.[12] The robustness exhibited by these different methodologies indicate that although lateral gene transfer has played major role in cellular evolution, massive lateral transfer events between distant groups has not taken place. This suggests not only that the early history of life has not been completely obliterated by lateral transfer of genes,[43] but also that the role of reticulate evolution in defining the LCA as a progenote swarm may have been overstated.

The Nature of the Cenancestral Genome: DNA or RNA

Since all extant cells are endowed with DNA genomes, the most parsimonious conclusion is that this genetic polymer was already present in the cenancestral population. Woese[44,45] has suggested otherwise, arguing for a progenote-like universal ancestor endowed with a rapidly evolving genome formed by disaggregated, small-sized RNA molecules. This possibility was supported at least in part by the findings of Mushegian and Koonin,[46] who suggested that the absence of eucaryal or archaeal homologs of key components of DNA replication and nucleotide biosynthesis in the minimal gene set which resulted from the comparison of the Haemophilus influenzae and Mycoplasma genitalium genomes indicated that the cenancestor had used RNA as genetic polymer. Such conclusion is weakened by the limited data-set analyzed, which consisted of only two parasitic bacterial genomes that have undergone extensive polyphyletic gene losses.[47] In a subsequent publication, however, Koonin and his collaborators analyzed a large set of primases, replicative polymerases, and other proteins involved in DNA replication, and have suggested an alternative scheme with a hybrid RNA/DNA cenancestral genetic system whose complex replication cycle involving reverse transcription.[48]

There are indeed manifold indications that RNA genomes existed during early stages of cellular evolution[49] but, as argued below, it is likely that double-stranded DNA genomes had become firmly established prior to the divergence of the three primary domains. The major arguments supporting this possibility are:

a. In sharp contrast with other energetically favorable biochemical reactions (such as phosphodiester backbone hydrolysis or the transfer of amino groups), the direct removal of the oxygen from the 2'-C ribonucleoride pentose ring to form the corresponding deoxy-equivalents is a thermodynamically much less-favored reaction, considerably reducing the likelihood of multiple, independent origins of biological ribonucleoride reduction;
b. demonstration of the monophyletic origin of ribonucleotide reductases (RNR) is greatly complicated by their highly divergent primary sequences and the different mechanisms by which they generate the substrate 3'-radical species required for the removal of the 2'-OH group. However, sequence analysis and biochemical characterization of archaebacterial RNRs have shown their similarities with their eubacterial and eukaryotic counterparts, confirming their ultimate monophyletic origin;[50-52] and
c. sequence similarities shared by many ancient, large proteins found in all three domains suggest that considerable fidelity existed in the operative genetic system of their common ancestor, but such fidelity is unlikely to be found in RNA-based genetic systems.[3]

While accepting a DNA component in the LCA genome, Leipe et al[48] have underlined the highly divergent character of the main components of the (eu)bacterial replication machinery when compared with their archaeal/eukaryotic counterpart. Although it is possible to recognize the evolutionary relatedness of various orthologous DNA informational proteins (i.e., ATP-dependent clamp loader proteins, topoisomerases, gyrases, and 5'-3' exonucleases) across the entire phylogenetic spectrum,[14,13,48] comparative proteome analysis has shown that (eu)bacterial replicative polymerases and primases lack homologues in the two other primary kingdoms. As argued by Leipe et al[48] these observations can be explained by assuming a dual, independent origin of the DNA replication machineries of the Bacteria, on the one hand, and of the Archaea/Eucaryal on the other. Further convolutions have been added to the plot by Forterre,[53] who argued that the evolutionary separation between the replication components resulted from the nonorthologous displacement by rapidly evolving viral or plasmid-encoded gene products soon after the divergence of the three primary domains, as well by Villarreal and DeFilippis,[54] who in a similar vein have suggested a viral origin of nucleated cell DNA polymerases.

Evolution of enzymes in biological systems often involves the acquisition of new catalytic or binding properties by an existing protein scaffold. This has not been the case for the major types of polymerases, as shown by the identification of several nonhomologous classes of polymerases: primases, DNA polymerases, DNA-dependent RNA polymerases, replicases, and poly(A) polymerase, among others.[55] The polyphyletic origin of different polymerases and the large sequence space explored by DNA polymerases probably reflect the energetically favorable character of the enzyme-mediated synthesis of phosphodiester bonds in the presence of a template.

All DNA polymerases whose tertiary structure has been determined share a common overall architectural feature comparable to a right hand shape. Detailed analysis of the three-dimensional structures of the pol I, pol α, and reverse transcriptase families have shown that their palm subdomain, which catalyzes the formation of the phosphodiester bond, is homologous in all of them, while the fingers and thumb subdomains are different in all four of the families for which structures are known.[55] Homologous palm subdomains have also been identified in the viral T7 DNA- and RNA polymerases,[56] indicating that it can catalyze the template-dependent polymerization of ribo- and of deoxyribonucleotides (Fig. 2). More recently, the construction of a database of aligned crystal structures of DNA pol families A and B has allowed the precise identification of the conserved motifs described by Poch et al[57] in the catalytic palm subdomain of DNA polymerase families A(I) and B(II), and leading to its identification in the eukaryotic DNA polymerase δ and ζ subunits.[58]

As summarized by Forterre,[53] a nucleic acid replication enzymatic machinery requires, at the very least, a replicase, a primase, and a helicase, which are currently described as nonorthologues between the bacterial and the archaea/eukaryotic branches. Given the central role that is assigned to nucleic acid replication in mainstream definitions of life,[59] the lack of conservation and polyphyly of several of its key enzymatic components is somewhat surprising. However, the ample phylogenetic distribution of the catalytic palm subdomain and the relative template- and substrate specificities of polymerases[60,61] and helicases, suggest an explanation for the evolution of the DNA replication machinery simpler that those advocated by Leipe et al,[48] Forterre,[53] and Villarreal and DeFilippis.[54]

Our scheme assumes that the conserved palm subdomain described above is one of the oldest recognizable components of an ancestral cellular polymerase that may have acted both as a replicase and a transcriptase during the RNA/protein world stage. Once the advent of

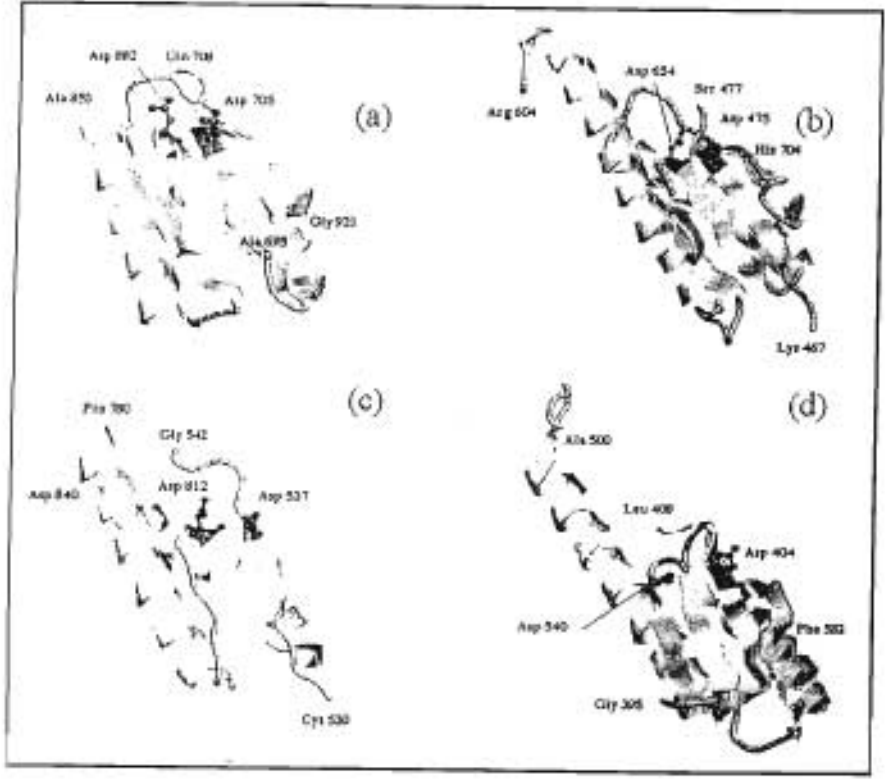


Figure 2. Conserved catalytic palm subdomain of the family I E. coli DNA polymerase I (a), the bacteriophage T7 DNA pol (b) and DNA-dependent RNA polymerase (c), and the family II Desulfurococcus DNA pol II (d). (Adapted from Brautigam and Steitz;[90] Cheetham et al;[91] Jeruzalmi and Steitz;[56] and Zhao et al.[92])

double-stranded DNA took place, relatively few mutations would have been required for the evolution of this RNA replicase into a DNA polymerase prior to the divergence of the three domains. Our hypothesis implies that this progenitor DNA polymerase was originally involved in the replication of the LCA genome, until its (eu)bacterial descendant (represented today by repair DNA pol II) underwent a nonorthologous displacement by the ancestor of the Escherichia coli replicative DNA pol III (DNA pol C) and its homologs. The structural homology of RNA- and DNA-helicase domains[62,63] suggest, on the other hand, the possibility of a nonspecific helicase inherited from the RNA/protein world that may have operated in unwinding double-stranded DNA until the evolution of the extant DNA helicases.

By analogy with the yeast and animal mitochondrial RNA polymerases, which play a dual role in transcription and in the initial priming required for DNA replication,[64] we propose that the original RNA polymerase described above catalyzed the formation of the RNA primer required for DNA replication. This hypothesis implies that extant bacterial and archaeal/eucaryotic primases are later independent evolutionary developments that displaced the cenancestral RNA polymerase from its primase function. As suggested above, this ancestral polymerase may have acted as a transcriptase during the RNA/protein stage, but the distribution of the highly conserved sequences of the oligomeric DNA-dependent RNA polymerase indicate that by the time the cenancestor diverged, a modern type of transcription had evolved. How this complex oligomeric transcriptase came into being can only be surmised at the time being.

Some Like It Very, Very Hot

The rooting of universal cladistic trees determines the directionality of evolutionary change and allows the recognition of ancestral from derived characters, i.e., primitive characters should appear in older, basal branches than do their derived counterparts. Determination of the rooting point of a tree normally imparts polarity to most or all characters.[65] It is, however, important to distinguish between ancient and primitive organisms. Organisms located near the root of universal rRNA-based trees are cladistically ancient, but they are not endowed with a primitive molecular genetic apparatus, nor seem to be more primitive in their metabolic abilities than their aerobic counterparts.

The situation is slightly different regarding the phylogenetic distribution of hyperthermophily, which appears to be a truly ancestral, primitive trait. 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.[66] Rooted universal phylogenies confirmed that hyperthermophiles are not randomly distributed in the universal tree, but are clearly located towards the lowest portion of molecular rRNA-based cladograms.[67] It is sometimes overlooked that the bacterial rooting of universal trees implies that hyperthermophilic bacteria such as Thermotoga and Aquifex are closer to the LCA than the oldest hyperthermophilic archaea, including the korearchaeota, which branch below the euryarchaeota/crenarchaeota split.[68] Some hyperthermophile sequences are displaced from their basal positions if molecular markers other than elongation factors or ATPase subunits are compared,[69] but the antiquity of hyperthermophiles appears to be well established,[45,67,70-72] and has received additional support from trees based on combined protein data sets from which sequences alignments that are candidates for HGT have been excluded.[12]

Backward extrapolation of the basal position of hyperthermophiles led not only to the hypothesis of a heat-loving LCA, but also of a high-temperature origin of life,[70] which according to some took place in extreme environments such as those found today in deep-sea vents[73] or in other sites in which mineral surfaces may have fueled the appearance of primordial chemoautolithotrophic biological systems.[74] However, all these views have been contested in one way or another, and are still open issues.[75] For instance, it is difficult to take for granted the possibility of hyperthermophilic universal ancestor endowed with a fragmented RNA genome proposed by Woese[33] with the extreme thermal fragility of RNA molecules.

The recognition that the deepest branches in rooted universal phylogenies are occupied by hyperthermophiles does not provides by itself conclusive proof of a heat-loving LCA. Analysis of the correlation of the optimal growth temperature of prokaryotes and the G+C nucleotide content of 40 rRNA sequences through a complex Markov model, has led Galtier et al[76] to conclude that the universal ancestor was a mesophile. If this is indeed the case, then the distribution of hyperthermophiles in rRNA-based phylogenies could be explained by: (a) lateral transfer of thermoadaptative traits;[77] (b) heat as a relic from early Archean high-temperature regimes that may have resulted from a severe impact regime;[78,79] (c) assuming that hyperthermophiles displaced older mesophiles when they adapted to lower temperatures, rather than being the sole survivors of an impact event.[80] It should be kept in mind, however, that since the time dimension is absent from the low G+C rRNA value inferred by Galtier et al,[76] it is possible that it corresponds not the cenancestor itself, but to one of its evolutionary predecessors, located along the trunk of the universal tree.

Thus, 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.[78,81,82] 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. Such adaptations may include histone-like proteins, RNA modifying enzymes, and reverse gyrase, a peculiar ATP-dependent enzyme that twists DNA into a positive supercoiled conformation.[81] 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,[67] which shows that at least five independent abandonment events of hyperthermophilic traits took place in widely separated branches of universal trees, one of which corresponds to the eukaryotic nucleocytoplasm.

Trimming the rRNA-Based Universal Trees

The conclusion that the LCA was a prokaryote-like organism similar to extant (eu)bacteria does not says much about its mode of energy acquisition and carbon sources. As summarized by Stetter,[67] the basal position of universal trees are occupied by heterotrophic and autotrophic hyperthermophiles, many of which live in sulphur-rich, extreme environments, with the deepest branches occupied by chemolithoautotrophs that have aerobic and anaerobic respiration. Direct extrapolation of these and other extremophile traits into the LCA has not been taken by granted by all. On the other hand, the irregular distribution of merabolic pathways and the large pool of sequences shared by extant species leads to a totipotent, phototroph LCA, unrealistically endowed with more biochemical attributes than some modern prokaryotes.[33,83] However, if multiple copies of every major gene family are assumed to have been already present in the LCA genome,[43] then the observed complex distribution patterns of bioenergetic and biosynthetic genes can be explained as the outcome of polyphyletic gene losses as the cenancestor descendants adapted to a wide variety of environments under different selection pressures.[38,39]

Although the timescale separating the LCA from the possible emergence of life is not a major problem given the rapid pace of prokaryotic evolution,[37] characterization of the cenancestral metabolic abilities can be hindered by several major problems. These include the horizontal acquisition of metabolic pathways, a possibility enhanced by likelihood of LGT of housekeeping genes,[29] and the fact that many open reading frames derived from complete genome sequencing projects remain unidentified (30 to 50% depending on the organism). It is possible that some of these ORFs correspond to rapidly evolving sequences encoding missing enzymes of incompletely reconstructed metabolic pathways.[84,85]

The inadequate biodiversity sampling that has shaped our current databases, which represent an extremely biased set of sequenced gene and genomes, also complicates our efforts. Quite understandably, medical and veterinarian interests have shaped the nature of extant genome databases from which many species are absent, perhaps even excluding members of every major biological group. Although clearly incomplete, the adequacy of fully sequenced genome databases for the reconstruction of ancestral states is probably greater than it is generally realized. There are, of course, many taxa we do not know about that are yet to be described. However, in spite of this strong limitation and of the extraordinary diversity of habitats and lifestyles, organisms share a surprising amount of enzymatic activities, metabolic routes, and basic biological functions, as reflected in genome replication, gene expression, and metabolic pathways. As the number of completely sequenced genomes has increased, the identification of new genes and functions common to all living beings has not expanded at the same rate (Fig. 3). The possibility that some of the enzymes of archaic pathways may have survived in unusual organisms suggest that considerable prudence should be exerted when attempting to describe the physiology of ancestral organisms. However, the sharp decline in the discovery of new, universally distributed sequences, which would correspond to an almost complete inventory of genes common to all living beings, should signal the approach to an almost complete universal set of genes (Fig. 3).

A more complicated issue is raised by the possibility that extant enzymes participated in alternative pathways which no longer exist or remain to be discovered,[86,87] a possibility that has begun to be explored by computer searches for alternative reaction pathways.[88] The discovery that carbamate kinase, which participates in fermentative ATP production, catalyzes the formation of carbamoyl phosphate in the archaea Pyrococcus furiosus and P. abyssi[89] shows that considerable attention should be given to the possibility that significant variations of the basic pathways may have existed in the past.

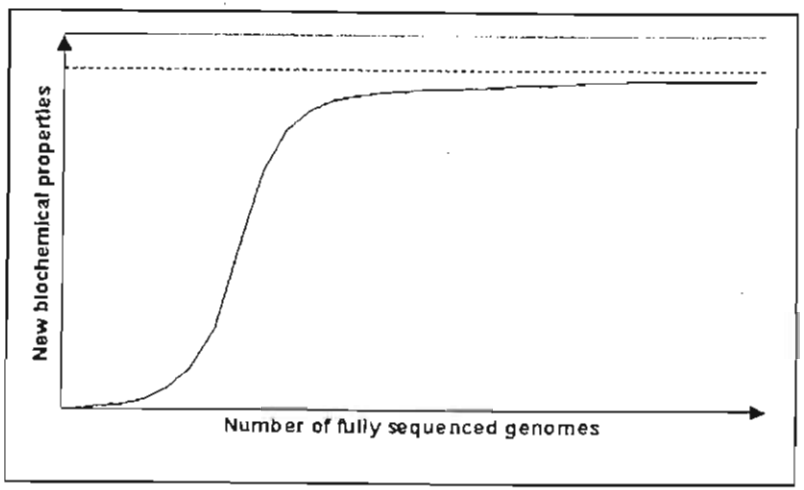


Figure 3. The discovery curve of universally distributed biochemical properties as inferred from proteome analysis. By analogy with the so called collector curve employed in ecology, it quantifies the analytic effort, assessed as the number of completely sequenced genomes analyzed, against the identification of additional biochemical features common to all cellular genomes.

Conclusions and Outlook

Understanding the biological attributes of the LCA and the evolutionary processes that shaped it has been defined as one of the major problems in evolutionary biology. This is not an overstatement, since it will assist in the comprehension of one of the major divergence events in the history of life, as well of paramount significance in understanding the different degrees of freedom that have been explored in the sequence- and three-dimensional spaces by the molecular components of central biological processes. Of course, current descriptions of the LCA are limited by the scant information available. It is hard, of course, to understand the evolutionary forces that acted on our distant ancestors, whose environments and detailed biological characteristics are forever beyond our ken.

Nevertheless, understanding the characteristics of the LCA may assist us in describing the entities that may have preceded it. Although we strongly favor an (eu)bacterial-like cenancestor, it is clear that biological evolution prior to the divergence of the three domains was not a continuous, unbroken chain of progressive transformation steadily proceeding towards the LCA. No evolutionary intermediate stages or ancient simplified version of the basic biological processes have been discovered in extant organisms. Did Woese's[33] differentiating communal progenote-like genetic entities exisred during this period?

Molecular cladistics and comparative genomics may provide clues to the genetic organization and biochemical complexity of the earlier entities from which the cenancesror evolved may be derived from the analysis of conserved ORFs. Genes involved in RNA metabolism, i.e., ORFs whose products synthesize, degrade, or interact with RNA, are among the most highly conserved sequences common to all known genomes, and provide insights inro an early stage in cell evolution during which RNA played a much more conspicuous biological role.[11,34,36] However, it is difficult to see how the applicability of molecular cladistics and comparative genomics can be extended beyond a threshold that corresponds ro 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 prob-

lem 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, it may be naive to attempt to describe the origin of life and the nature of the first living systems from the available rooted phylogenetic trees.

Acknowledgments

Work reported here was supported by project PAPIIT IN213598 (UNAM, Mexico). This paper was completed during a sabbatical leave of absence in which one of us (AL) enjoyed the hospitality of Dr. Ricardo Amils and his associates in the Universidad Autónoma de Madrid (Spain).

References

1. Woese CR, Kandler O, Wheelis ML et al. Towards a natural system of organisms, proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 1990; 87:4576-4579.
2. Woese CR, Fox GE. The concept of cellular evolution. J Mol Evol 1977; 10:1-6.
3. Lazcano A, Fox GE, Oró J. Life before DNA: The origin and early evolution of early Archean cells. In: R.P. Mortlock, ed. The Evolution of Metabolic Function. Boca Raton, FL: CRC Press, 1992:237-295.
4. Lazcano A. Cellular evolution during the Early Archean: What happened between progenote and the cenancestor? Microbiologia SEM 1995; 11:185-198.
5. Doolittle WF. The nature of the universal ancestor and the evolution of the proteome. Curr Opin Struct Biol 2000; 10:355-358.
6. Fitch WM, Upper K. The phylogeny of tRNA sequences provides evidence of ambiguity reduction in the origin of the genetic code. Cold Spring Harbor Symp Quant Biol 1987; 52:759-767.
7. Philippe H, Forterre P. The rooting of the universal tree of life is not reliable. J Mol Evol 1999; 49:509-523.
8. Line MA. The enigma of the origin of life and its timing. Microbiology 2002; 148:21-27.
9. Gee H. In search of deep time. New York: The Free Press, 1999.
10. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997; 278:631-637.
11. Tekaia F, Dujon B, Lazcano A. Comparative genomics: Products of the most conserved protein-encoding genes synthesize, degrade, or interact with RNA. Abstracts of the 9th ISSOL Meeting San Diego, California, USA: July 11-16, 1999:Abstract c46:53.
12. Brown JR, Douady CJ, Italia MJ et al. Universal trees based on large combined protein sequence datasets. Nat Genet 2001; 28:281-285.
13. Edgell RD, Doolittle WF. Archaea and the origin(s) of DNA replication proteins. Cell 1997; 89:995-998.
14. Olsen G, Woese CR. Archaeal genomics: an overview. Cell 1997; 89:991-994.
15. Iwabe N, Kuma K, Hasegawa M et al. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 1989; 86:9355-9359.
16. Gogarten JP, Kibak H, Dittrich P et al. Evolution of the vacuolar H+-ATPase, implications for the origin of eukayotes. Proc Natl Acad Sci USA 1989; 86:6661-6665.
17. Ouzonis C, Sander C. TFIIB, an evolutionary link between the transcription machineries of archaebacteria and eukaryotes. Cell 1992; 71:189-190.
18. Kaine BP, Mehr IJ, Woese CR. The sequence, and its evolutionary implications, of a Thermococcus celer protein associated with transcription, Proc Natl Acad Sci USA 1994; 91:3854-3856.
19. Brown JR, Doolittle WF. Archaea and the prokaryote to eukaryote transition. Microbiol Mol Biol Rev 1997; 61:456-502.
20. Koonin EV, Mushegian AR, Galperin MY et al. Comparisonof archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. Mol Microbiol 1997; 25:619-637.
21. Penny D, Poole A. The nature of the universal common ancestor. Curr Opin Genet Dev 1999; 9:672-677.
22. Harumen H, Fedorov A. The origin of the eukaryotic cell: a genomic investigation. Proc Natl Acad Sci USA 2002; 99:1420-1425.
23. Eisen JA, Hanawalt PC. A phylogenomic study of DNA repair genes, proteins, and processes. Mutation Res 1999; 435:171-213.
24. Reanney DC. Genetic error and genome design. Cold Spring Harbor Symp Quant Biol 1987; 52:751-757.
25. Bendich AJ, Drlica K. Prokaryotic and eukaryotic chromosomes: What's the difference. BioEssays 2000; 22:481-486.
26. Gesteland RF, Atkins JF, eds. The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World Cold Spring Harbor, New York:,Cold Spring Harbor Laboratory Press, 1993.
27. Doolittle WF. Phylogenetic classification and the universal tree. Science 1999; 284:2124-2128.
28. Ochman H, Lawrence JGM, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature 2000; 405:299-304.
29. Rivera MC, Jain R, Moore JE et al. Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci USA 1998; 95:6239-6244.
30. Alifano P, Fani R, Lió P et al. Histidine biosynthetic pathway and genes: structure, regulation, and evolution. Microbiol Rev 1996; 60:44-69.
31. Perez-Luz S, Rodríguez-Valera F, Lan R et al. Variation of the ribosomal operon 16S-23S gene spacer region in representatives of Salmonella enterica subspecies. J Bacteriol 1998; 180:2144-2151.
32. Yap WH, Zhang Z, Wang Y. Distinct types of rRNA operons exist in the genome of the actinomycete Thermonospora chromogena and evidence for horizontal transfer of an entire rRNA operon. J Bacteriol 1999; 181:5201-5209.
33. Woese CR. The universal ancestor. Proc Natl Acad Sci USA 1998; 95:6854-6859.
34. Delaye L, Lazcano A. RNA-binding peptides as molecular fossils. In: Chela-Flores J, Lemerchand G, Oró J, eds. Origins from the Big-Bang to Biology: Proceedings of the First Ibero-American School of Astrobiology (Dordrecht): Kluwer Academic Publishers, 2000:285-288.
35. Lazcano Araujo A. El último ancestro común. In: Martínez Romero E, Martínez Romero Y, eds. Microbios en Línea. UNAM, México, 2001:421-429.
36. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acid Res 2002; 30:1427-1464.
37. Lazcano A, Miller SL. How long did it take for life to begin and evolve to cyanobacteria? Jour Mol Evol 1994; 39:546-554.
38. Castresana J. Comparative genomics and bioenergetics. Biochem Biophys Acta 2001; 1506:147-162.
39. Snel B, Bork P, Huynen MA. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res 2002; 12:17-25.
40. Fitz-Gibbon ST, House CH. Whole genome-based phylogenetic analysis of free-living organisms. Nucleic Acids Res 1999; 27:4218-4222.
41. Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. Nat Genet 1999; 21:108-110.
42. Tekaia F, Lazcano A, Dujon B. The genomic tree as revealed from whole proteome comparisons. Genome Res 1999b; 9:550-557.
43. Glansdorff N. About the last common ancestor, the universal life-tree and lateral gene transfer: A reappraisal. Mol Microbiol 2000; 38:177-185.
44. Woese CR. The primary lines of descent and the universal ancestor. In: Bendall DS, ed. Evolution from Molecules to Men. Cambridge: Cambridge University Press, 1983:209-233.
45. Woese CR. Bacterial evolution. Microbiol Reviews 1987; 51:221-271.
46. Mushegian AR, Koonin EV. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 1996; 93:10268-10273.
47. Becerra A, Islas S, Leguina JI et al. Polyphyletic gene losses can bias backtrack characterizations of the cenancestor. J Mol Evol1997; 45:115-118.
48. Leipe DD, Aravind L, Koonin EV. Did DNA replication evolve twice independently? Nucleic Acid Res 1999; 27:3389-3401.
49. Lazcano A, Guerrero R, Margulis L et al. The evolutionary transition from RNA to DNA in early cells. J Mol Evol 1988a; 27:283-290.
50. Tauer A, Bennet SA. The B12-dependent ribonucleotide reductase from the archaebacterium Thermoplasma acidophila: An evolutionary solution to the ribonucleotide reductase conundrum. Proc Natl Acad Sci USA 1996; 94:53-58.
51. Riera J, Robb FT, Weiss R et al. Ribonucleotide reductase in the archaeon Pyrococcus furiosus: a critical enzyme in the evolution of DNA genomes. Proc Natl Acad Sci USA 1997; 94:475-478.
52. Freeland SJ, Knight RD, Landweber LF. Do proteins predate DNA? Science 1999 286:690-692.
53. Forterre P. Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins Mol Microbiol1999; 33:457-465.
54. Villarreal LP, DeFilippis VR. A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. J Virol 2000; 74:7079-7084.

55. Steitz TA. DNA polymerases: structural diversity and common mechanisms. J Biol Chem1999; 274:17395-17398.
56. Jeruzalmi D, Steitz TA. Structure of T7 RNA polymerase complexed to the transcriptional inhibitor T7 lysozyme. EMBO J 1998 17:4101-4113.
57. Poch O, Sauvaget I, Delarue M et al. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. EMBO J 1989; 8:3867-3874.
58. Delaye L, Vázquez H, Lazcano A. The cenancestor and its contemporary biological relics: the case of nucleic acid polymerases. In: Chela-Flores J, Owen T, Raulin F, eds. First steps in the origin of life in the Universe. Dordrecht: Kluwer Academic Publisher, 2001:223-230.
59. Koshland DE. The seven pillars of life. Science 2002; 295:2215-2216.
60. Lazcano A, Fastag J, Gariglio P et al. On the early evolution of RNA polymerase. J Mol Evol 1988b; 27:365-37.
61. Siegel RW, Bellon L, Beigelman L et al. Use of DNA, RNA, and chimeric templates by a viral RNA-dependent RNA polymerase: evolutionary implications for the transition from the RNA to the DNA world. J Virol 1999; 73:6424-6429.
62. Theis K, Chen PJ, Skorvaga M et al. Crystal structure of UvrB, a DNA helicase adapted for nucleotide excision repair. EMBO J 1999; 18:6899-6907.
63. Caruthers JM, Johnson ER, McKay DB. Crystal structure of yeast initiation factor 4A, a DEAD-box RNA helicase. Proc Natl Acad Sci USA 2000; 97:13080-13085.
64. Schinkel AH, Tabak HF. Mitochondrial RNA polymerase: dual role in transcription and replication. Trends Genet 1989; 5:149-154.
65. Scotland RW. Character coding. In: Florey PL, Humphries CJ, Kitching IL et al, eds. Cladistics: A practical course in systematics. Oxford: Clarendon Press, 1992:14-43.
66. Achenbach-Richter L, Gupta R, Stetter KO et al. Were the original eubacteria thermophiles? System Appl Microbiol 1987; 9:34-39.
67. Stetter KO. The lesson of archaebacteria. In: Bengtson S, ed. Early Life on Earth, Nobel Symposium No. 84. New York: Columbia University Press, 1994:114-122.
68. Barns SM, Delwiche CF, Palmer JD et al. Perspectives on archaeal diversity, thermophily and monophily from environmental rRNA sequences. Proc Natl Acad Sci USA 1996; 92:2441-2445.
69. Forterre P. A hot topic: The origin of hyperthermophiles. Cell 1996; 85:789-792.
70. Pace N. Origin of life —facing up to the physical setting. Cell 1991; 65:531-533.
71. Di Giulio M. The universal ancestor lived in a thermophilic or hyperthermophilic environment. J Theoret Biol 2000a; 203:203-213.
72. Di Giulio M. The stage of the genetic code structuring took place at a high temperature. Gene 2000b; 261:189-195.
73. Holm NG, ed. Marine Hydrothermal Systems and the Origin of Life. Dordrecht: Kluwer Academic Publ, 1992.
74. Wächtershäuser G. The case for the chemoautotrophic origins of life in an iron-sulfur world. Origins Life Evol Biosph 1990; 20:173-182.
75. Wiegel J, Adams MWW, eds. Thermophiles: The keys to molecular evolution and the origin of life. London: Taylor and Francis, 1998.
76. Galtier N, Tourasse N, Gouy M. A nonhyperthermophilic common ancestor to extant life forms. Science 1999; 283:220-221.
77. Forterre P, Bouthier de la Tour C, Philippe H et al. Reverse gyrase from hyperthermophiles: Probable transfer of a thermoadaptation trait from Archaea to Bacteria. Trends Genet 2000; 16:152-154.
78. Sleep NH, Zahnle KJ, Kastings JF et al. Annihilation of ecosystems by large asteroid impacts on the early Earth. Nature 1989; 342:139-142.
79. Gogarten-Boekels M, Hilario E, Gogarten JP. The effects of heavy meteoritic bombardments of the early evolution —the emergence of the three domains of life. Origins Life Evol Biosph 1995; 25:251-264.
80. Miller SL, Lazcano A. The origin of life —did it occur at high temperatures? J Mol Evol 1995; 41:689-692.
81. Confalonieri F, Elie C, Nadal M et al. Reverse gyrase, a helicase-like domain and a type I topoisomerase in the same polypeptide. Proc Natl Acad Sci USA 1993; 90:4753-4758.
82. Lazcano A. Biogenesis, some like it very hot. Science 1993; 260:1154-1155.
83. Olsen G, Woese CR. Lessons from an archeal genome: what are we learning from Methanococcus jannaschii? Trends Genet 1996; 12:377-379.
84. Bono H, Ogata H, Goto S et al. Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. Genome Res 1998; 8:203-210.
85. Velasco AM, Leguina JI, Lazcano A. Molecular evolution of the lysine biosynthetic pathways. J Mol Evol 2002; in press.
86. Zubay G. To what extent do biochemical pathways mimic prebiotic pathways? Chemtracts-Biochem. Mol Biol 1993; 4:317-323.
87. Becerra A, Lazcano A. The role of gene duplication in the evolution of purine nucleotide salvage pathways. Origins Life Evol Biosph 1998; 28:539-553.
88. Goto S, Bono H, Ogata H et al. Organizing and computing metabolic pathway data in terms of binary relations. In: Altman RB, Dunker K, Hunter L et al, eds. Pacific Symposium on Biocomputing. Singapore: World Scientific, 1996:175-186.
89. Alcántara C, Cervera J, Rubio V. Carbamate kinase can replace in vivo carbamoyl phosphate synthetase. Implications for the evolution of carbamoyl phosphate biosynthesis. FEBS Lett 2000; 484:261-264.
90. Brautigam CA, Steitz TA. Structural principles for the inhibition of the 3',5' exonuclease activity of Escherichia coli DNA polymerase I by phosphorothioates. J Mol Biol 1998; 277:363-377.
91. Cheetham GM, Jeruzalmi D, Steitz TA. Structural basis for initiation of transcription from an RNA polymerase-promoter complex. Nature 1999; 399:80-83.
92. Zhao Y, Jeruzalmi D, Moarefi I et al. Crystal structure of an archaeabacterial DNA polymerase. Struct Fold Des 1999; 7:1189-1199.

ELSEVIER

Physics, Mathematics
Computer Science & Astronomy

Professor A. Lazcano
UNAM
Facultad de Ciencias
Apdo. Postal 70-407
Cd. Universitaria
04510 Mexico D.F.
Mexico

Amsterdam, 22 March 2005

Dear Antonio,

Re: Physics of Life Reviews

I have pleasure in announcing the publication of *Physics of Life Reviews*, Vol. 2/1 which contains your article entitled "Prebiological evolution and the physics of the origin of life". Please find enclosed a copy of this issue; I hope you will like the presentation.
This copy is a complimentary one. You will receive your free offprints and any ordered ones from another department.

As (co-)author of this review, you are entitled to a fee of € 270.00. A cheque to this amount will be sent to you by Ms U. Bauwens of our Financial & Administrative Department, together with a copy of this letter. (This sum is to be divided equally between all the authors).
If you have not received the offprints or your fee within 8 weeks, please don't hesitate to contact me.

Thank you for the pleasant co-operation,
with best wishes,

Yvonne van Lieshout
Review journal co-ordinator

c.c.: Dr. L. Delaye
Ms U. Bauwens, Financial & Administrative Department
(phone +31 20 485 2024, fax +31 20 485 2722)

Available online at www.sciencedirect.com
SCIENCE DIRECT
Physics of Life Reviews 2 (2005) 47–64
PHYSICS of LIFE reviews
www.elsevier.com/locate/plrev

Prebiological evolution and the physics of the origin of life

Luis Delaye, Antonio Lazcano *

Facultad de Ciencias, UNAM, Apdo. Postal 70-407, Cd. Universitaria, 04510 Mexico, Mexico

Received 19 December 2004; accepted 23 December 2004

Available online 20 January 2005

Communicated by Di Mauro

Abstract

The basic tenet of the heterotrophic theory of the origin of life is that the maintenance and reproduction of the first living systems depended primarily on prebiotically synthesized organic molecules. It is unlikely that any single mechanism can account for the wide range of organic compounds that may have accumulated on the primitive Earth, suggesting that the prebiotic soup was formed by contributions from endogenous syntheses in reducing environments, metal sulphide-mediated synthesis in deep-sea vents, and exogenous sources such as comets, meteorites and interplanetary dust. The wide range of experimental conditions under which amino acids and nucleobases can be synthesized suggests that the abiotic syntheses of these monomers did not take place under a narrow range defined by highly selective reaction conditions, but rather under a wide variety of settings. The robustness of this type of chemistry is supported by the occurrence of most of these biochemical compounds in the Murchison meteorite. These results lend strong credence to the hypothesis that the emergence of life was the outcome of a long, but not necessarily slow, evolutionary processes. The origin of life may be best understood in terms of the dynamics and evolution of sets of chemical replicating entities. Whether such entities were enclosed within membranes is not yet clear, but given the prebiotic availability of amphiphilic compounds this may have well been the case. This scheme is not at odds with the theoretical models of self-organized emerging systems, but what is known of biology suggest that the essential traits of living systems could have not emerged in the absence of genetic material able to store, express and, upon replication, transmit to its progeny information capable of undergoing evolutionary change. How such genetic polymer first evolved is a central issue in origin-of-life studies.

* Corresponding author.
E-mail address: alar@correo.unam.mx (A. Lazcano).

1571-0645/$ – see front matter
doi:10.1016/j.plrev.2004.12.002

Keywords: Chemical evolution; Prebiotic synthesis; Prebiotic evolution; Organic compounds; Primitive terrestrial environment; Primitive replicators

Contents

1. Introduction

During a memorable 1939 lecture at the Royal Institution in London, wrote Max Perutz, the famous John D. Bernal stated that "all protein that we know now have been made by other proteins, and these in turn by others". How did such process got started? When Bernal repeated the same argument in a later discussion, Perutz [67] adds, "the physicist W. H. Bragg asked him where the first protein had come form. Instead of replying 'I do not know', Bernal skillfully sidestepped Bragg's awkward question".

Perutz does not writes how Bernal avoided the issue raised by Bragg, but the story reveals the strong scientific appeal that issues related to the nature of life and the origin of biological systems that has been brewing among physicists since the pre-DNA double helix times. Such trend, which was highlighted by Schrödinger's 1945 seminal book *What is Life?*, continues to this day, as shown by the manifold attempts to describe the emergence of life in terms of non-linear interactions and non-equilibrium constraints, the thermodynamics of irreversible processes, pattern formation, chaos, attractors, fractals and, more recently, complexity theory. Such approaches should be seen as open invitations to develop multidisciplinary research programs but, as noted by Fenchel [16], in some cases invocations to spontaneous generation appear to be lurking behind appeals to undefined "emergent properties" or "self-organizing principles" that are used as the basis for what many life scientists see as grand, sweeping generalizations with little relationship to actual biological phenomena.

The proposal of an heterotrophic origin of life is strongly supported by a number of rather successful prebiotic simulation experiments, as well as by the characterization of organic molecules of biochemical significance in meteorites and other extraterrestrial minor bodies rich in organic material. These results lend strong credence to the hypothesis that the emergence of life was the outcome of a long, but not necessarily slow, evolutionary processes. This conclusion is not at odds with the theoretical models of highly complex functionally organized systems favored nowadays by some physicists, but as of today none of these have provided manageable descriptions of the origin of life. Mainstream evolutionary

biologists and prebiotic chemists tend to be wary of explanations that assume that the emergence of life was the outcome of timeless mathematical or physical principles in which replication, selection, and adaptation play no role. Such lack of interest does not implies, of course, a belief that the natural processes that led to the first life forms were exempt from the constraints imposed by physics, or that explanations on the appearance of life should reduce themselves to the issue of the emergence of nucleic acids or their precursors. However, in spite of a number of mesmerizing theoretical and experimental analogs [39,94] what is known of biology suggest that the essential traits of living systems could have not emerged in the absence of genetic material able to store, express and, upon replication, transmit to its progeny information capable of undergoing evolutionary change. How such genetic polymer first evolved is one of the most basic questions in origin-of-life studies. Those involved in this field know they have plenty to be modest about, and they tend to be. For most life scientists, research on the origin of life should be addressed conjecturally, in an attempt to construct a coherent, non-teleological historical narrative with and inquiring and explanatory character [36]. How the current information on the distribution of abiotically synthesized organic compounds both in extraterrestrial environments and under simulated laboratory conditions can be combined with the idea of an RNA world is discussed in this review.

2. The physical setting of the origin of life

It is unlikely that the paleontological record will provide direct data on how life first appeared. There is no geological evidence of the environmental conditions on the Earth at the time of the origin of life, nor any fossil register of the evolutionary processes that preceded the appearance of the first cells. Direct information is lacking not only on the composition of the terrestrial atmosphere during the period of the origin of life, but also on the temperature, ocean pH values, and other general and local environmental conditions which may or may not have been important for the emergence of life. Moreover, the attributes of the first living organisms are unknown. They were probably simpler than any cell now alive, and may have lacked not only protein-based catalysis, but perhaps even the familiar genetic macromolecules, with their ribose-phosphate backbones. It is possible that the only property they shared with extant organisms was the structural complementarity between monomeric subunits of replicative genetic polymers able to transmit to its progeny information capable of undergoing evolutionary change. Hence, caution must be exercised in extrapolating molecular phyiogenies back into primordial times. Comparative genomics is a blooming field that has an extraordinary potential for our understanding early cellular evolution, but it cannot be applied to events prior to the evolution of protein biosynthesis. Older stages are not yet amenable to this type of analysis, and the organisms at the base of universal phylogenies are ancient species, not primitive unmodified microbes.

However, the traits shared by all known living beings are far to numerous and complex to assume that they evolved independently. Minor differences in the basic molecular processes of the three main cell lines can be distinguished, but all known organisms share the same genetic code and 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 provide direct evidence of the monophyletic origin of all known forms of life, while their variations can be easily explained as the outcome of divergent processes from an ancestral lifeform, *fons et origo* of all contemporary organisms. When and how did such ancestral form arise?

It is not possible to assign a precise chronology to the appearance of life. However, in the past few years estimates of the available time for this to occur have been considerable reduced. As shown by recent debates, determination of the biological origin of what have been considered the earliest traces of life is a rather contentious issue, an outcome of a scarce Archaean geological record with very few rocks older than 3.5 billion years. Those that remain have been so extensively altered by metamorphic processes that any direct life evidence of life predating this limit has apparently been largely obliterated, and most of the rocks which have been preserved have been metamorphosed to a considerable extent [89].

Nevertheless, there is evidence that life emerged on Earth as soon as it was possible to do so. It has been argued that the microstructures interpreted as cyanobacterial remnants in the 3.5 billion years-old Apex sediments of the Australian Warrawoona formation [78] could be the outcome of abiotic hydrothermal processes [6,24]. However, recent analysis of 3.4 billion years-old South African cherts indicates the existence of photosynthetic microbial mats in ancient marine environments [87]. Such rapid development speaks for the relatively short timescale required for the origin and early evolution of life on Earth, and suggests that the critical factor may have been the presence of liquid water, which became possible as soon as the planet's surface finally cooled down.

Water provides the medium for chemical reactions to take place, and the polymers required to carry out the central biological functions of replication and catalysis. How did it accumulate on the primitive Earth? The depletion of rare gases in the Earth's atmosphere compared to cosmic abundances shows that any primary atmosphere, if the planet ever had one, was rapidly lost [38]. Moreover, it is unlikely that water made its first appearance on Earth as a liquid. Soon after the Earth was formed, the release of the volatiles trapped within the accreting planetesimals very likely lead to a secondary atmosphere. Since current evidence suggests that the Earth's core formed when accretion was taking place, removal of metallic iron from the upper mantle must have lead to a highly reduced atmosphere of volcanic origin containing chemical species such as CH_4, NH_3 and H_2. Due to the high surface temperature, however, the bulk of the atmosphere would have consisted of superheated steam [38]. However, large impact events such as the one that lead to the Moon's formation would have eroded this primitive atmosphere, which would have been replaced by further outgassing events.

Moon-forming impacts must have been relatively rare, but it is generally accepted that during the latter stages of the accretion process the influx of comet-like bodies that originated from further out in the Solar System impacting the primitive Earth must have been considerable and could have led to the accumulation of significant amounts of water and other volatiles [64]. Cometary nuclei, which appear to be the most pristine materials surviving from the formation of the Solar System, may have supplied organic compounds that could played a role in the origin of life on Earth [1,7,8,64].

One reason for proposing an extraterrestrial origin of the components of the prebiotic soup is the CO_2-rich model of the primitive Earth's atmosphere [38]. Of course, the presence of an extraordinarily complex array of organic molecules in meteorites, comets, interplanetary dust and interstellar molecules argues for the robustness of organic chemistry in the Universe, but also raises the issue of their possible role in the origin of life. As noted below, it is likely that exogenous sources of organic compounds contributed to the synthesis of the primitive soup. The major sources of exogenous compounds would appear to be comets and dust, with asteroids and meteorites being minor contributors. Asteroids would have impacted the Earth frequently during the Hadean and early Archean, but the amount of organic material brought in would seem to be small, even if the asteroids are assumed to be Murchison meteorite-type object. Carbonaceous chondrites, a class of stony meteorites, are among the most primitive objects in the Solar System in terms of their elemental composition. The most extensively analyzed meteorites

for organic compounds include the Murchison and Murray meteorites, as well as the CI class Orgueil meteorite. The Murchison meteorite contains approximately 1.8% organic carbon, but most of this is a polymer, and there are only about 100 parts per million of amino acids (which represents, assuming a void volume of 10% and a density of approximately 2.0, 0.10 gm/kg meteorite, or 2.0×10^{-2} M of amino acids). The majority (up to 80%) of the soluble organic matter in meteorites is made up by polycyclic aromatic hydrocarbons (PAHs), followed by the carboxylic acids, the fullerenes and amino acids, which are about an order of magnitude less abundant [5]. The purines adenine, guanine, xanthine and hypoxanthine have also been detected, as well as the pyrimidine uracil in concentrations of 200–500 parts per billion in the CM chondrites Murchison and Murray and in the CI chondrite Orgueil [84,85, 90]. In addition, a variety of other nitrogen-heterocyclic compounds including pyridines, quinolines and isoquinolines were also identified in the Murchison meteorite [86], as well as sugar acids (polyols) [10] and membrane-forming lipidic compounds [12].

Comets are the most promising source of exogenous compounds [66]. At summarized elsewhere [3], it is reasonable to assume that the atmosphere that developed on the Earth over the period 4.4–3.8 billion years ago was essentially a mix of volatiles delivered by bodies such as cometary nuclei, combined with the products of outgassing processes from the interior of an already differentiated planet. This atmosphere was probably dominated by water steam until the surface temperatures dropped to ~100 °C (depending on the pressure), at which point water condensed out to form early oceans [93]. As the Earth had cooled down and the influx of myriads of comets and asteroids had settled down, the reduced chemical species, which were mainly supplied by volcanic outgassing and are very sensitive to UV radiation that penetrated through the atmosphere due to the lack of a protective ozone layer, were probably destroyed by photodissociation, although there might have been steady state equilibrium between these two processes that allowed a significant amount of these reduced species to be present in the atmosphere.

3. Primordial heterotrophy and the emergence of life

It is generally believed that after Louis Pasteur had disproved the spontaneous generation of microbes using his famous swam-necked flasks experiments, the discussion of life beginning's had been vanished to the realm of useless speculation. However, scientific literature of the first part of the 20th century shows the many attempts by major scientists to solve this problem. The list covers a rather wide range of explanations that go from the ideas of Pflüger on the role of hydrogen cyanide on the origin of life, to those of Svante Arrhenius on panspermia, and includes Leonard Troland's hypothesis of a primordial enzyme formed by chance events in the primitive ocean, Alfonso L. Herrera's sulfocyanic theory on the origin of cells, Harvey's 1924 suggestion of an heterotrophic origin in a high-temperature environment, and the provocative 1926 paper that Hermann J. Muller wrote on the abrupt, random formation of a single, mutable gene endowed with catalytic and autoreplicative properties [43].

In spite of their diversity, most of these explanations went unnoticed, in part because they were incomplete, speculative schemes largely devoid of direct evidence and not subject to fruitful experimental testing. Although some of these hypotheses considered life as an emergent feature of nature and attempted to understand its origin by introducing principles of historical explanation, the dominant view was that the first forms of life had been photosynthetic microbes endowed with the ability fix atmospheric CO_2 and to use it with water to synthesize organic compounds. A major scientific breakthrough occurred, however,

when Oparin [60] suggested a hetrototrophic origin of life that assumed that prior to the emergence of the first cells a prebiotic synthesis of organic compounds led to the accumulation of the primitive broth.

Such ideas were supported not only by the evidence of organic compounds in meteorites, but also by the striking 19th experimental demonstrations that biochemical compounds such as urea, alanine, and sugars could be formed under laboratory conditions, as had been demonstrated by Wöhler, Strecker and Butlerow, respectively. Oparin's proposal, which was based on his Darwinian credence in a gradual, slow evolution from the simple to the complex, stood in stood in sharp contrast with the then prevalent idea of an autotrophic origin of life. Since a heterotrophic anaerobe is metabolically simpler than an autotrophic one, the former would necessarily have evolved first. Thus, based on the simplicity and ubiquity of fermentative reactions, Oparin [60] suggested in a small booklet that the first organisms must have been heterotrophic bacteria that could not make their own food but obtained organic material present in the primitive milieu.

Careful reading of Oparin's [60] pamphlet shows that, in contrast to common belief, at first he did not assume an anoxic primitive atmosphere. In his original scenario he argued that while some carbides, i.e., carbon-metal compounds, extruded from the young Earth's interior would react with water vapor leading to hydrocarbons, others would be oxidized to form aldehydes, alcohols, and ketones (such as acetone). These molecules would then react among themselves and with NH_3 originating from the hydrolysis of nitrides (nitrogen-metals), to form "very complicated compounds", as Oparin wrote, from which proteins and carbohidrates would form. These ideas were further elaborated and refined in a more extensive book whose English translation was published in 1938 [61]. In this book Oparin's original proposal was revised, leading to the assumption of a highly reducing milieu in which iron carbides of geological origin would react with steam to form hydrocarbons. Their oxidation would yield alcohols, ketones, aldehydes, etc., that would then react with ammonia to form amines, amides and ammonium salts. The resulting protein-like compounds and other molecules would form a hot dilute soup, in which would aggregate to form colloidal systems such as coacervates, from which the first heterotrophic microbes evolved. Like many others at the time, Oparin did not address in his 1938 book the origin of nucleic acids, because their role in genetic processes was not even suspected. Because of this, inheritance of primordial genetic information was assumed by Oparin to be the result of growth and division in the coacervate drops he advocated as models of precellular systems.

4. Pyrite and the origin of life

Although by the late 19th century an autotrophic origin of life was part of mainstream biological thought, currently the best known alternative to the heterotrophic theory stems from the work of Wächtershäuser [92]. According to this hypothesis, life began with the appearance of an autocatalytic two-dimensional chemolithotrophic metabolic system based on the formation of the highly insoluble mineral pyrite. The synthesis in activated form of organic compounds such as amino acid derivatives, thioesters and keto acids is 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-organismal iron sulfide-based two-dimensional life, in which chemoautotrophic carbon fixation took place by a reductive citric acid cycle, or reverse Krebs cycle, of the type originally described for the photosynthetic green sulphur bacterium *Chlorobium limicola*. Molecular phylogenetic trees show that this mode of carbon fixation and its modifications (such as the reductive acetyl-CoA or the reductive

malonyl-CoA pathways) are found in anaerobic archaea and the most deeply divergent eubacteria, which has been interpreted as evidence of its primitive character [51]. This assumes, however, that the root of molecular phylogenetic trees can be extrapolated down to the very origin of life which, as argued below, is a rather contentious issue.

The reaction $FeS + H_2S = FeS_2 + H_2$ is a very favourable one. It has an irreversible, highly exergonic character with a standard free energy change $\Delta G^0 = -9.23$ kcal/mol, which corresponds to a reduction potential $E^0 = -620$ mV. Thus, the FeS/H_2S combination is a strong reducing agent, and has been shown to provide an efficient source of electrons for the reduction of organic compounds under mild conditions. Although pyrite-mediated CO_2 reduction to amino acids, purines and pyrimidines is yet to be achieved, the FeS/H_2S combination is a strong reducing agent that has been shown to reduce nitrate and acetylene, as well as to induce peptide-bonds that result from the activation of amino acids with carbon monoxide and (Ni, Fe)S [34,51]. Acetic acid and pyruvic acid have been synthesized from CO under simulated hydrothermal conditions in the presence of sulfide minerals [9,33]. However, the empirical support for Wächtershäuser's central tenets is meager. Life does not consist solely of metabolic cycles, and none of these experiments proves that enzymes and nucleic acids are the evolutionary outcome of multistep autocatalytic metabolic cycles surface-bounded to FeS/FeS_2 or some other mineral. As argued elsewhere [3], experiments using the FeS/H_2S combination are also compatible with a more general, modified model of the primitive soup in which pyrite formation is recognized as an important source of electrons for the reduction of organic compounds.

5. Prebiotic syntheses of amino acids and nucleobases: an optimistic assessment

The hypothesis that the first organisms were anaerobic heterotrophs is based on the assumption that abiotic organic compounds were a necessary precursor for the appearance of life. Experimental evidence in support of Oparin's proposal of chemical evolution came first from Harold C. Urey's laboratory, whom had been involved with the study of the origin of the Solar System and the chemical events associated with this process. Urey had also considered the origin of life in the context of his proposal of a highly reducing terrestrial atmosphere [88]. The first successful prebiotic amino acids synthesis was carried out with an electric discharge and a strongly reducing model atmosphere of CH_4, NH_3, H_2O, and H_2 [52]. The result of this experiment was a large yield of a racemic mixture of amino acids, together with hydroxy acids, short aliphatic acids, and urea. One of the surprising results of this experiment was that the products were not a random mixture of organic compounds; rather, a relatively small number of compounds were produced in substantial yield. Moreover, with a few exceptions, the compounds were of biochemical significance.

The mechanism of synthesis of the amino and hydroxy acids formed in the spark discharge experiment was investigated [52,53]. The presence of large quantities of hydrogen cyanide, aldehydes and ketones in the water flask, which were clearly derived from the methane, ammonia, and hydrogen originally included in the apparatus, showed not only that the amino acids were not formed directly in the electric discharge, but were the outcome of a Strecker-like synthesis that involved aqueous phase reactions of highly reactive intermediates. Detailed studies of the equilibrium and rate constants of these reactions demonstrated that both amino- and hydroxy acids can be synthesized at high dilutions of HCN and aldehydes in a simulated primitive ocean. The reaction rates depend on temperature, pH, HCN, NH_3, and aldehyde concentrations, and are rapid on a geological time scale; the half-lives for the hydrolysis of

the intermediate products in the reactions, amino- and hydroxy nitriles, are less than a thousand years at 0 °C, and there are no known slow steps [56].

A few years after the Miller experiment, Juan Oró, who had been studying the synthesis of amino acids from an aqueous solution of HCN and NH_3, reported the abiotic formation of adenine [63]. The synthesis is indeed remarkable. If concentrated solutions of ammonium cyanide are refluxed for a few days, adenine is obtained in up to 0.5% yield along with 4-aminoimidazole-5 carboxamide and the usual cyanide polymer [63,65]. This reaction, proceeds through the self-condensation of HCN to give diaminomaleonitrile, which according to [65], then reacts with formamidine to give adenine. Although in principle adenine may be consider as a mere pentamer of HCN, under dilute aqueous solutions adenine synthesis involves the formation and rearrangement of other precursors such as 2-cyano and 8-cyano adenine [91].

In the scheme suggested by Oró [63], the limiting step is the reaction of diaminomaleonitrile with formamidine, but as demonstrated by Ferris and Orgel [18], this can be bypassed by a two photon photochemical rearrangement of diaminomaleonitrile that proceeds readily with sunlight to give high yields of amino imidazole carbonitrile. An additional possibility is that tetramer formation may have occurred in the primitive Earth in an eutectic solution of HCN–H_2O, which could have existed in the polar regions of an Earth of the present average temperature. High yields of the HCN tetramer have been reported by cooling dilute cyanide solutions to temperatures between −10 and −30 °C for a few months [74]. Production of adenine by HCN polymerization is accelerated by the presence of formaldehyde and other aldehydes, which could have also been available on the prebiotic environment [91].

The prebiotic synthesis of guanine, the other major purine present in extant living systems, was first studied in an experimental setting involving high concentrations of a number of precursors, including ammonia [75]. It has been proposed that together with guanine, other purines including hypoxanthine, xanthine, and diaminopurine could have been produced in the primitive environment by variations of the adenine synthesis using aminoimidazole carbonitrile and aminoimidazole carboxamide [76]. A reexamination of the polymerization of concentrated NH_4CN solutions has shown that in addition to adenine, guanine is also produced at both −80 and −20 °C [47]. It is probable that most of the guanine obtained from the polymerization of NH_4CN is the product of diaminopurine, which reacts readily with water and undergoes a hydrolytic deamination to give guanine and some isoguanine. The yields of guanine in this "one-pot" reaction synthesis of purines yields are 10–40 less than those of adenine, guanine, and a simple set of amino acids dominated by glycine have also been detected in dilute solutions of NH_4CN which were kept frozen for 25 years at −20 and −78 °C, as well as in the aqueous products of spark discharge experiments from a reducing experiment frozen for 5 years at −20 °C [48]. Moreover, formamide, which is an hydrolytic product of HCN and is formed abundantly from the pyrolytic decomposition of HCN polymers, has been shown to react with HCN to produce adenine and formylpurine derivatives [72]. This reaction, which is enhanced in the presence of mineral catalyst, including silica, alumina, zeolite, and kaolin, is also known to yield cytosine and 4-hydroxypyrimidine [71,73].

The abiotic synthesis of cytosine in an aqueous phase from cyanoacetylene (HCC–CN) and cyanate (NCO^-) has been described [19,74]. Cyanoacetylene is abundantly produced by the action of a spark discharge on a mixture of methane and nitrogen, and cyanate can come from cyanogen (NCCN) or from the decomposition of urea (H_2N–CO–NH_2) However, since it is rapidly hydrolyzed to CO_2 and NH_3, the high concentrations of cyanate (>0.1 M) required in this reaction may be unrealistic.

Orotic acid, which is a biosynthetic precursor of uracil, was identified, albeit in low yields, among the hydrolytic products of hydrogen cyanide polymers [21]. On the other hand, the reaction of cyanoacetaldehyde, which is produced in high yields from the hydrolysis of cyanoacetylene, with urea, first studied

by Ferris et al. [20], produces no detectable levels of cytosine. However, when the same non-volatile compounds are concentrated in the laboratory modelling of "evaporating pond" conditions simulating primitive evaporating lagoons or pools on drying beaches on the early Earth, surprisingly high amounts of cytosine (>50%) are observed [68]. A related synthesis under evaporating conditions uses cyanoacetaldehyde with guanidine, which produce diaminopyrimidine [20] with very high yields [69]. Although it is unlikely that high amounts of diaminopyrimidine were present in the primitive Earth, both uracil and very low yields of cytosine result from its hydrolysis. The effectiveness of formamide as a prebiotic precursor of a mixture of both purines and pyrimidines in the presence of TiO_2 [71,73] suggest that in such environments simple minerals could have also promoted the synthesis of nucleobases in the primitive environment from hydrolytic products of HCN and other reactants that may have been easily available.

It is unlikely that high amounts of diaminopyrimidine were present in the prebiotic environment. However, a wide variety of other modified nucleic acid bases may have been available in the early Earth. The list includes isoguanine, which is a hydrolytic product of diaminopurine [47], as well as other modified purines which are the outcome of side reactions of both adenine and guanine with a number of different amines under the concentrated conditions of a drying pond [46], including a number of methylated bases.

6. How did organic compounds accumulate in the prebiotic soup?

The easiness of formation under reducing conditions ($CH_4 + N_2$, $NH_3 + H_2O$, or $CO_2 + H_2 + N_2$) in one-pot reactions of amino acids, purines, and pyrimidines strongly suggest these molecules were present in the prebiotic broth. Experimental evidence suggests that urea, alcohols, sugars formed by the non-enzymatic condensation of formaldehyde, a wide variety of aliphatic and aromatic hydrocarbons, urea, carboxylic acids, and branched and straight fatty acids, including some which are membrane-forming compounds, were also components of the primitive soup. The remarkable coincidence between the molecular constituents of living organisms and those synthesized in simulation experiments is too striking to be fortuitous, and the robustness of this type of chemistry is supported by the occurrence of most of these biochemical compounds in the 4.5×10^9 years-old Murchison carbonaceous meteorite, which also yields evidence of liquid water in its parent body [14].

These results are extremely encouraging, but it should be emphasized that the atmospheric composition that formed the basis of the Miller–Urey experiment is not considered today to be plausible by many researchers. Although it is generally agreed that free oxygen was absent from the primitive Earth, there is no agreement on the composition of the primitive atmosphere; opinions vary from strongly reducing ($CH_4 + N_2$, $NH_3 + H_2O$, or $CO_2 + H_2 + N_2$) to neutral ($CO_2 + N_2 + H_2O$). In general, non-reducing atmospheric models are favoured by planetary scientists, while prebiotic chemists lean towards more reducing conditions, under which the abiotic syntheses of amino acids, purines, pyrimidines, and other compounds are very efficient.

Although Miller and Urey originally rejected the idea of nonreducing conditions for the primitive atmosphere, a number of experiments were later on carried out in his laboratory using CO and CO_2 model atmospheres [77]. It was found that not only were the yields of the amino acids reduced, but that as the atmosphere became less reducing and more neutral, the yields of synthesized organic compounds decreased drastically and glycine was basically the only amino acid synthesized [56]. The presence of methane and ammonia appeared to be especially important for the formation of a diverse mixture of amino acids. The main problem in the synthesis of amino acids and other biologically relevant organic compounds with

nonreducing atmospheres is the formation of hydrogen cyanide (HCN), which is an intermediate in the Strecker pathway and an important precursor compound for the synthesis of nucleobases [21,63]. However, localized high concentrations of reduced gases may have existed around volcanic eruptions and in these localized environments reagents such as HCN, aldehydes and ketones could have been produced, which after dissolving into the primitive oceans could have taken part in the prebiotic synthesis of organic molecules.

Because of problems associated with the direct Miller–Urey type syntheses on the early Earth, different hypotheses for the abiotic synthesis of organic compounds has been proposed. One possibility that has been suggested resulted from the discovery of hydrothermal vents, which have been proposed as the site where prebiotic synthesis took place and life originated [11,32]. A further refinement of this hypothesis has led Everett Shock and his coworkers to argue, based on calculations of thermodynamic-based equilibria, that such environments favor the formation of compounds such as amino acids at high temperatures [81], especially in vents associated with off-axis systems [40].

As recognized long ago by Harvey [29], a major advantage of high temperatures is that the chemical reactions would go faster, and the primitive enzymes, once they appeared, could have been less efficient. However, the price paid is manifold: such high-temperature regimes would lead to (a) reduced concentrations of volatile intermediates, such as HCN, H_2CO and NH_3; (b) lower steady-state concentrations of prebiotic precursors like HCN, which at temperatures a little above 100 °C undergoes hydrolysis to formamide and formic acid and, in the presence of ammonia, to NH_4HCO_2; (c) instability of reactive chemical intermediates like amino nitriles ($RCHO(NH_2)CN$), which play a central role in the Strecker synthesis of amino acids; and (d) loss of organic compounds by thermal decomposition and diminished stability of genetic polymers [4,54,55].

Survival of nucleic acids is limited by the hydrolysis of phosphodiester bonds [50], and the stability of Watson–Crick helices (or their pre-RNA equivalents) is strongly diminished by high-temperatures. For an RNA-based biosphere the reduced thermal stability on the geologic timescale of ribose and other sugars is the worst problem [42], but the situation is equally bad for pyrimidines, purines and some amino acids. As reviewed elsewhere [55], the half-life of ribose at 100 °C and pH 7 is only 73 min, and other sugars (2-deoxyribose, ribose-5-phosphate, and ribose 2,4-biphosphate) have comparable half-lives [42]. The half-life for hydrolytic deamination of cytosine at 100 °C lies between 19 and 21 days [25,45,80], although at 100 °C the half-life of uracil is approximately 12 years [45]. At 100 °C the thermal stability of purines is also reduced: between 204 to 365 days for adenine [23,45,79], with comparable values for guanine [45]. These results imply that if the origin of life was sufficiently long, all the complex organic compounds in the ocean, whether derived from home-grown synthesis or from exogenous delivery, would be destroyed by passage through the hydrothermal vents. It is thus possible that hydrothermal vents are much more effective in regulating the concentration of critical organic molecules in the oceans rather than playing a significant role in their direct synthesis.

The difficulties involved with the endogenous synthesis of amino acids and nucleobases have led to the development of alternatives. It is likely, for instance, that geological sources of hydrogen, such as pyrite, may have been available; in the presence of ferrous iron, a sulfide ion (SH^-) would have been converted to a disulfide ion (S^{2-}), thereby releasing molecular hydrogen [51]. In addition, an analysis of Oro's 1961 suggestion on the role of cometary nuclei as sources of volatiles to the primitive Earth, has lead to the reassessment of the proposal that the exogenous delivery of organic matter by asteroids, comets and interplanetary dust particles (IDPs) could have played a significant role in the prebiotic accumulation of the compounds necessary for the origin of life [8]. If this idea is correct, impacts on the early Earth

could have led to devastating conditions which made it difficult for life to originate, but also delivered the raw material necessary for setting the stage for the origin of life. It is also possible that the impacts of iron-rich asteroids enhanced the reducing conditions, and that cometary collisions led to localized environments favouring organic synthesis. Based on what is known about prebiotic chemistry and meteorite composition, if the primitive Earth was non-reducing, then the organic compounds required must have been brought in by interplanetary dust particles, comets, and meteorites, a hypothesis that requires that a significant percentage of meteoritic amino acids and nucleobases could survive the high temperatures associated frictional heating during atmospheric entry, and become part of the primitive broth.

This eclectic view in which the prebiotic soup is formed by contributions from endogenous syntheses, extraterrestrial organic compounds delivered by comets and meteorites, and pyrite-mediated CO reduction does not contradict the heterotrophic theory. Even if the ultimate source of the organic molecules required for the origin of life turns out to be comets and meteorites, recognition of their extraterrestrial origin is not a rehabilitation of panspermia (e.g., the hypothesis that life existed elsewhere in the Universe and had been transferred from planet to planet, eventually gaining a foothold on the Earth), but an acknowledgement of the role of collisions in shaping the primitive terrestrial environment.

7. Prebiotic polymers and the RNA world

Regardless of its ultimate sources, the organic material that may have accumulated on the early Earth before life existed very likely consisted of a wide array of different types of compounds, including many of the simple compounds that play a major role in biochemistry today. How these abiotic organic constituents were assembled into polymers and then into the first living entities is currently one the most challenging areas of research in the study of the origin of life. There is no evidence of abiotically produced oligopeptides or oligonucleotides in the Murchison meteorite, but condensation reactions clearly took place in the primitive Earth. Synonymous terms like 'primitive soup', 'primordial broth', or 'Darwin's warm little pond' have led in some cases to major misunderstandings, including the simplistic image of a worldwide ocean, rich in self-replicating molecules and accompanied by all sorts of biochemical monomers. The term 'warm little pond', which has long been used for convenience, refers to parts of the hydrosphere where the accumulation and interaction of the products of prebiotic synthesis may have taken place. These include not only membrane-bound systems, but also oceanic sediments, intertidal zones, shallow ponds, fresh water lakes, lagoons undergoing wet-and-dry cycles, and eutectic environments (e.g., glacial ponds), where evaporation or other physicochemical mechanisms (such as the adherence of biochemical monomers to active surfaces) could have raised local concentrations and promoted polymerization [3].

Simple organic compounds dissolved in the primitive oceans or other bodies of water would need to be concentrated by some mechanism. Selective adsorption of molecules onto mineral surfaces could have promoted their polymerization, as suggested by laboratory simulations using a variety of simple compounds and activated monomers [17,30,31]. The potential importance of mineral assisted catalysis is demonstrated by the montmorillonite promoted polymerization of activated adenosine and uridine derivatives producing 25–50-mer oligonucleotides [17], the general length range considered necessary for primitive biochemical functions.

Since absorption onto surfaces involves weak non-covalent van der Waals interactions, the mineral based concentration process and subsequent polymerization would be most efficient at cool temperatures

[49,82]. However, as the length of polymers formed on mineral surfaces increases, they become more firmly bound to the mineral [22,62]. In order for these polymers to be involved in subsequent interactions with other polymers or monomers they would need to be released. This could be accomplished by warming the mineral although this would also tend to hydrolyze the absorbed polymers, or by concentrated salt solutions [30], a process that could take place in tidal regions during evaporation or freezing of seawater and that would have led to the release of polymers.

As summarized elsewhere [2], direct concentration of dilute solutions of monomers could also be accomplished by evaporation and by eutectic freezing of dilute aqueous solutions. The evaporation of tidal regions and the subsequent concentration of their organic constituents has been proposed in the synthesis of a variety simple organic molecules [57]. Salty brines may have also been important in the formation of peptides and perhaps other important biopolymers as well. As summarized by Rode [70], salt-induced peptide formation reaction may provide an abiotic route for the formation of peptides directly from amino acids in concentrated NaCl solutions containing Cu(II). Yields of di- and tripeptides in the 0.4–4% range have been reported using starting amino acid concentrations in the 40–50 mM range. Clay minerals such as montmorillonite apparently promote the reaction, which could have taken place in evaporating tidal pools and where the required concentrated salty brines would have been easily available. It has been shown that the freezing of dilute solutions of activated amino acids at $-20\,^{\circ}C$ yields peptides at higher yields than in experiments with highly concentrated solutions at 0 and 25 °C [49], and recent studies have shown that eutectic freezing is especially effective in the non-enzymatic synthesis of oligonucleotides [37].

It is very unlikely, however, that the RNA world would have arisen from such process. How the ubiquitous nucleic acid-based genetic system of extant life originated is one of the major unsolved problems in contemporary biology. The discovery of catalytically active RNA molecules gave considerable credibility to prior suggestions of that the first living organisms were largely based on ribozymes, an hypothetical stage called the RNA world [27,35]. This possibility is now widely accepted, but the chemical lability of RNA components suggests that this molecule was not a direct outcome of prebiotic evolution, but may have been one of the evolutionary outcomes of what are now referred to as pre-RNA worlds. However, the chemical nature of the first genetic polymers and the catalytic agents that may have formed the pre-RNA worlds that bridged the gap between the prebiotic broth and the RNA world are completely unknown and can only be surmised. Modified nucleic acid backbones have been synthesized, which either incorporate a different version of ribose or lack it altogether. Experiments on nucleic acid with hexoses instead of pentoses, and on pyranoses instead of furanose [15], suggests that a wide variety of informational polymers is possible, even when restricted to sugar phosphate backbones. One possibility that has not been explored is that the backbone of the original informational macromolecules may have been atactic (e.g., disordered) kerogen-like polymers such as those formed in some prebiotic simulations. There are other possible substitutes for ribose, including open chain, flexible molecules that lack asymmetric carbons. One of the most interesting chemical models for a possible precursor to RNA involves the so-called peptide nucleic acids (PNAs), which have a polypeptide-like backbone of achiral 2-aminoethyl-glycine, to which nucleic acid bases are attached by an acetic acid [58]. Such molecules form very stable complementary duplexes, both with themselves and with nucleic acids. Although they lack ribose, their functional groups are basically the same as in RNA, so they may also be endowed with catalytic activity.

Identification of adenine, guanine, uracil and other nucleobases in the Murchison meteorite supports the idea that these bases were present in the primitive environment. However, it is likely that other hetero-

cycles capable of forming hydrogen bonding were also available. The Watson–Crick base-pair geometry permits more than the four usual nucleobases, and simpler genetic polymers may not only have lacked the sugar-phosphate backbones, but may also have depended on alternative non-standard hydrogen bonding patterns. The search for experimental models of pre-RNA polymers will be rewarding but difficult; it requires the identification of potentially prebiotic components and the demonstration of their non-enzymatic template-dependent polymerization, as well as coherent hypothesis of how they may have catalyzed the transition to an RNA world.

8. The transition towards a DNA/RNA/protein world

RNA molecules adsorbed onto clays such as montmorillonite, which can catalyze the formation of RNA oligomers, can be encapsulated into fatty acid vesicles whose formation in turn is accelerated by the clay. By incorporating additional fatty acid micelles, these vesicles can grow and divide while still retaining a portion of their contents needed to support RNA replication. In this manner, some of the basic machinery needed for RNA self-replication could have been compartmentized into proto-type cells [28].

As hypothesized elsewhere [2], it is possible that by the time RNA-based life appeared on Earth, the supplies of simple abiotic organic compounds derived from the sources discussed above had been greatly diminished. Many of the components of the primordial soup may have been extensively converted into polymers including those associated with living entities, and the raw materials needed to sustain life may have been largely exhausted. This implies that the origin of simple metabolic-like pathways must have been in place in order ensure a supply in the ingredients needed to sustain the existence of the primitive living entities. In this case, some metabolic pathways needed to produce essential components required by primitive living entities were perhaps originally non-enzymatic or semi-enzymatic autocatalytic processes that later became fine tuned as ribozyme-mediated and protein-based enzymatic processed began to dominate [44].

All known organisms 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. No ancient incipient stages or evolutionary intermediate of these molecular structures are known but, as discussed below, in some cases the existence of graded intermediates can be deduced.

It is possible that the invention of protein synthesis and the encapsulation of reaction machinery needed for replication may have taken place during the RNA world [2]. The fact that RNA molecules are capable of performing by themselves all the reactions involved in peptide-bond formation suggests that protein biosynthesis evolved in an RNA world [95], i.e., that the first ribosome lacked proteins and was formed only by RNA. This possibility is supported by the crystallographic data that has shown that ribosome catalytic site where peptide bond formation takes place is composed solely of RNA [59]. As underlined by Kumar and Yarus [41], four of the central reactions involved in protein biosynthesis are catalyzed by ribozymes, and their complementary nature suggests suggestive that they may have first appeared in the RNA world. If this was the case, then the origin of a primitive nucleobase code used for protein biosynthesis had its origin in the RNA world although the bases used in the early code could have been different from the ones used today [68].

Clues to the genetic organization of primitive forms of translation are also provided by paralogous genes, which are sequences that diverge not through speciation but after a duplication event. For instance, the presence in all known cells of pairs of homologous genes encoding two elongation factors, which are GTP-dependent enzymes that assist in protein biosynthesis, provide evidence of the existence of a more primitive, less-regulated version of protein synthesis took place with only one elongation factor. In fact, the experimental evidence of in vitro translation systems with modified cationic concentrations lacking both elongation factors and other proteinic components [26,83] strongly supports the possibility of an older ancestral protein synthesis apparatus prior to the emergence of elongation factors.

The same is true of other enzymes. The high levels of genetic redundancy detected in all sequenced genomes imply not only that duplication has played a major role in the accretion of the complex genomes found in extant cells, but also that prior to the early duplication events revealed by the large protein families, simpler living systems existed which lacked the large sets of enzymes and the sophisticated regulatory abilities of contemporary organisms. The variations of traits common to extant species can be easily explained as the outcome of divergent processes from an ancestral lifeform that existed prior to the separation of the Bacteria, the Archaea and the Eucarya, i.e., the last common ancestor (LCA) or cenancestor. Universal gene-based phylogenies ultimately reach such single universal entity, which very likely was part of a population of similar entities existed throughout the same period. They may have not survived, but some of their genes did if they became integrated via lateral transfer into the LCA genome. As reviewed elsewhere [13], the cenancestor should be seen as one of the last evolutionary outcomes of a series of ancestral events including lateral gene transfer, gene losses, and paralogous duplications that took place before the separation of the three major cell lineages. Recognition that cellular genomes are historical documents recording at least part of past evolutionary events has allowed important insights into simpler biological systems that appear to have lacked DNA genomes, but that can be considered basically RNA/proteins cells far removed, if not time, at in complexity with respect the first living systems.

9. Conclusions

The understanding of the origin of life requires, wrote John D. Bernal several decades ago, requires a scientist with a deep knowledge in geology, chemistry, biology, astrophysics, theoretical physics, paleontology and philosophy. Since such polymaths are rare, we must either work in multidisciplinary teams or focus our attention in a particular issue within the framework and methodologies of one of these fields. It is true that physical and biological sciences should be seen as conceptual allies. However, Darwinism has successfully resisted reduction to physics, and the development of complex system systems dynamics advocated by many theoreticians has failed to provide manageable descriptions of the origin of life. The emergence of life may be best understood in terms of the dynamics and evolution of sets of chemical replicating entities. Whether such entities were enclosed within membranes is not yet clear, but given the prebiotic availability of amphiphilic compounds this may have well been the case.

As implied here, the most successful applications of physical sciences in the understanding of prebiotic evolution have resulted from those areas directly related to the reconstruction of the primitive environment, i.e., astrophysics, planetary sciences, and the like, as well as those pertaining the formation and stability of monomers and polymers of biochemical significance, including the physicochemistry of membrane-forming compounds. As emphasized in this review, the study of the emergence of life remains a chemical problem in which the transition from the results of purely physical and chemical processes

on the synthesis, accumulation and stability of simple biochemical monomers and polymers gave rise in still poorly understood processes to replicative systems capable of undergoing natural selection.

It is likely that no single mechanism can account for the wide range of organic compounds that may have accumulated on the primitive Earth, and that the prebiotic soup was formed by contributions from endogenous syntheses in a reducing atmosphere, metal sulfide-mediated synthesis in deep-sea vents, and exogenous sources such as comets, meteorites and interplanetary dust. Of course, not all prebiotic pathways are equally efficient, but the wide range of experimental conditions under which organic compounds can be synthesized demonstrates that prebiotic syntheses of the building blocks of life are robust, i.e., the abiotic reactions leading to them do not take place under a narrow range defined by highly selective reaction conditions, but rather under a wide variety of experimental settings. Our ideas on the prebiotic synthesis of organic compounds are based largely on experiments in model systems. The robustness of this type of chemistry is supported by the occurrence of most of these biochemical compounds in the Murchison meteorite. This makes it plausible, but does not proves, that similar synthesis took place on the primitive Earth. For all the uncertainties surrounding the emergence of life, it appears to us that the formation of the prebiotic soup is one of the most firmly established events that took place in the primitive Earth.

Thus, if convincing processes can be demonstrated for the origin of life on Earth, then it is reasonable to conclude that life is the natural outcome of an evolutionary process, and that it may have appeared elsewhere in the Universe. Although we do not know how the transition from the non-living to the living took place, most of the modern scenarios start out with relative simple organic molecules, now known to be widely distributed, which are readily synthesized, and hypothesized to undergo further evolutionary changes leading into self-maintaining, self-replicative systems from which the current DNA/protein-based biology resulted.

Acknowledgements

Support from UNAM-DGAPA Proyecto PAPIIT IN 111003-3 (UNAM, Mexico) is gratefully acknowledged.

References

[1] Anders E. Prebiotic organic matter from comets and asteroids. Nature 1989;342:255–7.
[2] Bada JL, Lazcano A. The origin of life—some like it hot, but not the first biomolecules. Science 2002;296:1982–3.
[3] Bada JL, Lazcano A. In: Ruse M, editor. The Harvard companion of evolution. Cambridge: Harvard Univ. Press; 2005. Submitted for publication.
[4] Bernhardt G, Luedmann HD, Jaenicke R, Koeing H, Stetter KO. Biomolecules are unstable under black smoker conditions. Naturwissenschaften 1984;71:583–6.
[5] Botta O, Bada JL. Extraterrestrial organic compounds in meteorites. Surv Geophys 2002;23:411–67.
[6] Brasier M, Green OR, Jephcoat AP, Kleppe AK, van Kranendonk MJ, Lindsay JF, Steele A, Grassineau NV. Questioning the evidence for Earth's earliest fossils. Nature 2002;416:76–9.
[7] Chyba CF. Impact delivery and erosion of planetary oceans in the early inner Solar System. Nature 1990;343:129–33.
[8] Chyba CF, Sagan C. Endogenous production, exogenous delivery, and impact-shock synthesis or organic compounds, an inventory for the origin of life. Nature 1992;355:125–32.

[9] Cody GD, Boctor NZ, Filley TR, Hazen RM, Scott JH, Sharma A, Yoder Jr HS. Primordial carbonylated iron–sulfur compounds and the synthesis of pyruvate. Science 2000;289:1337–40.
[10] Cooper G, Kimmich N, Belisle W, Sarinana J, Brabham K, Garrel L. Carbonaceous chondrites as a source of sugar related organic compounds for the early Earth. Nature 2001;414:879–83.
[11] Corliss JB, Baross JA, Hoffman SE. An hypothesis concerning the relationship between submarine hot springs and the origin of life on Earth. Oceanol Acta 1981;4(Suppl.):59–69.
[12] Deamer DW, Pashley RM. Amphiphilic components of the Murchison carbonaceous chondrite, surface properties and membrane formation. Origins Life Evol Biosph 1989;19:21–38.
[13] Delaye L, Becerra A, Lazcano A. In: Ribas de Pouplana L, editor. The genetic code and the origin of life. Landes Bioscience: Georgetown; 2004. p. 34–47.
[14] Ehrenfreund P, Irvine W, Becker L, Blank J, Brucato J, Colangeli L, Derenne S, Despois D, Dutrey A, Fraaije H, Lazcano A, Owen T, Robert F. Astrophysical and astrochemical insights into the origin of life. Reports Prog Phys 2002;65:1427–87.
[15] Eschenmoser A. Chemical etiology of nucleic acid structure. Science 1999;284:2118–24.
[16] Fenchel T. Origin and early evolution of life. Oxford: Oxford Univ. Press; 2002.
[17] Ferris JP. Montmorillonite catalysis of 30–50 mer oligonucleotides, laboratory demonstration of potential steps in the origin of the RNA World. Origins Life Evol Biosph 2002;32:311–32.
[18] Ferris JP, Orgel LE. An unusual photochemical rearrangement in the synthesis of adenine from hydrogen cyanide. J Am Chem Soc 1966;88:1074.
[19] Ferris JP, Sanchez RP, Orgel LE. Studies in prebiotic synthesis: III. Synthesis of pyrimidines from cyanoacetylene and cyanate. J Mol Biol 1968;33:693–704.
[20] Ferris JP, Zamek OS, Altbuch AM, Freiman H. Chemical evolution: XVIII. Synthesis of pyrimidines from guanidine and cyanoacetaldehyde. J Mol Evol 1974;3:301–9.
[21] Ferris JP, Joshi PD, Edelson EH, Lawless JG. HCN, a plausible source of purines, pyrimidines, and amino acids on the primitive Earth. J Mol Evol 1978;11:293–311.
[22] Ferris JP, Hill AR, Liu R, Orgel LE. Synthesis of long prebiotic oligomers on mineral surfaces. Nature 1996;381:59–61.
[23] Frick L, Mac Neela JP, Wolfenden R. Transition state stabilization by deaminases, rates of nonenzymatic hydrolysis of adenosine and cytidine. Bioorg Chem 1987;15:100–8.
[24] Garcia-Ruiz JM, Hyde ST, Carnerup AM, Christy AG, van Kranendonk MJ, Welham NJ. Self-assembled silica-carbonate structures and detection of ancient microfossils. Science 2003;302:1194–7.
[25] Garrett ER, Tsau J. Solvolyses of cytosine and cytidine. J Pharm Sci 1972;61:1052–61.
[26] Gavrilova LP, Kostiashkina OE, Koteliansky VE, Rutkevitch NM, Spirin AS. Factor-free, nonenzymic, and factor-dependent systems of translation of polyuridylic acid by *Escherichia coli* ribosomes. J Mol Biol 1976;101:537–52.
[27] Gilbert W. The RNA world. Nature 1986;319:618.
[28] Hanczyc MM, Fujikawa SM, Szostak JW. Experimental models of primitive cellular compartments, encapsulation, growth, and division. Science 2003;302:618–22.
[29] Harvey RB. Enzymes of thermal algae. Science 1924;60:481–2.
[30] Hill AR, Böhler C, Orgel LE. Polymerization on the rocks, negatively charged D/L amino acids. Origins Life Evol Biosph 2001;28:235–43.
[31] Holm NG, editor. Marine hydrothermal systems and the origin of life. Dordrecht: Kluwer Academic; 1992.
[32] Holm NG, Andersson EM. Abiotic synthesis of organic compounds under the conditions of submarine hydrothermal systems: a perspective. Planet Space Sci 1995;43:153–9.
[33] Huber C, Wächtershäuser G. Activated acetic acid by carbon fixation on, Fe, Ni, S under primordial conditions. Science 1997;276:245–7.
[34] Huber C, Wächtershäuser G. Peptides by activation of amino acids with CO on, Ni, Fe, S surfaces and implications for the origin of life. Science 1998;281:670–2.
[35] Joyce GF. The antiquity of RNA-based evolution. Nature 2002;418:214–21.
[36] Kamminga H. The origin of life on Earth, theory, history, and method. Uroboros 1991;1:95–110.
[37] Kanavarioti A, Monnard PA, Deamer DW. Eutectic phases in ice facilitate nonenzymatic nucleic acid synthesis. Astrobiology 2001;1:481.
[38] Kasting JF. Earth's early atmosphere. Science 1993;259:920–6.
[39] Kauffman SA. The origins of order: self organization and selection in evolution. New York: Oxford Univ. Press; 1993.

[40] Kelley DS, Karson JA, Blackman DK. An off-axis hydrothermal vent field near the Mid-Atlantic Ridge at 30° N. Nature 2001;241:145–9.
[41] Kumar RK, Yarus M. RNA-catalyzed amino acid activation. Biochemistry 2001;40:6998–7004.
[42] Larralde R, Robertson MP, Miller SL. Rates of decomposition of ribose and other sugars, implications for chemical evolution. Proc Natl Acad Sci USA 1995;92:8158–60.
[43] Lazcano A. A.I. Oparin, the man and his theory. In: Plogazov BF, Kurganov BI, Kritsky MS, Gladilin KL, editors. Evolutionary biochemistry and related areas of physicochemical biology. Moscow: Bach Institute of Biochemistry and ANKO Press; 1995. p. 49–56.
[44] Lazcano A, Miller SL. On the origin of metabolic pathways. J Mol Evol 1999;49:424–31.
[45] Levy M, Miller SL. The stability of the RNA bases: implications for the origins of life. Proc Natl Acad Sci USA 1998;95:7933–8.
[46] Levy M, Miller SL. The prebiotic synthesis of modified purines and their potential role in the RNA world. J Mol Evol 1999;48:631–7.
[47] Levy M, Miller SL, Oró J. Production of guanine from NH_4CN polymerizations. J Mol Evol 1999;49:165–8.
[48] Levy M, Miller SL, Brinton K, Bada JL. Prebiotic synthesis of adenine and amino acids under Europa-like conditions. Icarus 2000;145:609–13.
[49] Liu R, Orgel LE. Efficient oligomerization of negatively-charged D/L amino acids at −20 °C. J Am Chem Soc 1997;119:4791–2.
[50] Lindhal T. Instability and decay of the primary structure of DNA. Nature 1993;362:709–15.
[51] Maden BEH. No soup for starters? Autotrophy and origins of metabolism. Trends Biochem Sci 1995;20:337–41.
[52] Miller SL. A production of amino acids under possible primitive Earth conditions. Science 1953;117:528.
[53] Miller SL. Production of some organic compounds under possible primitive Earth conditions. J Am Chem Soc 1955;77:2351–61.
[54] Miller SL, Bada JL. Submarine hot springs and the origin of life. Nature 1988;334:609–11.
[55] Miller SL, Lazcano A. The origin of life—did it occur at high temperatures? J Mol Evol 1995;41:689–92.
[56] Miller SL, Lazcano A. In: Schopf JW, editor. Life's origin, the beginnings of biological evolution. Berkeley: California Univ. Press; 2002. p. 78.
[57] Nelson KE, Robertson MP, Levy M, Miller SL. Concentration by evaporation and the prebiotic synthesis of cytosine. Origins Life Evol Biosph 2001;31:221–9.
[58] Nielsen P. Peptide nucleic acid, PNA, a model structure for the primordial genetic material? Origins Life Evol Biosph 1993;23:323–7.
[59] Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. The structural basis of ribosome activity in peptide bond synthesis. Science 2000;289:920–30.
[60] Oparin AI. Proiskhozhedenie zhizni. Moscow: Mosckovskii Rabochii; 1924 (reprinted and translated in Bernal JD, the origin of life. London: Weidenfeld and Nicolson; 1967).
[61] Oparin AI. The origin of life. New York: MacMillan; 1938.
[62] Orgel LE. Polymerization on the rocks, theoretical introduction. Origins Life Evol Biosph 1998;28:227–34.
[63] Oró J. Synthesis of adenine from ammonium cyanide. Biochem Biophys Res Comm 1960;2:407–12.
[64] Oró J. Comets and the formation of biochemical compounds on the primitive earth. Nature 1961;190:442–3.
[65] Oró J, Kimball AP. Synthesis of purines under primitive Earth conditions. I. Adenine from hydrogen cyanide. Arch Biochem Biophys 1961;94:221–7.
[66] Oró J, Lazcano A. Comets and the origin and evolution of life. In: Thomas PJ, Chyba CF, McKay CP, editors. Comets and the origin and evolution of life. New York: Springer-Verlag; 1997. p. 3–27.
[67] Perutz M. I wish I'd made you angry earlier, essays on science, scientists, and humanity. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2003.
[68] Robertson MP, Miller SL. An efficient prebiotic synthesis of cytosine and uracil. Nature 1995;375:772–4.
[69] Robertson MP, Levy M, Miller SL. Prebiotic synthesis of diaminopyrimidine and thiocytosine. J Mol Evol 1996;43:543–50.
[70] Rode BM. Peptides and the origin of life. Peptides 1999;20:773–86.
[71] Saladino R, Crestini C, Costanzo G, DiMauro E. Advances in the prebiotic chemistry of nucleic acid bases. Implications for the origins of life. Curr Org Chem 2004;8.
[72] Saladino R, Crestini C, Costanzo G, Negri R, DiMauro E. Bioorg Med Chem 2001;9:1249–53.

[73] Saladino R, Ciambecchini U, Crestini C, Costanzo G, Negri R, DiMauro E. ChemBiochem 2003;4:514–21.
[74] Sanchez RA, Ferris JP, Orgel LE. Conditions for purine synthesis, did prebiotic synthesis occur at low temperatures? Science 1966;153:72–3.
[75] Sanchez RA, Ferris JP, Orgel LE. Studies in prebiotic synthesis: II. Synthesis of purine precursors and amino acids from aqueous hydrogen cyanide. J Mol Biol 1967;30:223–53.
[76] Sanchez RA, Ferris JP, Orgel LE. Studies in prebiotic synthesis: IV. The conversion of 4-aminoimidazole-5-carbonitrile derivatives to purines. J Mol Evol 1968;38:121–8.
[77] Schlesinger G, Miller SL. Prebiotic synthesis in atmospheres containing CH_4, CO, and CO_2. I. Amino acids. J Mol Evol 1983;19:376–82.
[78] Schopf JW. Microfossils of the early Archaean Apex chert, new evidence for the antiquity of life. Science 1993;260:640–6.
[79] Shapiro R. The prebiotic role of adenine, a critical analysis. Origins Life Evol Biosph 1995;25:83–98.
[80] Shapiro R, Klein RS. The deamination of cytidine and cytosine by acidic buffer solutions, mutagenic implications. Biochemistry 1966;5:2358–62.
[81] Shock EL. Geochemical constraints on the origin of organic compounds in hydrothermal systems. Orig Life Evol Biosph 1990;20:331–67.
[82] Sowerby SJ, Mörth C-M, Holm NG. Effect of temperature on the adsorption of adenine. Astrobiology 2001;1:481–8.
[83] Spirin AS. Ribosome structure and protein synthesis. Benjamin–Cummings: Menlo Park; 1986, 414 pp.
[84] Stoks PG, Schwartz AW. Uracil in carbonaceous meteorites. Nature 1979;282:709–10.
[85] Stoks PG, Schwartz AW. Nitrogen-heterocyclic compounds in meteorites: significance and mechanisms of formation. Geochim Cosmochim Acta 1981;45:563–9.
[86] Stoks PG, Schwartz AW. Basic nitrogen-heterocyclic compounds in the Murchison meteorite. Geochim Cosmochim Acta 1982;46:309–15.
[87] Tice MM, Lowe DR. Photosynthetic microbial mats in the 3.416-Myr-old ocean. Nature 2004;431:522–3.
[88] Urey HC. On the early chemical history of the Earth and the origin of life. Proc Natl Acad Sci USA 1952;38:351–63.
[89] van Zullen MA, Lepland A, Arrhenius G. Reassessing the evidence for the earliest traces of life. Nature 2002;418:627–30.
[90] Van der Velden W, Schwartz AW. Search for purines and pyrimidines in the Murchison meteorite. Geochim Cosmochim Acta 1977;41:961–8.
[91] Voet AB, Schwartz AW. Prebiotic adenine synthesis from HCN, evidence for a newly discovered major pathway. Bioorganic Chem 1983;12:8–17.
[92] Wächtershäuser G. Before enzymes and templates, theory of surface metabolism. Microbiol Rev 1988;52:452–84.
[93] Wilde SA, Valley JW, Peck WH, Graham CM. Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. Nature 2001;409:175–8.
[94] Zhabotinski AM. Self-oscillation concentrations. Moscow: Nauka; 1974.
[95] Zhang B, Cech TR. Peptidyl-transferase ribozymes, trans reactions, structural characterization and ribosomal RNA-like features. Chem Biol 1998;5:539–53.

Date: Sat, 09 Apr 2005 13:47:05 -0500 (CDT)
From: alar@correo.unam.mx
To: infinito@servidor.unam.mx
Cc: abb2@correo.unam.mx
Subject: Reenviar: OLEB 261

[The following text is in the "iso-8859-1" character set.]
[Your display is set for the "iso8859-1" character set.]
[Some characters may be displayed incorrectly.]

Aceptado --felicidades! Incluye esto en tu tesis, caro Ludovicus, y tu en tu Curriculum, Herr Susanito.

Il patrone, repartidor de alegrias

----- Mensaje reenviado por "Alan W. Schwartz" <alan@sci.kun.nl> -----
Date: Sat, 09 Apr 2005 16:22:46 +0200
From: "Alan W. Schwartz" <alan@sci.kun.nl>
Reply-To: "Alan W. Schwartz" <alan@sci.kun.nl>
Subject: OLEB 261
To: alar@correo.unam.mx

Dear Antonio,

This will acknowledge receipt and acceptance of your revised manuscript (*OLEB 261*).
Thank you very much for your contribution to the journal.

Please note, for future reference, that hard copies are not generally required for either reviewing or for publication, although you may be asked to supply Latex files by the publisher. At the moment this is not (yet) standard procedure. Also, unless files are extremely large (i.e. more than about 5 MB), files should be sent to me as email attachments. Please also note my home address below, which should be used for any material (for example, copies on disk) which you may wish to send to me by mail.

With best regards,

Alan

A.W. Schwartz
Lindenheuvel 12
Hilversum 1217 JX
The Netherlands

----- Fin del Mensaje reenviado -----

The Last Common Ancestor: what's in a name?

Luis Delaye, Arturo Becerra, and Antonio Lazcano*

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

* corresponding author

Abstract

Twenty completely sequenced cellular genomes from the three major domains were analyzed using twice one-way BLAST searches in order to define the set of the most conserved protein-encoding sequences to characterize the gene complement of the last common ancestor of extant life. The resulting set is dominated by different putative ATPases, and by molecules involved in gene expression and RNA metabolism. DEAD-type RNA helicase and enolase genes, which are known to be part of the RNA degradosome, are as conserved as many transcription and translation genes. This suggests the early evolution of a control mechanism for gene expression at the RNA level, providing additional support to the hypothesis that during early cellular evolution RNA molecules played a more prominent role. Conserved sequences related to biosynthetic pathways include those encoding putative phosphoribosyl pyrophosphate synthase and thioredoxin, which participate in nucleotide metabolism. Although the information contained in the available databases corresponds only to a minor portion of biological diversity, the sequences reported here are likely to be part of an essential and highly conserved pool of proteins domains common to all organisms.

Key words: last common ancestor, cenancestor, RNA/protein world, progenote

1. INTRODUCTION

One of the major achievements of molecular cladistics has been the evolutionary comparison of small subunit ribosomal RNA (rRNA) sequences, which has allowed the construction of an 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). The variations of traits common to these major groups can be easily explained as the outcome of divergent processes from an ancestral lifeform that existed prior to the separation of the three major biological domains, i.e., the last common ancestor (LCA) or cenancestor (Fitch and Upper, 1987). No paleontological remains will bear testimony of its existence, as the search for a fossil of the cenancestor is bound to prove fruitless. However, insights on its nature can in principle be deduced from the molecular record.

From a cladistic viewpoint, the LCA is merely an inferred inventory of features shared among extant organisms, all of which are located at the tip of the branches of molecular trees. From an evolutionary point of view, however, 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. However, the conclusion that the LCA was a progenote, i.e., a hypothetical biological entity in which phenotype and genotype still had an imprecise, rudimentary linkage relationship (Woese and Fox, 1977), was disputed some time ago when the analysis of homologous traits found among some of its descendants suggested that it was not a direct, immediate descendant of the RNA world, a protocell or any other pre-life progenitor system. Under the implicit assumption that lateral gene transfer (LGT) had not been a major driving force in the distribution of homologous traits in the three domains, it was concluded that the LCA was a complex organism, much like extant bacteria (Lazcano et al., 1992; Lazcano, 1995).

A decade ago many were convinced that the LCA was very much like extant prokaryotes, but the inventory of shared traits based on sequence comparisons was small. It was surmised that the sketchy picture developed with the limited data bases would be confirmed by completely sequenced cell genomes from the three primary domains. This has not been the case: the availability of an increasingly large number of completely sequenced cellular genomes has sparked new debates, rekindling the discussion on the nature of the ancestral entity (Doolittle 2000). This is shown, for instance, in the diversity of names that have been coined to describe it: progenote (Woese and Fox, 1977), cenancestor (Fitch and Upper, 1987), LUCA, a term first coined to describe the last universal common ancestor (Philippe and Forterre, 1999), and then as an acronym for the last universal cellular ancestor (Forterre, 2002), and LCC, last common community (Line, 2002), among others. These terms are not truly synonymous, and they reflect the current controversies on the nature of the universal ancestor and the evolutionary processes that shaped it.

Lateral gene transfer and the reconstruction of early cell evolution

In the past few years the analysis of an increasingly large number of completely sequenced cellular genomes has revealed major discrepancies with the topology of rRNA trees. Very often these differences have been interpreted as evidence of lateral gene transfer (LGT) events between widely separated species, questioning the feasibility of the reconstruction and proper understanding of early biological history (Doolittle, 1999, 2000). There is clear evidence that genomes have a mosaic-like nature whose components may come from many different phylogenetically separated donor species (Ochman et al., 2000; Zhaxybayeva and Gogarten, 2004; Zhaxybayeva et al., 2004). Depending on their different advocates, a wide spectrum of mix-and-match recombination processes have been described, ranging from the lateral transfer of few genes via conjugation, transduction or transformation, to cell fusion events involving organisms from the same or even different domains (Rivera and Lake, 2004).

Defining the nature of LCA is one of the central goals of the study of the early evolution of life on Earth, and several attempts have been made in this direction. Proper description of the LCA is still an unfinished task, mainly because of the complexity of the evolutionary process that connect extant organisms with it, the lack of a full understanding of the major evolutionary events that have taken place along the history of life on Earth, and the limitations inherent of the methodological attempts to reconstruct its nature. In this paper, we survey some of the difficulties encountered in the characterization of the last common ancestor, and summarize ongoing discussions on its nature. We also attempt a reconstruction of the gene complement of the LCA based on the conservation of proteins in a database of twenty completely sequenced cellular genomes from the three major domains, from which endosymbionts and obligate parasites were excluded.

3. MATERIAL AND METHODS

To avoid biases in the backtrack characterization of the LCA due to secondary gene losses, a sample of complete proteomes from the three domains of life from non-endosymbiotic or non parasitic species was downloaded from the Kyoto Encyclopedia of Genes and Genomes KEGG data base (ftp://ftp.genome.ad.jp/pub/kegg/) (Kanehisa et al., 2000). The following species were analized: Bacteria, *Bacillus subtilis*, *Streptococcus pneumoniae*, *Thermoanaerobacter tengcongensis* (Firmicutes), *Escherichia coli* K-12 (Proteobacteria), *Fusobacterium nucleatum* (Fusobacteria), *Synechocystis* sp. (Cyanobacteria), *Aquifex aeolicus* (Aquificae), *Deinococcus radiodurans* (Deinococcus-Thermus); Archaea, *Archaeoglobus fulgidus*, *Methanococcus jannaschii*, *Halobacterium sp.*, *Thermoplasma acidophilum*, *Pyrococcus horikoshii* (Euryarchaeota), *Aeropyrum pernix*, *Sulfolobus solfataricus*, *Pyrobaculum aerophilum* (Crenarchaeota); and Eucarya, *Caenorhabditis elegans* (Animalia), *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* (Fungi), *Arabidopsis thaliana* (Plantae).

In order to construct the set of the most conserved set of proteins common to these proteomes, a one-way BLAST search strategy was performed using BLAST algorithm (Altschul, et al., 1997) as summarized in Figure 1. The efficacy of this analysis depends on the ability of each BLAST search to find all homologs of the query sequence A. Although this methodology may miss homologs common to all three lineages, it has the advantage of constructing a census of only the most conserved sequences, or of the most conserved domains in proteins. The order in which these genomes were first analyzed by one way BLAST search was as follows: *A. aeolicus, A. fulgidus, A. pernix, A. thaliana, B. subtilis, C. elegans, D. radiodurans, E. coli, F. nucleatum, Halobacterium sp., M. jannaschii, P. aerophilum, P. horikoshii, S. cerevisiae, S. pneumoniae, S. pombe, S. solfataricus, Synechocystis sp., T. acidophilum, T. Tengcongensis.* A second one way BLAST search was then performed in opposite direction. Only one false negative sequence was identified, b1740, that corresponds to a NAD synthetase (glutamine-hydrolysing). Since domains are the structural and evolutionary units of proteins, we have extracted the highly conserved sequences from *E. coli, M. jannaschii* and *S. cerevisiae*, which are among the most well-studied organisms, and identified the protein domains conserved between each group of homologous sequences using the Pfam database (Bateman et al., 2004) (http://www.sanger.ac.uk/Software/Pfam/). This allowed the identification of multidomain proteins while avoiding the problem of false positives. To avoid the difficulties with sequences that can be classified under multiple categories, such as enolase (i.e., sugar metabolism or as a component of the degradosome), Table 1 follows, only in part, the cellular functional classes described by KEGG, with emphasis on a broad-scale description of sequences that interact with RNA.

4. RESULTS

Because our interest is centered on the construction of a census of the most conserved proteins, only the genomes of *E. coli, S. cerevisiae* and *M. jannaschii* were analyzed in detail. The sequences in the resulting set have been classified according to functional categories which we have modified from those used in KEGG to avoid diluting sequences involved in RNA metabolism among other categories (Table I). There are 283 highly conserved proteins from *S. cerevisiae*, 245 from *E. coli*, and 145 from *M. jannaschii*. This set represents the most conserved sequences common to all genomes studied here. As shown in Table I, the list of highly conserved molecular traits includes sequences related to informational process like transcription and translation and several kinds of metabolic enzymes.

As expected from previous studies, different groups of genes involved in different RNAs (Klenk et al., 1993) and translation (Olsen and Woese; 1997; Koonin, 2003; Harris et al., 2003) exhibit a high level of conservation, while the only replication-related conserved ORFs are the bacterial clamp-loading protein complex (Edgell and Doolittle, 1997) (*dnaX*, b0470), and its archaeal/eukaryotic homologs (replication factor-C, MJ1422, MJ0884, YJR068W, YNL290W, YOL094C), which belongs to the AAA (ATPases Associated with diverse cellular Activities) family (Table I). As shown in Table I, the set of sequences compiled using the methodology outlined here is overwhelmingly dominated by (a) molecules related to translation, RNA synthesis (i.e, transcription), translation and degradation; and (b) proteins with ATP-binding and hydrolyzing activities which can be grouped in relatively few discrete sets (ABC transporter subunits, RNA DEAD helicases, AAA-type ATPases, and HIT superfamily of nucleotide-binding proteins). Given the ubiquity and diversity of these different proteins with ATPase activity, it is perhaps not surprising that the sequences encoding the hydrophilic subunits of F-type ATP synthases and their homologs in the three domains are also highly conserved.

As reported by others (Mushegian and Koonin, 1996; Koonin, 2003; Harris et al., 2003) the resulting repertoire includes few isolated sequences from incompletely represented basic biological processes, such as energy metabolism, nucleotide and amino acid biosynthetic pathways, transcription, translation, and folding of proteins, as well as some sequences related to replication, repair, and cellular transport. As discussed below, this pattern of primary sequence conservation can be explained by manifold processes that include polyphyletic losses, the metabolic idiosyncrasies of diverse species, and different rates of molecular evolution.

5. DISCUSSION

The methodological approach developed here is straightforward: we searched for sequences present in all genomes analyzed that have changed slowly enough to be still recognizable using the one-way BLAST strategy described above. Because a simple BLAST search may not find all possible homologs of a given sequence, the results shown here represent a first approximation of the gene complement of the LCA as described from the standpoint of *S. cerevisiae*, *E. coli* and *M. jannaschii*, three species that have been selected because of the considerable information that exists on their biology. A more complete census of highly conserved traits would require a detailed analysis of each set of homologous proteins using more sensitive approaches like profile-based methods and, eventually, information derived from tertiary structure databases.

Reconstructions of gene complements of distant ancestors are mere statistical approximations of biological past, since their accuracy depends on manifold factors including the possible biases in the construction of genome databases, the levels of horizontal gene transfer, the significant variations in substitution rates of different proteins, and the degree of secondary looses, as well as methodological caveats. As argued here, in spite of these limitations the available data provides significant insights into (a) the existence of an ancient RNA/protein world; (b) the biological complexity of the LCA; and (c) and evidence pertaining to the chemical nature of the cenancestral genome (i.e. RNA or DNA).

The high proportion of cenancestral RNA-related ORFs suggest the prior-existence of an RNA/protein world

In order to avoid the bias introduced by secondary gene losses (Becerra et al., 1997), we have not included in our analysis genomes from obligate parasites or endosymbiotic organisms, which would lead to an underestimation of the number of genes inherited from the LCA. As demonstrated by other analyses, proteins that interact with RNA in one way or other are among the most highly conserved sequences (Delave and Lazcano, 2000; Anantharaman et al., 2002). This is shown in Figure 2, where more than 80% of the conserved domains correspond to proteins that interact directly with RNA (such as ribosomal proteins, DEAD-type helicases, aminoacyl tRNA synthetases, and elongation factors, among others), or take part in RNA and nucleotide biosyntheses, including the DNA-dependent RNA polymerase β and β' subunits, dimethyladenosine transferase, adenyl-succinate lyases, dihydroorotate oxidase, and ribose-phosphate pyrophosphokinase, among many others (Table I). This percentage includes sugar metabolism-related sequences (see below).

Nonetheless, few metabolic genes are part of the conserved ORF product set. These include many sugar metabolism-related sequences, such as the enolase-encoding genes noted above, as well as homologs of thioredoxin (*trxB*, mj1536), phosphoribosyl-pyrophosphate synthase (*prs*, b1207), and UDP-galactose 4-epimerase (*galE*, b0759) genes. Very likely, the evolutionary conservation of the *trxB* and *prsA* genes is best understood in terms of the key roles they play in nucleotide biosynthesis. The role of UDP-galactose 4-epimerase in complex carbohydrate synthesis via the interconversion of the galactosyl and glucosyl groups is well-known. Although the uniqueness of the enzyme mechanism has been acknowledged, it is possible that the conservation of UDP-galactose 4-epimerase is due to an undescribed participation in other basic processes, as in the case of enolase.

ATP-dependent RNA helicases are universally-distributed, highly conserved proteins which participate in a variety of cellular functions involving the unwinding and rearrangement of RNA molecules, including translation initiation, RNA splicing, ribosome assembly, mRNA nucleocytoplasmic transport, and degradosome-mediated mRNA decay (Schmid and Linder 1992). The degradosome is a multienzymatic complex involved in mRNA processing and breakdown, that includes polynucleotide phosphorylase (which shares an RNA-binding domain with RNAse E), polyphosphate kinase (PPK), ATP-dependent DEAD/H-type RNA helicase (RhlB), and enolase, a glycolytic enzyme that catalyzes the conversion of 2-phosphoglycerate to phosphoenolpyruvate and water (Blum et al., 1997). Reports showing that PPK is not essential for *E. coli* survival and may be a later evolutionary addition involved in degradosome regulation (Blum et al., 1997) are consistent with the absence of the corresponding gene from the set of highly similar ORFs.

Although RNA hydrolysis is an exergonic process, degradosome-mediated mRNA turnover plays a key role as a regulatory mechanism for gene expression in both prokaryotes and eukaryotes (Blum et al., 1997). A possible explanation for the conservation of DEAD-type RNA helicases may lie in their role in protein biosynthesis and in mRNA degradation. This possibility is supported by the phylogenetic relatedness of the RhlB and DeaD sequences (Schmid and Linder, 1992) and by the surprising conservation of the *eno*-like sequences. If this interpretation is correct, then it could be argued that degradosome-mediated mRNA turnover is an ancient control mechanism at RNA level that was established prior to the divergence of the three primary kingdoms. Together with other lines of evidence, including the observation that the most highly conserved gene clusters in several (eu)bacterial genomes are regulated at RNA-level (Siefert et al., 1997), the results reported here are fully consistent with the hypothesis that during early stages of cell evolution RNA molecules played a more conspicuous role in cellular processes.

The LCA, a progenote or a bacterial-like cenancestor?

Our ability to reconstruct the gene complement of the LCA depends in part on the levels of lateral gene transfer during early cell evolution, as well as on the degree of differential gene losses across cellular lineages (Becerra et al., 1997). If lateral gene transfer was rampant during early evolution, then there is a risk of overestimating the number of genes of the LCA. The opposite outcome can be expected if, on the other hand, differential losses have been much more common in evolution.

Analysis of an increasingly large number of genes and genomes has revealed major discrepancies with the topology of rRNA trees. As summarized by Brown (2003), very often these differences have been interpreted as evidence of LGT events between different species, questioning the feasibility of the reconstruction and proper understanding of early biological history (Doolittle, 1999). *There is evidence that genomes have a mosaic-like nature whose components come* from a wide variety of sources (Ochman et al., 2000). However, not all sequences are equally prone to such phenomenon, nor has the evolution of all prokaryotic species been equally affected by LGT (Zhaxybayeva et al., 2004). Most transfers take place between closely related species, and interdomain LGT events are quite rare. Accordingly, if the species that carries the ancestral sequence was not the organismal ancestor itself, then very likely was its close relative (Zhaxybayeva and Gogarten, 2004).

Driven in part by the impact of lateral gene acquisition as revealed by the discrepancies of different gene phylogenies with the rRNA tree, Woese (1998) proposed that the LCA was not a single organism, but rather a highly diverse population of metabolically complementary, cellular progenotes endowed with multiple, small linear chromosome-like genomes that benefited from massive multidirectional horizontal transfer events. According to this idea, which is reminiscent of a similar hypothesis proposed independently by Kandler (1994), the essential features of translation and the development of metabolic pathways took place before the earliest branching event, but what led to the three domains was not a single ancestral lineage, but a rapidly differentiating community of genetic entities. This communal ancestor occupied as a whole the node located at the bottom of the universal tree, in which the decrease of sequence exchange and increasing genetic isolation eventually lead to the observed tripartite division of the biosphere.

We suggest a different scenario. The LCA was of course not alone: company must have been kept by its siblings, a population of entities similar to it that existed throughout the same period. They may have not survived, but some of their genes did if they became integrated via lateral transfer into the LCA genome. The cenancestor is one of the last evolutionary outcomes of a tree trunk of unknown length, during which the history of a long but not necessarily slow (Lazcano and Miller, 1994) series of ancestral events including lateral gene transfer, gene losses, and duplications probably played a significant role in the accretion of complex genomes (Lazcano et al., 1992; Castresana, 2001; Snel et al., 2002).

The gene complement of the LCA can also be considered a mosaic in that not all universally distributed genes are of equal antiquity. For instance, the evidence suggesting that DNA evolved after RNA and proteins (Lazcano et al., 1988; Freeland et al., 1999) implies that the translational machinery is older than ribonucleotide reductases or DNA polymerases. However, the extraordinary similarity of the basic traits shared by all extant cells suggest that they must have been integrated by the time of the LCA.

The genetic entities that formed the communal ancestor proposed by Woese (1998) may have been extremely diverse, but an indication of their ultimate monophyletic origin from a sole progenitor is provided by universally distributed features such as the genetic code and the basic features of the gene expression machinery. Did this hypothetical communal progenote ancestor diverged sharply into the three domains soon after the appearance of the code and the establishment of translation? Not necessarily. The origin of the mutant sequences ancestral to those found in all extant species, and the divergence of the Bacteria, Archaea, and Eucarya were not synchronous events, i.e., the separation of the primary domains took place later, perhaps even much later, than the appearance of the genetic components of their last common ancestor. Moreover, 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. What we have been calling the root of the universal tree is in fact the tip of its trunk: inventories of LCA genes include sequences that originated in different pre-cenancestral epochs (Delaye and Lazcano, 2000; Becerra-Bracho et al., 2000; Anantharaman et al., 2002). As noted by Fox et al (1982), a major phylogenetic issue is the relationship between the timing of the transition from the age of progenotes and the branching order between the three cell domains. The gene complement of the LCA described here corresponds to a cellular entity that evolved after the age of progenote had come to an end, but prior to the divergence of the Archaea, Bacteria and Eucarya.

It is currently difficult to propose a unifying hypothesis. However, the scheme outlined here is supported by gene content trees, which exhibit a broad-level agreement with rRNA-based phylogenies (Fitz-Gibbon and House, 1999; Snel et al., 1999; Tekaia, et al., 1999). Such trees are not cladograms but phenograms, i.e., they are merely hierarchical *representations of similarities and differences in gene content, where the presence or absence of a sequence is counted* as a character. Since different lineages evolve at different rates, such overall similarity may be an equivocal indicator of

genealogical relationships. Nevertheless, these trees are consistent with rRNA phylogenies, and do not support the hypothesis of massive LGT between distant species. The robustness exhibited by these different methodologies indicates that although LGT has played an important role in cellular evolution, it has not obliterated the early history of life (Glansdorff 2000), and that the role of reticulate evolution in defining the LCA as a progenote swarm may have been overstated.

The LCA, a DNA or a RNA genome?

Since all extant cells are endowed with DNA genomes, the most parsimonious conclusion is that this genetic polymer was already present in the cenancestral population. Woese (1983, 1987) has suggested otherwise, arguing for a progenote-like universal ancestor endowed with a rapidly evolving genome formed by disaggregated, small-sized RNA molecules. This possibility appeared to be supported by the findings of Mushegian and Koonin (1996), who argued that the absence of eucaryal or archaeal homologs of key components of DNA replication and nucleotide biosynthesis in the minimal gene set which resulted from the comparison of the *Haemophilus influenzae* and *Mycoplasma genitalium* genomes suggested that the cenancestor had used RNA as genetic polymer. Such conclusion is weakened by the limited data set analyzed, which consisted of only two parasitic bacterial genomes that have undergone extensive polyphyletic gene losses (Becerra et al., 1997). In a subsequent publication, however, Koonin and his collaborators analyzed a large set of primases, replicative polymerases, and other proteins involved in DNA replication, and suggested an alternative scheme with a hybrid RNA/DNA cenancestral genetic system whose complex replication cycle involved reverse transcription (Leipe et al., 1999).

The idea that RNA preceded DNA as cellular genetic material has been proposed independently by many authors (see, for instance, Oparin, 1961; Rich, 1962; Haldane, 1965; Reanney, 1979). However, it is likely that double-stranded DNA genomes had become firmly established prior to the divergence of the three primary domains. The major arguments supporting this possibility are:

(a) in sharp contrast with other energetically favorable biochemical reactions (such as phosphodiester backbone hydrolysis or the transfer of amino groups), the direct removal of the oxygen from the 2'-C ribonucleotide pentose ring to form the corresponding deoxy-equivalents is a thermodynamically much less-favored reaction, considerably reducing the likelihood of multiple, independent origins of biological ribonucleotide reduction;

(b) demonstration of the monophyletic origin of ribonucleotide reductases (RNR) is greatly complicated by their highly divergent primary sequences and the different mechanisms by which they generate the substrate 3'-radical species required for the removal of the 2'-OH group. However, sequence analysis and biochemical characterization of archaebacterial RNRs have shown their similarities with their eubacterial and eukaryotic counterparts, suggesting that the most distributed enzymes are of monophyletic origin (Tauer et al., 1996; Riera et al., 1997; Freeland et al., 1999); and

(c) sequence similarities shared by many ancient, large proteins found in all three domains suggest that considerable fidelity existed in the operative genetic system of their common ancestor, but such fidelity is unlikely to be found in RNA-based genetic systems (Lazcano et al., 1992).

While accepting a DNA component in the LCA genome, Leipe et al., (1999) have underlined the highly divergent character of the main components of the (eu)bacterial replication machinery when compared with their archaeal/eukaryotic counterpart. Although it is possible to recognize the evolutionary relatedness of various orthologous DNA informational proteins (i.e., ATP-dependent clamp loader proteins, topoisomerases, gyrases, and 5'-3' exonucleases) across the entire phylogenetic spectrum, comparative proteome analysis has shown that (eu) bacterial replicative polymerases and primases lack homologues in the two other primary kingdoms. As argued by Leipe et al., (1999) these observations can be explained by assuming a dual, independent origin of the DNA replication machineries of the Bacteria, on the one hand, and of the Archaea/Eucaryal on the other.

We think this is unlikely. Nucleic acid replication enzymatic machinery requires, at the very least, a replicase, a primase, and a helicase (Forterre, 1999), which are currently described as non-orthologues between the bacterial and the archaea/eukaryotic branches. Given the central role that is assigned to nucleic acid replication in mainstream definitions of life (Koshland 2002), the lack of conservation and polyphyly of several of its key enzymatic components is somewhat surprising. However, we believe that there may be an explanation for the evolution of the DNA replication machinery simpler that the one advocated by Leipe et al., (1999). Our hypothesis implies that this progenitor DNA polymerase was originally involved in the replication of the LCA genome, until its (eu)bacterial descendants underwent a non-orthologous displacement by the ancestor of the *Escherichia coli* replicative DNA pol III (DNA pol C) and its homologs. Based on the conservation and versatily of functions of the palm domain of DNA polymerase II and its homologs, we suggest that the Archaeal-Eucaryal replication machinery is in fact older than the current Bacterial one.

Conclusions and outlook

The variations of traits common to extant species can be easily explained as the outcome of divergent processes from an ancestral life form that existed prior to the separation of the three major biological domains, i.e., the last common ancestor (LCA) or cenancestor. However, if the term "universal distribution" is restricted to its most obvious sense, i.e., that of traits found in all completely sequenced genomes, then quite unexpectedly the resulting repertoire is formed by relatively few features and by incompletely represented biochemical processes (Tatusov et al., 1997; Tekaia et al., 1999; Brown et al., 2001; Delaye et al., 2002). Quite surprisingly, some of the most likely *a priori* candidates for strict universality, such as those sequences involved in DNA replication, have also turned out to be not only poorly preserved but also, in some cases, of polyphyletic origin (Edgell and Doolittle, 1997; Olsen and Woese 1997; Böhlke et al., 2000).

The traits described here as inherited from the LCA represent only those sequences that can be identified at the primary structure level. Because it is known that tertiary structure level is more conserved across evolutionary distances, attempts to reconstruct the LCA gene complement using a fold-recognition algorithm may enhance the census of such cenancestral molecular traits. For instance, the lack of detection of the complete set of sequences encoding F-type ATPases does not indicate these multimeric enzymes were absent in the LCA (Gogarten and Taiz, 1992; Castresana et al., 1994), but should be interpreted instead as an indication of the different rates of evolution of the sequences and the limits of the methodologies described here. Nonetheless, the fact that are no major inconsistencies between the data presented here and those reported by other authors who have used different methodologies (Delaye and Lazcano, 2000; Anantharaman et al., 2002; Harris et al., 2003; Koonin, 2003; Delaye et al., 2004) underlines the robustness of our own results.

The dataset reported in Table I includes (a) genes that have undergone lateral transfer and (b) sequences that although highly conserved, have originated in different evolutionary epochs. However, the over-representation of highly conserved sequences related to RNA metabolism, i.e., ORFs whose products synthesize, degrade, or interact with polyribonucleotides (Table I), is best understood in terms of an early evolutionary period during which RNA played a more prominent role in biological processes, i.e., an RNA/protein world. Degradasome components (DEAD helicase and enolase) are as highly conserved as molecules involved in RNA biosynthesis. It can also be argued that the conservation of enolase and DEAD-type RNA helicases discussed here constitutes additional evidence of the early development of gene expression control mechanisms at the RNA level. These conclusions, however, do not imply that the cenancestor was endowed with an RNA genome, nor support the possibility of an RNA-based origin of life. Indeed, the conservation of genes involved in ribonucleotide reduction such as *trxB*, combined with other independent lines of evidence, argues for the presence of a DNA genome in the LCA.

The analysis of the dataset reported here is consistent with a prokaryotic root of universal phylogenies, and indicates that the cenancestor was much more complex that expected for a progenote. There is no contradiction between this conclusion and the relatively few metabolic genes that are conserved. In fact, most of them synthesize or interact with ribonucleotides or sugar compounds. Conserved sequences related to metabolic pathways include homologs of phosphoribosyl pyrophosphate synthase and thioredoxin, among others, which are involved in nucleotide biosynthesis. Although the information contained in the available databases corresponds only to a minor portion of biological diversity, the sequences reported here are likely to be part of an essential and highly conserved pool of proteins common to all organisms.

Acknowledgments

Work reported here was supported by project DGPA IN-111003 (UNAM, Mexico). We are deeply indebted to two anonymous reviewers for their many suggestions and help in improving the results presented here. This paper was completed during a leave of absence in which one of us (A.B.) enjoyed the hospitality of Dr. Janet Siefert (Rice University, Houston) thanks to the support of DGPA-UNAM. Due thanks are given to the Departamento de Supercómputo, DGSCA-UNAM, for its constant support and technical assistance.

Figure captions

Fig. 1. The one way BLAST search. (a) a given sequence (white circle) is considered as highly conserved if it has at least one homolog (gray circles) in all genomes (G_n) in our dataset. Divergent homologous sequences (black circles) may be mismatched or remain unidentified. Sequences detected as highly conserved (HC) by the BLAST search (e-value <0.0001), are represented by gray circles, are not used as query sequences in additional searches once they have been identified as HC; (b) in some cases, a previously undetected homolog will be matched again with the same family of homologous sequences. To avoid this kind of redundancy, visual inspection was performed to merge all the corresponding sequences in one single family; (c) due to protein domain fusions, in some cases false HC protein families will be constructed. A Pfam protein domain analysis was performed in *E. coli*, *M. jannaschii* and *S. cerevisiae* to identify them and eliminate the false HC family from the final dataset; and (d) simple BLAST searches are likely to miss some homologous sequences. In order to make a more comprehensive search, it would be necessary to use more sensitive methods like profile-based algorithms or three-dimensional structure comparisons, which are not feasible for the time being. As described in the text, to construct the database reported here, steps (a), (b) and (c) were applied.

Fig. 2. Prevalence of highly conserved sequences and protein domains related to RNA metabolism in the dataset estimated here.

Table I. List of highly conserved genes in *E. coli*, *M. jannaschii* and *S. cerevisiae* genomes. Sequences involved in nucleotide-, sugar-, or nucleic acid metabolism are located in the shaded part of the table. Conserved Pfam domains in the three genomes are shown. Sequences with the same domain organization and function are underlined. Sequences are grouped according to single functional categories. *Different groups of homologous are joined in a single functional category. ** Hypothetical ORF.

Description	Pfam domain	*E. coli*	*M. iannaschii*	*S. cerevisiae*
Transcription				
RNA polymerase β	RNA_pol_Rpb2_1; RNA_pol_Rpb2_2; RNA_pol_Rpb2_3; RNA_pol_Rpb2_6; RNA_pol_Rpb2_7	b3987	MJ1040, MJ1041	YOR151C, YOR207C, YPR010C
RNA polymerase β'	RNA_pol_Rpb1_1, RNA_pol_Rpb1_2; RNA_pol_Rpb1_3, RNA_pol_Rpb1_4; RNA_pol_Rpb1_5	b3988	MJ1042, MJ1043	YDL140C, YOR116C, YOR341W
Translation				
aminoacyl-tRNA synthetases class I	tRNA-synt_1c, tRNA-synt_1c_C; Arg_tRNA_synt_N; tRNA-synt_1d, tRNA-synt_1d_C, tRNA-synt_1	b2400, b0144, b1876, b3384, b2114, b0642, b0026, b4258, b0526	MJ1377, MJ0237, MJ1415, MJ1263, MJ0947, MJ1007, MJ0633	YOL033W, YGL245W, YOR168W, YDR341C, YHR091C, YDR268W, YOL097C, YGR264C, YGR171C, YPL040C, YBL076C, YGR094W, YLR382C, YPL160W, YNL247W
aminoacyl-tRNA synthetases class II	tRNA-synt_2d; tRNA_anti; tRNA-synt_2; tRNA-synt_2b; HGTP_anticodon; tRNA-synt_2c; DHHA1	b1714, b1713, b0930, b1866, b4129, b0893, b0194, b2697, b2890, b4155, b1719, b2514	MJ0487, MJ1108, MJ1555, MJ1077, MJ0564, MJ1238, MJ0228, MJ1197, MJ1000	YFL022C, YPR047W, YLR060W, YHR019C, YLL018C, YCR024C, YPL104W, YDR023W, YHR011W, YKL194C, YOR335C, YNL040W, YDR037W, YNL073W, YER087W, YIL078W, YHR020W, YPR033C
tRNA pseudouridine 55 synthase	TruB_N	b3166	MJ0148	YNL292W, YLR175W
dimethyladenosine transferase	RrnaAD	b0051	MJ1029	YPL266W
elongation factors (EG-G, EF-Tu, and other GTP binding proteins)	GTP_EFTU; GTP_EFTU_D2,	b3340, b3339, b3980, b4375, b2569, b3871, b3168, b3590, b2751	MJ1048, MJ0324, MJ0495, MJ0262, MJ1261, MJ0325	YDR385W, YOR133W, YBR118W, YPR080W, YLR069C, YJL102W, YOR187W, YNL163C, YKL173W, YDR172W, YKR084C, YLR289W, YAL035W, YOL023W, YER025W, YLR244C, YBL091C, YER078C, YFR006W
methionine aminopeptidase	Peptidase_M24	b0168, b2385, b3847, b2908	MJ1329, MJ0805	YLR244C, YBL091C, YER078C, YFR006W
ribosomal-protein-alanine acetyltransferase	Acetyltransf_1	b4373, b2434**, b1448, b4012	MJ1530, MJ1207	YHR013C
Sua5, RNA-binding	Sua5_yciO_yrdC	b3282	MJ0062	YGL169W
several different RNA-binding proteins containing the S1 domain	S1	b3164, b0911	MJ0117	YJR007W, YMR229C
ribosomal proteins (small subunit) *	Ribosomal_S5; Ribosomal_S5_C; Ribosomal_S2; KH_2; Ribosomal_S3_C; S4; Ribosomal_S7; Ribosomal_S8; Ribosomal_S9; Ribosomal_S10, Ribosomal_S11; Ribosomal_S12, Ribosomal_S13; Ribosomal_S19	b3303, b0169, b3314, b3296, b3341, b3306, b3230, b3321, b3297, b3342, b3298, b3316	MJ0475, MJ0982, MJ0461, MJ0190, MJ1047, MJ0470, MJ0195, MJ0322, MJ0191, MJ1046, MJ0189, MJ0180	YGL123W, YBR251W, YLR048W, YGR214W, YHL004W, YNL178W, YPL081W, YBR189W, YNL137C, YHR148W, YJR123W, YJL190C, YLR367W, YDL083C, YMR143W, YBR146W, YHL015W, YJL191W, YCR031C, YNR036C, YGR118W, YPR132W, YDR450W, YML026C, YNL081C, YOL040C, YNR037C
ribosomal proteins (large subunit) *	Ribosomal_L1, Ribosomal_L2; Ribosomal_L2_C; Ribosomal_L6; Ribosomal_L11_N; Ribosomal_L11; Ribosomal_L5; Ribosomal_L5_C; Ribosomal_L14	b3984, b3317, b3305, b3983, b3308, b3310	MJ0510, MJ0179, MJ0176, MJ0471, MJ0469, MJ0466	YGL135W, YPL220W, YEL050C, YFR031C-A, YIL018W, YGR220C, YNL067W, YGL147C, YDR237W, YGR085C, YPR102C, YKL170W, YBL087C, YER117W

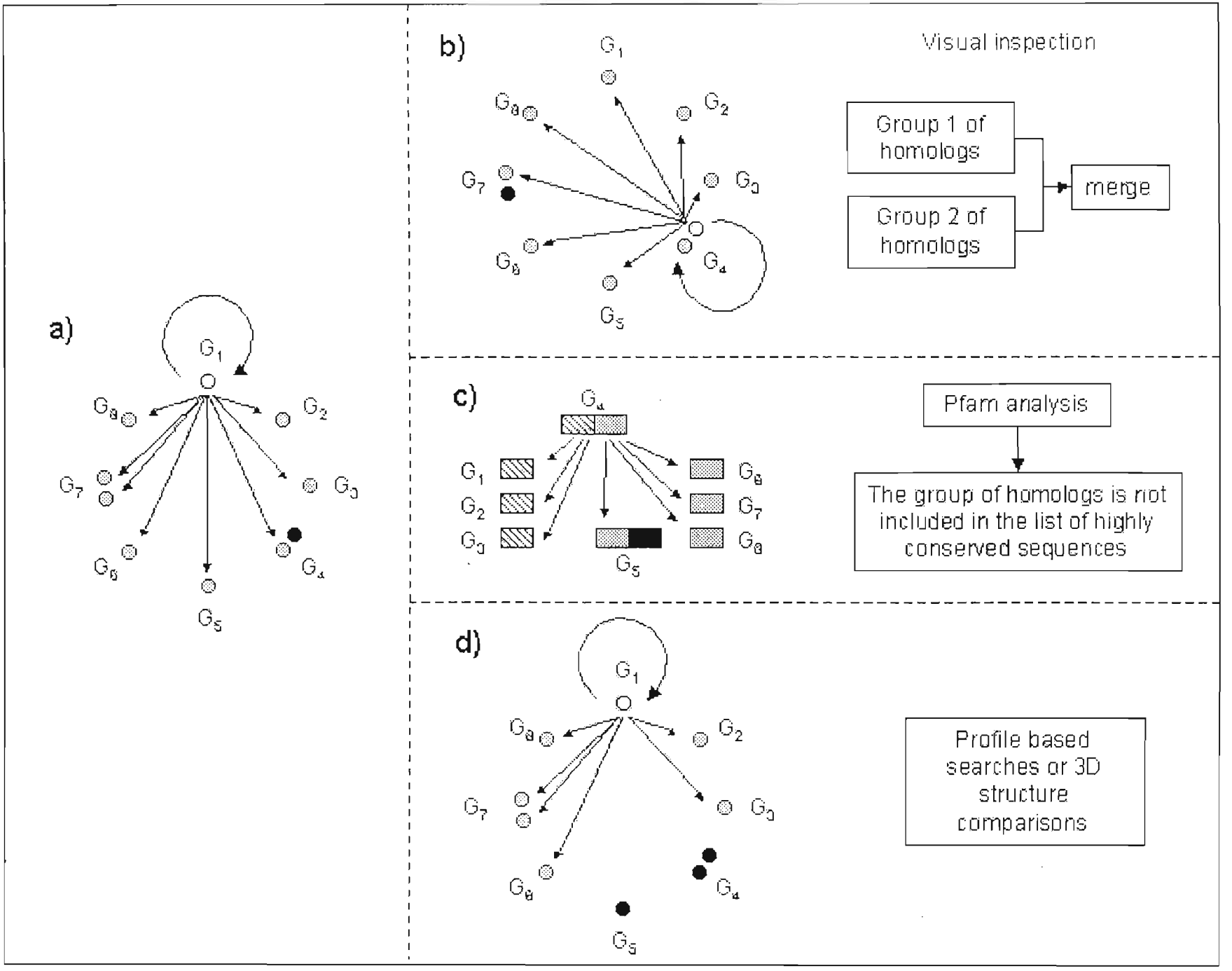
Table I.

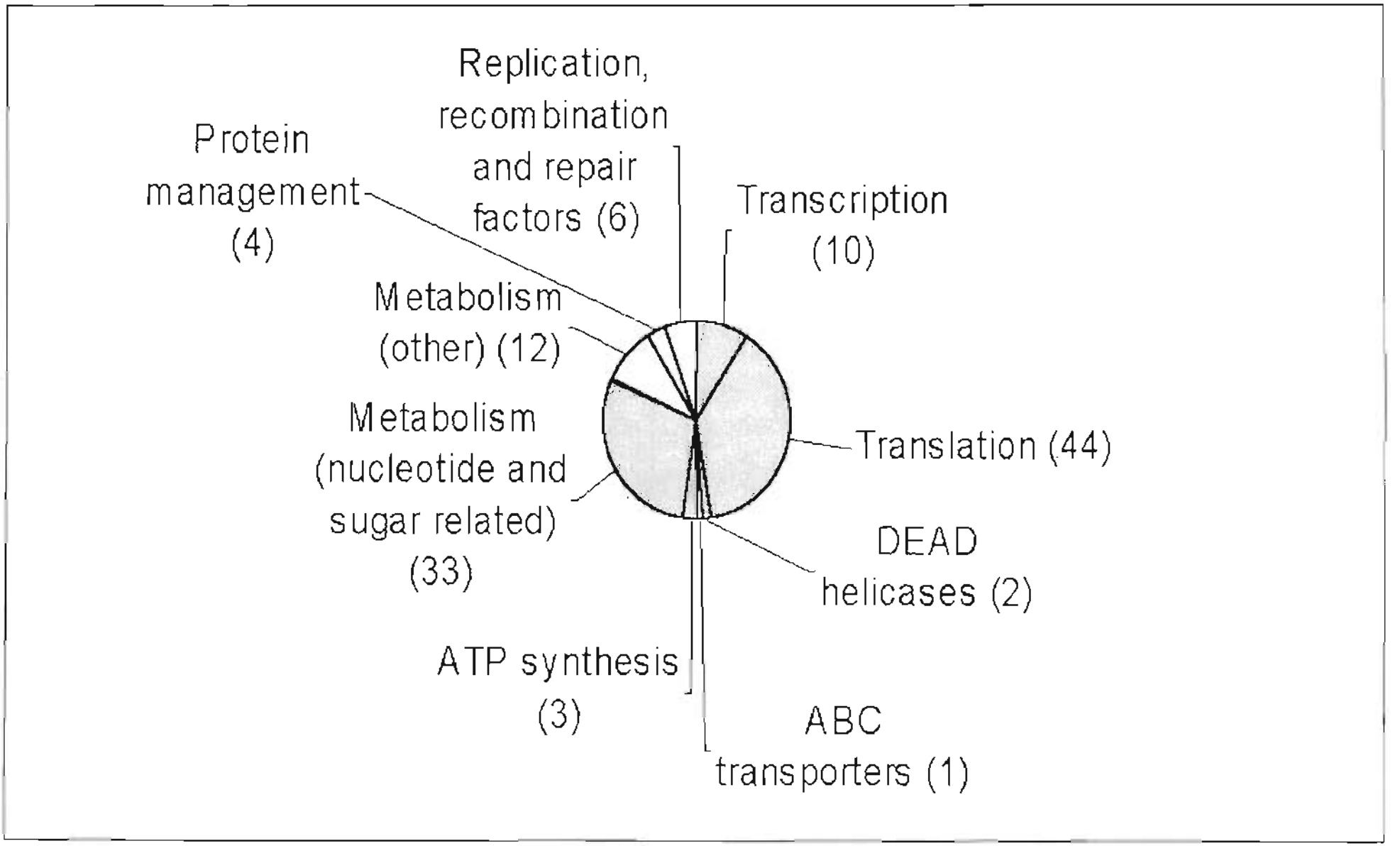
Description	Pfam domain	*E. coli*	*M. jannaschii*	*S. cerevisiae*
Metabolism				
thymidylate kinase [EC:2.7.4.9]	Thymidylate_kin	b1098	MJ0293	YJR057W
dihydroorotate oxidase [EC:1.3.3.1]	DHO_dh	b0945, b2147**	MJ0654	YKL216W
orotate *phosphoribosyltransferase* [EC:2.4.2.10]	Pribosyltran	b3642	MJ1109, MJ1655	YML106W, YMR271C
aspartate [EC:2.1.3.2] and ornithine [EC:2.1.3.3] carbamoyl- transferase catalytic chain	OTCace_N; OTCace	b4245, b4254, b0273, b2870**	MJ1581, MJ0881	YJL130C, YJL088W
carbamoyl-phosphate synthase small chain [EC:6.3.5.5]	CPSase_sm_chain; GATase	b0032, b3360, b2507, b1263	MJ1019, MJ0238, MJ1575, MJ1131	YJL130C, YOR303W, YKL211C, YMR217W
ribose-phosphate pyrophosphokinase [EC:2.7.6.1]	Pribosyltran	b1207	MJ1366	YER099C, YBL068W, YHL011C, YKL181W, YOL061W
IMP dehydrogenase [EC:1.1.1.205] and *hypothetical proteins*	IMPDH (1rst. half); CBS, CBS; *IMPDH (2nd. half)*	b2508	MJ1616, MJ0188, MJ1232, MJ0653, MJ0100, MJ1225, MJ0922, MJ0392, *MJ0868*, MJ1404, MJ0556	YAR073W, YHR216W, YML056C
adenylo succinate lyase [EC:4.3.2.2]	Lyase_1	b1131, b3960, b1611, b4139	MJ0929, MJ0791	YLR359W, YPL262W
amidophosphoribosyltransferase [EC:2.4.2.14]	GATase_2; Pribosyltran	b2312	MJ0204	YMR300C
pyruvate kinase [EC:2 7.1.40]	PK; PK_C	b1676, b1854	MJ0108	YAL038W, YOR347C
thioredoxin reductase and other reductases [EC:1.8.1.9]	Pyr_redox	b0888, b0606, b0116, b3500, b0304, b3962, b3365, b2711, b2763, b2542	MJ1536, MJ0649, MJ0551	YHR106W, YDR353W, YFL018C, YPL091W, YPL017C, YJR137C
glucosamine–fructose-6-phosphate aminotransferase *(isomerizing) (EC:2.6.1.16)*	GATase_2; SIS; SIS	b3729, b3371**	MJ1420, MJ1116	YKL104C, YMR085W, YMR084W
metal dependent hydrolase superfamily [EC:3.5.-.-]	Amidohydro_1	b2873**	MJ1490	YIR027C
UDP-galactose 4-epimerase [EC:5.1.3.2] and others	Epimerase	b0759, b3619, b2041, b3788	MJ0211, MJ1055	YBR019C
nucleotidyltransferase activity [EC: 2.7.7.-]	NTP_transferase; Hexapep	b2039, b3789, b1236, b2042, b3730, b3430	MJ1101, MJ1334	YDL055C, YDR211W
hypothetical nucleoside-triphosphatase [EC:3.6.1.15]	Ham1p_like	b2954**	MJ0226	YJR069C**
enolase [EC:4.2.1.11]	Enolase_N; Enolase_C	b2779	MJ0232, MJ0198**	YGR254W, YHR174W, YMR323W, YOR393W, YPL281C
phosphoglycerate kinase [EC:2.7.2.3]	PGK	b2926	MJ0641	YCR012W
phosphomannomutase [EC:5.4.2.8]	PGM_PMM_I; PGM_PMM_II; PGM_PMM_III; PGM_PMM_IV	b2048, b0688, b3176	MJ1100, MJ0399	YMR278W, YMR105C, YKL127W
sugar transferases	Glycos_transf_1	b2044, b3631	MJ1607, MJ1178, MJ1059, MJ1069	YPL175W
sugar transferases	Glycos_transf_2	b2254, b2351, b0363, b1022, b3615	MJ1222, MJ0544	YPL227C, YPR183W
nucleotide-binding proteins	HIT	b1103**	MJ0866**	YDL125C, YDR305C
phosphoglycerate dehydrogenase [EC:1.1.1.95]	2-Hacid_dh, 2-Hacid_dh_C; ACT	b2913, b1380, b3553, b2320, b1033	MJ1018	YER081W, YIL074C, YOR388C, YNL274C, YGL185C, YPL113C
NAD synthetase [EC:6.3.1.5, 6.3.5.1]	NAD_synthase	b1740	MJ1352	YHR074W
flavoprotein enzymes	Flavoprotein	b3639	MJ0913	YKL088W, YKR072C, YOR054C
UPP synthetase [EC:2.5 1.-]	Prenyltransf	b0174	MJ1372	YMR101C, YBR002C
tryptophan synthase (β-chain) [EC:4 2.1 20]	PALP	b1261, b3117, b2421, b3772, b2871, b2414	MJ1037, MJ1465	YGL026C, YCL064C, YKL218C, YGR155W, YER086W, YGR012W
tryptophan synthase (α-chain) [EC:4 2.1.20]	Trp_synlA	b1260	MJ1038	YGL026C
histidinol-phosphate aminotransferase [EC:2.6.1.9]	Aminotran_1_2	b2021, b2379, b0600, b2290, b1439, b1622, b4340	MJ0955, MJ0001, MJ1391, MJ0684, MJ1479	YIL116W, YJL060W, YDR111C, YLR089C
Polyprenyl synthetase [EC: 2.5.1-]	polyprenyl_synt	b0421, b3187	MJ0860	YJL167W, YPL069C, YBR003W
probable glyoxylase II [EC 3.1.2.6]	Lactamase_B	b0927**, b0212	MJ0888**	YDR272W
probable peroxiredoxin [EC:1.6.4.-]	AhpC-TSA	b0605, b2480	MJ0736	YIL010W, YBL064C, YML028W, YDR453C

Table I. (*continuation*).

Description	Pfam domain	*E. coli*	*M. jannaschii*	*S. cerevisiae*
DEAD helicases	DEAD; Helicase_C	b0797, b3162, b1343, b3780, b2576, b3822, b1653	MJ0669, MJ1401, MJ1574, MJ0294, MJ0383, MJ1124	YJL138C, YKR059W, YDR021W, YPL119C, YOR204W, YGL078C, YNL112W, YDL160C, YLL008W, YHR065C, YDL084W, YHR169W, YJL033W, YDR243C, YOR046C, YBR237W, YMR290C, YGL171W, YFL002C, YDL031W, YDR194C, YLR276C, YBR142W, YKR024C, YGL064C, YNR038W, YDR291W, YGL251C, YER172C, YMR190C, YGR271W
ATP synthesis (*atpA*, *atp8*)	ATP-synt_ab_N; ATP-synt_ab; ATP-synt_ab_C	b3734, b3732, b1941	MJ0217, MJ0216	YBL099W, YJR121W, YDL185W, YBR127C
Replication, recombination and repair factors				
ATPase family proteins (clamp-loading, γ τ subunits)	AAA	b0470, b3178, b0892,	MJ1422, MJ0884, MJ1156, MJ1176, MJ1494	YJR068W, YNL290W, YOL094C, YMR089C, YER017C, YDL126C *YBR030C, YPR024W, YGR270W*, YPR173C, YKL145W, YGL048C, YOR259C, YDL007W, YOR117W, YDR394W, YLR397C, YNL329C, YLL034C, YKL197C, YGR028W, YPL074W, YER047C, YDR375C, YBR186W
ribonuclease HII	RNase_HII	b0183	MJ0135	YNL072W
endonuclease III	HhH-GPD (1rst. half); HHH; HhH-GPD (2nd. half);	b1633, b2961	MJ1434, MJ0613	YAL015C, YOL043C
DNA topoisomerase I and III	Toprim; Topoisom_bac	b1274, b1783	MJ1652, MJ1512	YLR234W
ABC transporters	ABC_tran	b0448, b1290, b1291, b1496, b1682, b1709, b1756, b2201, b3479, b4058, b4096, b0066, b0127, b0151, b0199, b0262, b0366, b0449, b0490, b0495, b0588, b0652, b0760, b0794, b0809, b0820, b0829, b0855, b0864, b0879, b0886, b0887, b0914, b0933, b0949, b1117, b1126, b1246, b1247, b1318, b1441, b1483, b1484, b1513, b1858, b1900, b1917, b2129, b2149, b2180, b2306, b2422, b2547, b2677, b3201, b3271, b3352, b3450, b3454, b3455, b3463, b3480, b3486, b3540, b3541, b3567, b3725, b3749, b4035, b4087, b4097, b4106, b4228, b4287, b4391	MJ1023, MJ1088, MJ1242, MJ1267, MJ1367, MJ1508, MJ1572, MJ1662	YKR104W, YLL015W, YNR070W, YOR011W, YOR328W, YPL058C, YPL147W, YCR011C, YDR091C, YFR009W, YGR281W, YHL035C, YKL209C, YLL048C, YLR188W, YLR249W, YMR301C, YNL014W, YOL075C, YOR153W, YPL226W, YPL270W
Protein management				
signal recognition particle protein	SRP54_N; SRP54; SRP_SPB	b2610 b3464	MJ0101, MJ0291	YPR088C, YDR292C
chaperonin Cpn60	Cpn60_TCP1	b4143	MJ0999	YLR259C, YJR064W, YJL111W, YOL143W, YIL142W, YDR188W, YJL014W, YDR212W, YJL008C

Table I. (*continuation*).


a)
b)
Visual inspection
Group 1 of homologs
Group 2 of homologs
merge
c)
Pfam analysis
The group of homologs is not included in the list of highly conserved sequences
d)
Profile based searches or 3D structure comparisons


Replication, recombination and repair factors (6)
Protein management (4)
Transcription (10)
Metabolism (other) (12)
Metabolism (nucleotide and sugar related) (33)
Translation (44)
DEAD helicases (2)
ATP synthesis (3)
ABC transporters (1)

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, Z., Miller, W., and Lipman, D. J.: 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res* **25**: 3389-3402

Anantharaman, V., Koonin, E. V., and Aravind, L.: 2002, Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acid Res.* **30**: 1427-1464

Bateman, A., Coin, L., Durbin, R., Finn, RD., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R.: 2004, The Pfam protein families database. *Nucleic Acids Res.* **32**: 138-141

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

Becerra-Bracho, A., Velasco, A. M., Islas, S., Silva, E., Lloret, S. and Lazcano, A.: 2000, Molecular biology and the reconstruction of microbial phylogenies: des laisions dangereuses? *In* J. Chela-Flores, G. Lemerchand, and J. Oró (eds), *Origins from the Big-Bang to Biology: Proceedings of the First Ibero-American School of Astrobiology*, Kluwer Academic Publishers, Dordrecht, pp. 135-150

Blum, E., Py, B., Carpousis, A. J.,and Higgins, C. F.: 1997, Polyphosphate kinase is a component of the *Escherichia coli* RNA degradosome. *Mol. Microbiol.* **26**: 387-398

Böhlke, K., Pisani, F. M., Vorgias, C. E., Frey, B., Sobek, H., Rossi, M., and Antranikian, G.: 2000, PCR performance of the B-type DNA polymerase from the thermophilic euryarchaeon *Thermococcus aggregans* improved by mutations in the Y-GG/A motif. *Nucleic Acid Res.* **28**: 3910-3917

Brown, J. R.: 2003, Ancient horizontal gene transfer. *Nature Rev. Genet.* **4**: 121-132

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., and Stanhope, M. J.: 2001, Universal trees based on large combined protein sequence data sets. *Nature Genet.* **28**: 281-285

Castresana, J.: 2001, Comparative genomics and bioenergetics. *Biochem. Biophys. Acta* **1506**: 147-162

Castresana, J., Lubben, M., Saraste, M., and Higgins, D. G.: 1994, Evolution of cytochrome oxidase, and enzyme older than atmospheric oxygen. *EMBO J.* **13**: 2516-2525

Delaye, L. and Lazcano, A. : 2000, RNA-binding peptides as molecular fossils *In* J. Chela-Flores, G. Lemerchand, and J. Oró (eds), *Origins from the Big-Bang to Biology: Proceedings of the First Ibero-American School of Astrobiology* : Kluwer Academic Publishers. Dordrecht,, pp. 285-288

Delaye, L., Becerra, A., and Lazcano, A.: 2002, The nature of the last common ancestor *In* Lluis Ribas de Pouplana (ed), *The Genetic Code and the Origin of Life* : Landes Bioscience, Georgetown, in press

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

Doolittle, W. F.: 2000, The nature of the universal ancestor and the evolution of the proteome. *Curr. Opinion Struct. Biol.* **10**: 355-358

Edgell, D. R.and Doolittle, W. F.: 1997, Archaea and the origin: s, of DNA replication proteins. *Cell* **89**. 995-998

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

Fitz-Gibbon, S. T. and House, C. H.: 1999, Whole genome-based phylogenetic analysis of free-living organisms. *Nucleic Acids Res.* **27**: 4218-4222

Forterre, P.: 1999, Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. *Mol. Microbiol.* **33**: 457-465

Fox, G. E., Luehrsen, K. R., and Woese, C. R.: 1982, Archaebacterial 5S ribosomal RNA. *Zbl. Bakt. Hyg. I Abt. Org.* **C3**: 330-345

Forterre, P.: 2002, *The origin of DNA genomes and DNA replication proteins. Curr. Opin. Microbiol.* **5**: 525-532

Freeland, S. J., Knight, R. D., and Landweber, L. F.: 1999, Do proteins predate DNA? *Science* **286**: 690-692

Glansdorff, N.: 2000, About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. *Mol. Microbiol.* **38**: 177-185

Gogarten, J. P. and Taiz, L.: 1992, *Evolution of proton-pumping ATPase: rooting the tree of life. Photosyn. Res.* **33**: 137-146

Haldane, J. B. S.: 1965, Data needed for the blueprint of the first organism. *In* Fox, S. W. (ed), *The Origin of Prebiological Systems and their Molecular Matrices* : Academic Press, New York, pp. 11-15

Harris, J.K., Kelley, S.T., Spiegelman, G.B., and Pace, N.R.: 2003, The genetic core of the universal ancestor. *Genome Res.* **13**: 407-412

Kandler, O.: 1994, The early diversification of life. *In* Stefan Bengtson : (ed), *Early Life on Earth: Nobel Symposium No. 84* : Columbia University Press/Nobel Foundation, New York, pp. 152-160

Kanehisa, M. and Goto, S.: 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**: 27-30

Klenk, H.-P., Palm, P., & Zillig, W.: 1993, DNA-dependent RNA polymerases as phylogenetic markers molecules. *Syst. Appl. Microbiol.* **16**, 138-147

Koonin, E.V.: 2003, Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews* **1**: 127-136

Koshland, D. E.: 2002, The seven pillars of life. *Science* **295**: 2215-2216

Lazcano, A., Guerrero, R., Margulis, L., and Oró, J.: 1988, The evolutionary transition from RNA to DNA in early cells. *J. Mol. Evol.* **27**: 283-290

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

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

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

Leipe, D.D., Aravind, L., and Koonin, E. V.: 1999, Did DNA replication evolve twice independently? *Nucleic Acid Res.* **27**: 3389-3401

Line, M. A.: 2002, The enigma of the origin of life and its timing. *Microbiology* **148**: 21-27

Mushegian, A. R. and Koonin, E. V.: 1996, A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA* **93**: 10268-10273

Ochman, H., Lawrence, J. G., and Groisman, E. A.: 2000, Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299-304

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

Oparin, A. I.: 1961, *Life: its nature, origin and development*: Oliver and Boyd, Edinburgh,

Philippe, H. and Forterre, P.: 1999, The rooting of the universal tree of life is not reliable. *J. Mol. Evol.* **49**: 509-523

Reanney, D. C.: 1979, RNA splicing and polynucleotide evolution. *Nature* **227**: 597-600

Rich, A.: 1962, On the problems of evolution and biochemical information transfer. *In* Kasha, M. and Pullman, B.: eds, *Horizons in Biochemistry*. Academic Press, New York, pp. 103-126

Riera, J., Robb, F. T., Weiss, R., and Fontecave, M.: 1997, Ribonucleotide reductase in the archaeon *Pyrococcus furiosus*: a critical enzyme in the evolution of DNA genomes. *Proc. Natl. Acad. Sci. USA* **94**: 475-478

Rivera, M. C. and Lake, J. A.: 2004, The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**: 152-155

Schmid, S. R. and Linder, P.: 1992, D-E-A-D protein family of putative RNA helicases. *Mol. Microbiol.* **6**: 283-292

Seifert, J. L., Martin, K. A., Abdi, F., Wagner, W. R., and Fox, G. E.: 1997, Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *J. Mol. Evol.* **45**, 467-472

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

Snel, B., Bork, P., and Huynen, M. A.: 2002, Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* **12**: 17-25

Tauer, A. and Benner, S. A.: 1996, The B_{12}-dependent ribonucleotide reductase from the archaebacterium *Thermoplasma acidophila*: An evolutionary solution to the ribonucleotide reductase conundrum. *Proc. Natl. Acad. Sci. USA* **94**: 53-58

Tatusov, R. L., Koonin, E. V., and Lipman, D. J.: 1997, A genomic perspective on protein families. *Science* **278**: 631-637

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

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

Woese, C. R.: 1983, The primary lines of descent and the universal ancestor. *In* D. S. Bendall: (ed), *Evolution from Molecules to Men* : Cambridge University Press, Cambridge,, pp. 209-233

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

Woese, C. R.: 1998, The universal ancestor. *Proc. Natl. Acad. Sci. USA* **95**: 6854-6859

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

Zhaxybayeva O. and Gogarten, J. P.: 2004, Cladogenesis, coalescence and the evolution of the three domains of life. *Trends Genet.* **20**: 182-187

Zhaxybayeva O., Lapierre, P., and Gogarten, J. P.: 2004, Genome mosaicism and organismical lineages. *Trends Genet.* **20**: 254-260

On the early evolution of sensory responses: when did life first begin to perceive its surroundings?

Antonio Lazcano, Arturo Becerra, and Luis Delaye

Facultad de Ciencias
Universidad Nacional Autónoma de México
Apartado Postal 70-407
Cd. Universitaria, 04510 México, D.F.
MEXICO
E-mail: alar@correo.unam.mx

Abstract

Recognition of the sensing properties of RNA molecules involved in the control of several major metabolic routes, including biosynthesis of purines, vitamins and amino acids, have been interpreted as evidence of regulatory mechanism in the RNA world. Alarmones, on the other hand, are modified ribotides that form part of regulatory systems that are activated when cells sense stress conditions. The chemical nature of alarmones and their relatively simple biosynthetic pathways, which involve nucleotide-modifying enzymes, suggest that some of them like cyclic AMP and ZTP may have appeared in an RNA/protein world, prior to the evolutionary development of DNA genomes. This suggests an early start for intra- and extracellular chemically-based communication systems based on cAMP and other alarmones. The phylogenetic distribution of alarmones and their biosynthetic enzymes suggest a complex evolutionary history that includes the independent emergence of signal molecules. However, profile-based computer searches have demonstrated the presence of its adenylyl cyclase 2 (AC2) homologs in the three major biological lineages (Bacteria, Archaea and Eucarya), suggesting that this cAMP-synthesizing enzyme was already present in the last common ancestor of all extant lifeforms. Biochemical studies and genomic analysis have failed to identify ppGpp and pppGpp and their biosynthetic enzymes in Archaea. These results, combined with the diversity of non-homologous eubacterial ACs, indicate that Bacteria have explored the widest range of alarmone biosyntheses and utilization.

I. Introduction

Living beings do not live in blissful isolation. They are constantly exchanging matter and energy with their surroundings and with other organisms, whether of their same species or not, with whom they share their environments. Such interactions involve a wide array of different sensory systems and signaling mechanisms, which in some cases partake not only in extracellular responses but also in the regulation of metabolic processes within cells. While the evolutionary history of animals and plants wears testimony to a number of major innovations that have shaped their different sensory systems, there is also conclusive evidence that many different traits of their sensory systems, including cell-to-cell signaling, were assembled from components that originated in the prokaryotic ancestors of nucleated cells.

The traditional depiction of bacteria as mere infectious agents and their 19th century classification as thallophyta, i.e, lower plants, conceived them as passive microbes with little or no abilities to perceive their surroundings and to respond to outside stimuli. Although this simplistic scheme has been challenged by an acknowledgement of the unparallel metabolic diversity of prokaryotes and their surprising geographical distribution, it is often not realized that they are also endowed with elaborate sensory systems that allow them to monitor and respond to their environment. It is reasonable to assume that early lifeforms must have developed very rapidly receptors capable of sensing their environment, taking cues from intra- and extracellular chemical signals. This possibility is supported by the existence of naturally occurring mechanisms for transcription control that depend solely on the sensing of small molecules by RNA and natural metabolite-responsive ribozymes, i.e., riboswitches. Characterization of these riboswitches, which are known to be present in the three major biological domains, has demonstrated the existence of regulatory properties in RNA molecules and the possibility of the development of sensing abilities prior to the emergence of proteins (Winkler et al., 2004).

Stress conditions are also sensed by alarmones (*alarm* + hor*mones*, cf. Watson et al., 1987), which are small signal metabolites that are rapidly synthesized when specific intracellular starving conditions develop due to lack of amino acids or sugars, or in response to environmental insults such as heat, ethanol and a wide variety of oxidants. Alarmones are modified purine-ribotides (Figure 1) and include cyclic AMP (cAMP, adenosine 3',5'-cyclic monophosphate), cGMP (guanosine 3',5'-cyclic monophosphate), AppppA (diadenosine tetraphosphate), ZTP (5-amino- 4-imidazole carboxamide riboside 5'-triphosphate), which signals folate deficiency (Bochner and Ames, 1982), as well as ppGPP (guanosine tetraphosphate) and pppGpp (guanosine pentaphosphate), which play a direct role in the inhibition of stable RNA synthesis in Bacteria, (Stephens et al., 1975). As argued here, the chemical nature of some alarmones like cAMP, its biological distribution, and mode of action all support the possibility that they are part of a stress-sensing regulatory system established in RNA/protein world prior to the evolutionary development of DNA genomes (Figure 2).

II. The evolutionary history of chemical signals and sensory systems: some cautionary notes

The awareness that genes and 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 simultaneous 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. This is particularly true of rRNA. Comparison of small subunit ribosomal RNA (16/18S rRNA) sequences led to the construction of a trifurcated, unrooted tree in which all known organisms can be grouped in one of three major monophyletic cell lineages (Woese and Fox, 1977), now referred to as the domains Bacteria, Archaea, and Eucarya (Woese et al., 1990).

Analysis of an increasingly large number of completely sequenced cellular genomes has revealed major discrepancies with the topology of rRNA trees. There are manifold reasons for such discrepancies, including inadequate biodiversity sampling, polyphyletic gene losses, convergence and polyphily, and unequal rates of evolution, among others. However, very often these differences have been interpreted as evidence of horizontal gene transfer (HGT) events between different species, questioning the feasibility of the reconstruction and proper understanding of early biological history (Doolittle, 1999). There is clear evidence that genomes have a mosaic-like nature whose components come from a variety of sources (Ochman et al., 2000). Depending on their different advocates, a wide spectrum of mix-and-match recombination processes have been described, ranging from the lateral transfer of few genes via conjugation, transduction or transformation, to cell fusion events involving organisms from different domains.

Nevertheless, there is evidence that the historical record of past evolutionary events has not been completely lost. Comparisons of combined ortholog protein data sets that exclude sequences that may have undergone lateral transfer are consistent with rRNA trees (Brown et al., 2001). Genomic trees also exhibit an excellent broad-level agreement

with rRNA-based phylogenies (Fitz-Gibbon and House, 1999; Snel et al., 1999; Tekaia et al., 1999). Genomic trees are not cladograms but phenograms, i.e., they are hierarchical representations of similarities and differences in gene content, where the presence or absence of a sequence is counted as a character. Since different lineages evolve at *different rates, such overall similarity may be an equivocal indicator of genealogical relationships.* The robustness exhibited by these different methodologies indicates that although lateral gene transfer has played major role in cellular evolution, massive lateral transfer events between distant groups has not completely uprooted the tree of life.

Because of their chemical nature, alarmones cannot be used to construct such phylogenies. Nonetheless, important insights on their evolutionary history can be accomplished by comparing the phylogeny and genomic distribution of the enzymes involved in their biosyntheses. Of course, reticulate phylogenies greatly complicate the understanding of the evolution of sensory systems. It has been argued that a significant number of the genes involved in animal cell-to-cell signaling, which is essential in the nervous, neuroendocrine and immune systems, have undergone massive horizontal transfer (Iyer et al., 2004), and there is evidence of lateral transfer between Gram positive bacteria and Archaea of the chemotactic sensory system (Faguy and Jarrell, 1999). Sequences involved in alarmone biosynthesis have not escaped this fate. However, it appears that the last common ancestor (LCA) of all extant lifeforms, i.e., the cenancestor (Figure 2) was a modern-type of prokaryote already endowed with alarmones and elaborate sensory systems. This conclusion is supported by the evolutionary conservation of the intracellular distribution of methyl-accepting chemotaxic proteins, which partake in the chemotactic response system of prokaryotes, and which concentrate in the cell poles of an evolutionary diverse array of bacterial and archaeal species (Gestwicki et al., 2000). When and how did these molecular sensory and regulatory systems first evolve?

III. The evolution and phylogenetic distribution of alarmones

The presence of alarmones has been reported in all three primary biological lineages. Current evidence indicates that their highest structural and functional diversity is found in the Bacteria (Table 1). It is likely that this conclusion is not shaped by inadequate biodiversity sampling of the Archaea and the Eucarya, or by our incomplete knowledge of the basic biochemical processes of these groups. The biosynthesis of guanosine 5'-diphosphate 3'-diphosphate (ppGpp or guanosine tetraphosphate), and guanosine 5'-triphosphate 3'-diphosphate (pppGpp, guanosine pentaphosphate), which act as general signals for amino acid starvation conditions by adjusting the rates of rRNA and tRNA syntheses (Stephens et al., 1975), is mediated by GTP pyrophosphokinase (Table 1). This is a diphosphotransferase encoded by the *relA* gen, which was first described in enterobacteria. Biochemical analysis failed to identify ppGpp and pppGpp among Euryarchaeota and Crenarcheaota (Cellini et al., 2004), suggesting that they are a typical eubacterial trait. The search for *relA* homologs using a BLAST analysis (Atschul et al, 1997) has confirmed this result (data not shown), as no related molecules are found in the completely sequenced archaeal genomes available as of August 2004. The presence of *relA*-related sequences in red algae *Cyanidioschyzon merolae* and in *Arabidposis thaliana* may be due to lateral gene transfer, as no other eukaryotic homologs have been found. Diadenosine tetraphosphate (AppppA, or diadenosine 5',5'''-P^1,P^4-tetraphosphate), which accumulates in the presence of a bacteriostatic quinone and other oxidizing agents (Lee et al., 1983) can be synthesized by aminoacyl-tRNA synthase (Lee et al., 1983) and by ATP adenylyl transferase. A BLAST search (Atschul et al, 1997) using as query sequence the *Saccharomyces cerevisiae* ATP adenylyl transferase, however, indicated the presence of homologs only among cyanobacteria, ascomycetes, the red algae *C. merolae* and the slime mold *Dictyostelum discoideum* (Table 1).

By contrast, ApppA, which is synthesized by diadenosine triphosphatase 5',5'-P^1,P^4-triphosphatase, appears to be one of the most widely distributed alarmones, as suggested by the presence of the corresponding gene in the three primary domains. The same may be true of ZTP (Figure 1), which has been implicated in folate starvation signaling (Bochner and Ames, 1982). Its biosynthesis involves the transfer of a pyrophosphate group of 5-phosphoribosyl-1-pyrophosphate to the riboside monophosphate by 5-phosphoribosyl-1-pyrophosphate synthetase, i.e., PRPP synthetase (Sabina et al., 1984), an almost universally distributed, highly conserved enzyme that plays a key role in purine and histidine biosyntheses (Figure 3). The search for homologs in all completely sequenced cellular genomes indicated that it is absent in the chlamydial and rickettsial genomes, a result which can be explained as one of of the many secondary losses that these parasitic species have undergone.

Cyclic AMP appears to the most widely distributed alarmone (Table 1). This ubiquity probably reflects both its antiquity and the polyphyletic origin of adenylyl cyclase (AC) enzymes involved in its biosynthesis (Danchin, 1993). ACs have divided on the basis of conserved motifs into several classes, which include cGMP biosynthetic enzymes (Sismeiro et al., 1998; Baker and Kelly, 2004), which appear to be descendants of ACs (Shenoy and Visweswariah, 2004). Structural analysis of class III ACs have led to their classification into six different families of proteins, with no sequence similarities among them (Shenoy and Visweswariah, 2004). Class III adenylyl cyclases are also known as the universal class, but this name is not reflected in their biological distribution. As summarized by Shenoy and Visweswariah. (2004), they are absent from several major Bacteria clades (the *Bacillus/Clostridium* and the *Deinococcus* groups, from the Chlamydiae, and the ε- and δ-proteobacteria). Highly divergent class III ACs are present

in few methanogenic euryarchaeota (*Methanosarcina acetivorans* strain C2A, *Methanopyrus kandleri*, and *Methanothermobacter thermoautotrophicum*), but otherwise they are absent from the Archaea, (Table 1).

Biochemical reports of the occurrence of cyclic AMP in archaea (Leichtlin et al., 1986) very likely reflected the activity of adenylyl cyclase 2 (AC2). Sequence analysis demonstrated that this enzyme, first described in the bacteria *Aeromonas hydrophila*, has no sequence similarity to previously reported ACs (Sismeiro et al., 1998). Analysis of protein databases revealed the existence of AC2-related sequences in *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum* and *Archaeoglobus fulgidus*, which lead to suggestions that its presence in *A. hydrophila* was due to lateral gene transfer from a methanogen (Sismeiro et al., 1998). A profile-based search (Pfam) in a dataset that includes approximately 140 completely sequenced cellular genomes (data not shown) suggests otherwise. As summarized in Table 1, with few exceptions, which include three euryarchaeal species (*Thermoplasma acidophilum*, *T. volcanium*, and *Picrophilus torridus*) and the episymbiotic archaeon *Nanoarchaeum equitans*, AC2 sequences are universally distributed. This ubiquity suggest that the last common ancestor (LCA) of all extant lifeforms, i.e., the cenancestor, was already endowed with AC2 signal molecules.

IV. Sensing in the RNA world

Since all extant cells are endowed with DNA genomes, the most parsimonious conclusion is that this genetic polymer was already present in the cenancestral population. Since the late 1960's, however, Woese (1967), Crick (1968) and Orgel (1968) argued independently that RNA had preceded not only DNA but also proteins (Figure 2), a stage now referred to as the RNA world (Gilbert, 1986; Joyce, 2002). As reviewed elsewhere (Lazcano et al., 1988), the possibility that RNA played a much more conspicuous role in early biological evolution is strongly supported by (a) the genetic information storage capacity of RNA molecules; (b) the central role that different RNAs in protein synthesis; (c) the evidence that many coenzymes are in fact ribonucleotide derivatives; (d) that fact that with few exceptions, all cellular and most viral DNA genomes require RNA primers for their replication; (e) the evidence that ribonucleotides are metabolic precursors of deoxyribonucleotides; and (f) the increasingly large number of reported catalytic activities of RNA molecules, which include their direct involvement in ribosome-mediated synthesis of peptide-bonds (Nissen et al., 2000).

The roots of cAMP utilization might lie in the RNA world. The existence of naturally occurring mechanisms for transcription control that depend solely on the sensing of small molecules by RNA and natural metabolite-responsive ribozymes, i.e., riboswitches (Winkler et al., 2002; Vitreschak et al., 2004), demonstrate that RNA molecules are endowed with regulatory properties which may have functioned as metabolite sensors prior to the emergence of proteins (Winkler et al., 2004). *In vitro* selection of aptamers, which bind to the adenosine moiety of ATP, NAD, and acetyl-CoA (Sassanfar and Szostak, 1993; Burgstaller and Famulok, 1994; Burke and Gold, 1997), as well as the existence of cAMP-, cGMP- and cCAMP-responsive ribozymes (Koizumi et al., 1999) suggest that cyclic AMP may have been part of the RNA world. These results, combined with the engineering of hammerhead ribozymes that self-destruct by self-cleavage reactions in the presence of cAMP or cGMP (Koisumi et al., 1999b), give additional credibility to the possibility of an ancestral all-RNA-based metabolism with regulatory mechanisms based on binding of cyclic nucleotides and other effector molecules. Since the intramolecular cyclization reaction required for cAMP from its AMP precursor is mechanistically equivalent to polynucleotide elongation, it could have been catalyzed in an RNA world devoid of proteins by a ribozyme with polymerizing abilities.

V. Alarmones and the RNA/protein world

Comparative genomics does not provide direct evidence of an RNA world or its sensing abilities. However, at the time being phylogenetic analysis of some molecular markers can be extrapolated to a period of cellular evolution in which protein biosynthesis was already in operation, i.e., an RNA/protein world. Older stages are not yet amenable to molecular phylogenetic analysis, but clues of the RNA/protein world stage can be inferred when whole genome analysis is performed. ORFs whose products synthesize, degrade, or interact with RNA, are among the most highly conserved sequences common to all known genomes. Their conservation strongly validates the idea of an RNA/protein world that existed prior to the extant DNA-encoding cells, and provide insights into an early stage in cell evolution during which RNA played a much more conspicuous biological role (Tekaia et al., 1999; Delaye and Lazcano, 2000; Anantharaman et al., 2002). The divergence of AC2 homologs puts them apart from this set of highly conserved sequences. Nonetheless, the profile-based searches (Pfam) reported here (Table 1) suggest that AC2 may have been part of the catalytic repertoire of the last common ancestor, and may have evolved among its ancestors in the RNA/protein world. The hypothesis of pre-DNA AC2s in RNA/protein cells is consistent with its use of a ribonucleotide as a substrate, which may be interpreted as a vestige from an era when RNA and ribonucleotides played a more conspicuous role in biological processes. This hypothesis is supported by the intrinsic simplicity of the intramolecular cyclization reaction required for cAMP from its AMP precursor, which is involves a nucleophilic attack of the 3'-OH ribose moiety associated with pyrophosphate release.

The structural homology between the class III bacterial cAMP-synthesizing adenylyl cyclases and the universally-distributed highly conserved palm domain of manifold polymerases (Artymiuk et al., 1997; Bryant et al., 1997; Aravind and Koonin, 1999), could also be interpreted as evidence that these enzymes also evolved in the RNA/protein world (Figure 4). This raises the possibility that the LCA endowed with two phylogenetic unrelated cAMP-synthesizing enzymes. This hypothesis, however, is at odds with the phylogenetic distribution of class III ACs, which, with the sole exception of few methanogens, are not found in Archaea. It could be argued that absence of these set of enzymes in the Archaea is due to a secondary loss that took place soon after their divergence from the Bacteria and the Eucarya. We favor an alternative, less-parsimonious explanation, according to which class III AC resulted from the evolutionary recruitment of a polymerase palm domain homolog in the line leading to Bacteria. This hypothesis is consistent with the surprising diversity of non-homologous bacterial adenylyl cyclases (Table 1), that demonstrates the polyphyletic origin of enzymes involved in the intramolecular cyclizations reactions required for cAMP biosynthesis.

VI. Conclusions

Although how the transition from the non-living to the living took place is not yet known, there is strong evidence suggesting that the RNA world may have been the first organized biological world. The evolutionary advantages that the early emergence of simple sensing mechanisms and RNA-based control mechanisms, specially once spatial separation appeared, are easy to understand, and may explain the existence of naturally occurring mechanisms for transcription control that depend on riboswitches.

As argud here, its is possible that cAMP and perhaps other alarmones are fossils of an intermediate stage in the evolution of chemical signaling and metabolite sensing that followed the RNA-dependent sensing mechanisms that may have existed in the RNA world. This hypothesis implies an early start for intra- and extracellular chemically-based communication systems dependent on cAMP and other alarmones. Like nucleic acid polymerases, ACs appear to be of polyphyletic origin. It is likely that this reflects the pervasive roles of nucleotides and their availability in the intracellular environment. It is somewhat surprising, however, that pyrimidine-ribotides are not part of the inventory of intra- and extracellular alarmones. It has been proposed that the reverse cyclase reaction may generate ATP from cAMP (Barzu and Danchin, 1994). It is also possible, however, that the abundance of purine moieties in alarmones (Table 1) reflects the intrinsic chemical stability of purines as compared to pyrimidines.

The presence of cAMP and the widely distributed AC2s found in all three major biological domains (Bacteria, Archaea, and Eucarya) demonstrates that cyclic AMP has been conserved across prokaryotes and nucleated cells. The fact that more than one biosynthetic enzyme to cAMP have been identified emphasizes the fact that cyclic AMP has been selected as a secondary messenger by manifold convergent processes. The phylogenetic distribution of ACs defies a simple explanation. The available information indicates that the genes involved in cAMP synthesis have undergone branch-specific gene duplications, lateral transfer events, and secondary losses, and appear to be of polyphyletic origin. Such diversity and unequal distribution indicates not only the need for more detailed studies of alarmone evolution, but also points out to the wide diversity of sensing strategies and signal mechanisms that the biosphere has successfully explored during four billion years of constant change.

Acknowledgements

Work reported here was supported by project DGPA IN-111003 (UNAM, Mexico). Part of this work was completed during leave of absence at the Department of Statistics, Rice University, where one of us (A.B.) enjoyed the hospitality of Dr. Janet Seifert. Due thanks are given by L.D. to the 2000 Planetary Biology Internship NASA program for its support for a training visit at the laboratory of Dr. Peter Gogarten.

Figure captions

Figure 1. Alarmones are modified ribotides, and include cyclic AMP (cAMP, adenosine 3`,5`-cyclic monophosphate), ZTP (5-amino- 4-imidazole carboxamide riboside 5'-triphosphate), ApppA (diadenosine 5`,5```-P^1,P^3 triphosphate), and AppppA (diadenosine 5`,5````-P^1,P^4 tetrahosphate)

Figure 2. Origin and early evolution of life. Adenylyl cyclase 2 is hypothesized to be the oldest alarmone-synthezing enzyme which first evolved in the RNA/protein stage.

Figure 3. Biosynthesis of purines, histidine, and ZTP. The diagram emphasizes the key role of PRPP synthetase

Figure 4. Biological distribution of the DNA polymerase β palm domain and its homologs.

References

Anantharaman, V., Koonin, E. V., and Aravind, L. 2002. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acid Res.* **30**: 1427-1464

Aravind, L. and Koonin, E. V. 1999. DNA polymerase β-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history. *Nucleic Acid Res.* **27**: 1609-1618

Artymiuk, P. J., Poirrette, A. R., Rice, D. W., and Willett, P.1997. A polymerase I palm in adenylyl cyclase? *Nature* **388**: 33-34

Atschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein databases search program. Nucleic Acid Res. 25: 3389-3402

Baker, D. A. and Kelly, J. M. 2004. Structure, function and evolution of microbial adenylyl and guanylyl cyclases. *Mol. Microbiol.* **52**: 1229-1242

Barzu, O. and Danchin, A. 1994. Adenylyl cyclases: a heterogeneous class of ATP-utilizing enzymes. *Prog. Nucleic Acid Res. Mol. Biol.* **49**: 241-283

Bochner, B. R. and Ames, B. N. 1982. ZTP (5-amino 4-imidazole carboxamide riboside 5'-triphosphate: a proposed alarmone for 10-formyl-tetrahydrofolate deficiency. *Cell.* **29**: 929-937

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., and Stanhope, M. J. 2001. Universal trees based on large combined protein sequence datasets. *Nature Genetics.* **28**: 281-285

Bryant, S. H., Maderj, T., Janin, J., Liu, Y., Ruoho, A. E., Zhang, G., and Hurley, J. H. 1997. A polymerase I palm in adenylyl cyclase? *Nature* **388**: 34

Burke, D. H. and Gold, L. 1997. RNA aptamers to the adenosine moiety of S-adenosyl methionine: structural inferences from variations on a theme and the reproducibility of SELEX. *Nucleic Acid Res.* **25**: 2020-2024

Burgstaller, P. and Famulok, M. (1994) Isolation of RNA aptamers for biological cofactors by *in vitro* selection. *Angewandte Chemie* **106**: 1163-1166

Cellini, A., Scoarughi, G. L., Poggiali:, P., Santino, I. Sessa, R., Donini, P., Cimmino, C. 2004. Stringent control in the archael genus *Sulfolobus*. *Res. Microbiol.* **155**: 98-104

Crick, F.H. C. 1968. The origin of the genetic code. *J. Mol. Biol.* **38**: 367-379

Danchin, A. 1993. Phylogeny of adenylyl cyclases. *Adv. Second Messenger Phosphoprotein Res.* **27**: 109-162

Delaye, L. and Lazcano, A. (2000) RNA-binding peptides as molecular fossils *In* J. Chela-Flores, G. Lemerchand, and J. Oró (eds)*Origins from the Big-Bang to Biology: Proceedings of the First Ibero-American School of Astrobiology* (Klüwer Academic Publishers, Dordrecht), pp. 285-288

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

Faguy, D. M. and Jarrell, K. F. 1999. A twisted tale: the origin and evolution of motility and chemotaxis in prokaryotes. *Microbiology.* **15**: 279-281

Fitz-Gibbon, S. T. and House, C. H. 1999. Whole genome-based phylogenetic analysis of free-living organisms. *Nucleic Acids Res.* **27**: 4218-4222

Gestwicki, J. E., Lamanna, A. C., Harshey, R. M., McCarter, L. L., Kiessling, L. L. and Adler, J. 2000. Evolutionary conservation of methyl-accepting chemotaxis protein location in *Bacteria* and *Archaea*. *J. Bacteriol.* **182**: 6499-6502

Gilbert, W. 1986. The RNA world. *Nature.* **319**: 618

Iyer, L. M., Aravind, L., Coon, S. L., Klein, D. C., and Koonin, E. V. 2004. Evolution of cell-cell signaling in animals: did late horizontal gene transfer from bacteria have a role? *Trends Genet.* **20**: 292-299

Joyce, G. F. 2002. The antiquity of RNA-based evolution. *Nature* **418**: 214-221

Koizumi, M., Kerr, J. N., Soukup, G. A., and Breaker, R. R. 1999a. Allosteric enzymes sensitive to the second messengers cAMP and cGMP. *Nucleic Acid Symp. Ser* **42**: 275-276

Koizumi, M., Soukup, G. A., Kerr, J. N. and Breaker, R. R. 1999b. Allosteric selection of ribozymes that respond to the second messengers cGMP and cAMP. *Nature Struct. Biol.* **6**: 992-994

Lazcano, A., Guerrero, R., Margulis, L., Oró, J. 1988. The evolutionary transition from RNA to DNA in early cells. *J. Mol. Evol.* **27**: 283-290

Lee, P. C., Bochner, B. R., and Ames, B.N. 1983. Diadenosine 5',5'''-P^1,P^4-tetraphosphate and related adenylylated nucleotides in *Salmonella typhimurium*. *J. Biol. Chem* . **258**: 6827-6834

Leichtling, B. H., Rickenberg, H. V., Seely, R. J., Fahrney, D. E., and Pace, N. R. 1986. The occurrence of cyclic AMP in archaebacteria. *Biochem. Biophys. Res. Commun.* **136**: 1078-1082

Nissen, P., Hansen, J., Ban., N., Moore, P. B. and Steitz, T. A. 2000. The structural basis of ribosome activity in peptide bond synthesis. *Science.* **289**: 920-930

Ochman, H., Lawrence, J. G.n and Groisman, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature.* **405**: 299-304

Orgel, L. E. 1968. Evolution of the genetic apparatus. *J. Mol. Biol.* **38**: 381-393

Orgel, L. E. 2004. Prebiotic chemistry and the origin of the RNA world. *Crit. Rev. Biochem. Mol. Biol.* **39**: 99-123

Sabina, R. L., Holmes, E. W., and Becker, M. A. 1984. The enzymatic synthesis of 5-amino-4-imidazolecarboxamide riboside triphosphate(ZTP). *Science* **223**: 1193-1195

Sassanfar, M. and Szostak, J. W. 1993. An RNA motif that bind ATP. *Nature* **364**: 550-553

Shenoy, A. R. and Visweswariah, S. S. 2004. Class III nucleotide cyclases in bacteria and archaebacteria: lineage-specific expansion of adenylyl cyclases and a dearth of guanylyl cyclases. *FEBS Lett.* **561**: 11-21

Sismeiro, O., Trotot, P., Biville, F., Vivares, C., and Danchin, A. (1998) *Aeromonas hydrophila* adenylyl cyclase 2: a new class of adenylyl cyclases with thermophilic properties and sequence similarities to proteins from hyperthermophilic archaebacteria. *J. Bacteriol.* **180**: 3339-3344

Snel, B., Bork, P., and Huynen, M. A. 1999. Genome phylogeny based on gene content. *Nat. Genet.* **21**: 108-110

Stephens, J. C., Artz, S. W., and Ames, B. C. 1975. Guanosine 5'-diphosphate 3'-diphosphate (ppGpp): positive effector for histidine operon transcription and general signal for amino-acid deficiency. *Proc. Natl. Acad. Sci. USA.* **72**: 4389-4393

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

Vitreschak, A. G., Rodionov, D. A., Mironov, A. A., and Gelfand, M. S. 2004. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet.* **20**: 44-50

Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A., and Weiner, A. M. (1987) *Molecular Biology of the Gene* vol. I (4th ed., Benjamin Cummings Publ., Menlo Park), p.498

Winkler, W., Nahvi, A., and Breaker, R. R. 2002. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**: 952-956

Winkler, W., Nahvi, A., Roth, A., Collins, J. A. Breaker, R. R. 2004. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428**: 281-286

Woese, C. R. 1967. *The Genetic Code: the molecular basis for gene expression* (Harper and Row, New York)

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

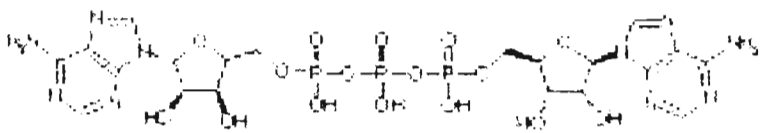
Woese. C. R., Kandler, O., and Wheelis, M. L. 1990. Towards a natural system of organisms, proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87**: 4576-4579

Alarmone	Enzyme	First characterized	Phylogenetic distribution		
			Bacteria	Archaea	Eucarya
cAMP	AC*[1] class I (enterobacterial)	*Escherichia coli*	γ-Proteobacteria		
	AC class II (calmodulin-activated)	*Bacillus anthracis*	Proteobacteria, and one Firmicutes		
	AC class III (palm domain)	*Rattus norvegicus*	Several Bacteria*[2]	Few Euryarchaea*[3]	Several Eucarya*[4]
	AC class IV (AC2)	*Aeromonas hydrophila*	Several Bacteria*[5]	Several Archaeas*[6]	Several Eucarya*[7]
	AC class V	*Prevotella ruminicola*	*P. ruminicola* (Bacteroides)		
	AC class VI	*Rhizobium etli*	α-Proteobacteria, few Spirochaetes		
ZTP	PRPP synthetase (5 -phosphoribosyl-I-PPsynthetase)	*Homo sapiens*	Most Bacteria	Most Archaea	Most Eucarya
ppGpp	GTP pyrophosphokinase	*Salmonella thiphimurium*	Several bacteria*[8]		*A. gambiae*, red algae and plants*[9]
AppppA	ATP adenylyltransferase	*Sacharomyces cerevisiae*	Cyanobacteria		Ascomycetes, red algae and slime mold
ApppA	Diadenosine 5, 5-P1, P3-triphosphatase	*Sacharomyces cerevisiae*	Most Bacteria	Most Archaea	Most Eucarya

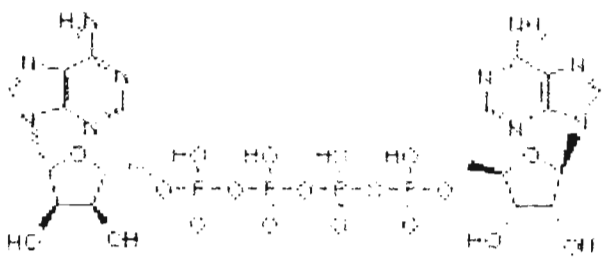
Table 1. Phylogenetic distribution of alarmone synthesizing enzymes. Red alga refers to *Cyanidioschyzon merolae*; slime mold refers to *Dictyostelium discoideum*. *[1]AC: adenylate cyclases. *[2]Cyanobacteria, Actinobacteria, Proteobacteria, Planctomycetes, Spirochaetes. Not present in *Bacillus/Clostridium*, *Deinococcus*, Chlamidiae, and ε- and δ-Proteobacteria (Shenoy, et al., 2004). *[3]Present in *Methanosarcina acetivorans* str. C2A, *Methanopyrus kandlery*, and *Methanothermobacter thermoautotrophicus* (Shenoy, et al., 2004). *[4]Lacking in *A. thaliana*. *[5] Cyanobacteria, Actinobacteria, Bacteroidetes, Proteobacteria, Planctomycetes, Spirochaetes, Firmicutes, Bacillales.*[6] AC2 is not present in the Archaeas *Thermoplasma acidophilum*, *T. volcanium* and *Picrophilus torridus*. *[7]Metazoa, Mycetozoa, Diplomonadida and Viridiplantae (lacking in Fungi). *[8]Aquificae, Actinobacteria, Deinococcus-thermus, Bacteroidetes, Proteobacteria, Spirochaetes, Firmicutes, Cyanobacteria, Thermotogae, Fusobacteria, Chlorobi. *[9]*Anopheles gambiae*, Diptera.

cAMP

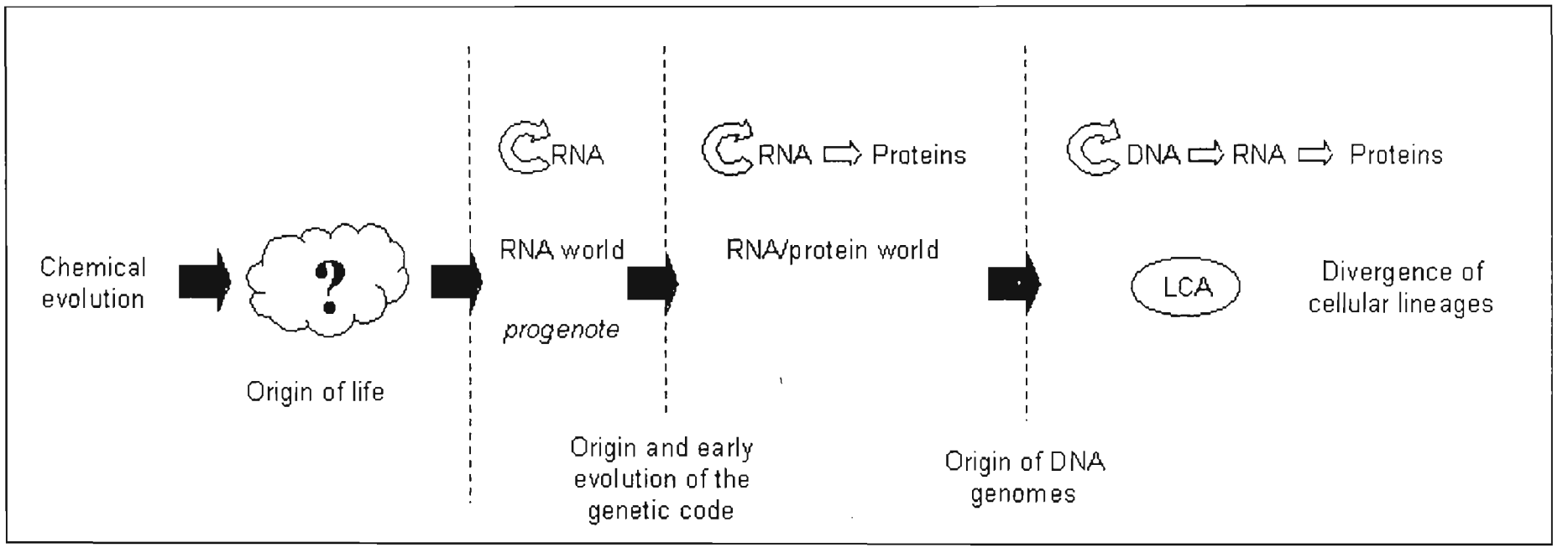
ZTP

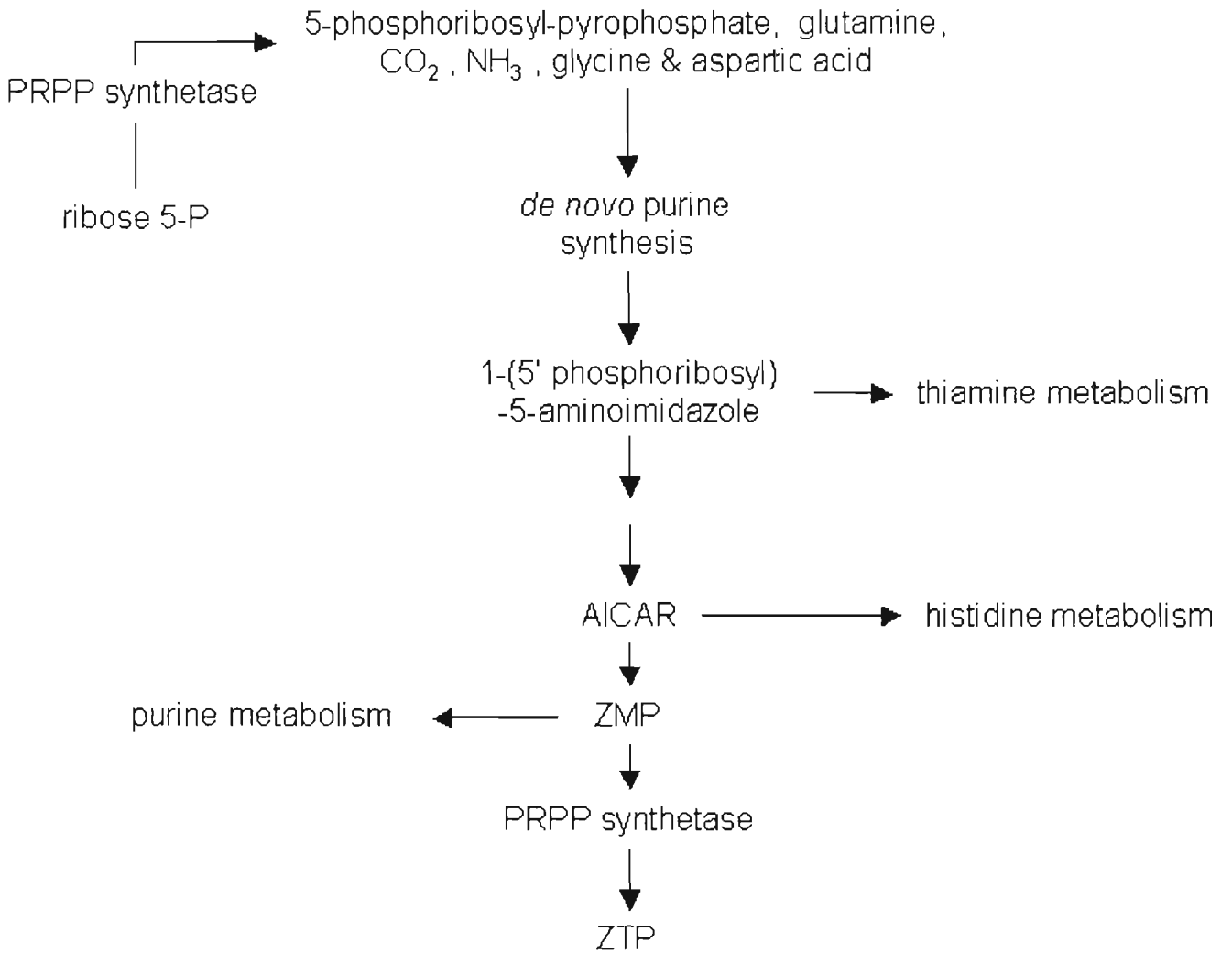


ApppA



AppppA


Chemical evolution
?
Origin of life
RNA
RNA world
progenote
Origin and early evolution of the genetic code
RNA ⇨ Proteins
RNA/protein world
DNA ⇨ RNA ⇨ Proteins
LCA
Origin of DNA genomes
Divergence of cellular lineages


5-phosphoribosyl-pyrophosphate, glutamine,
CO_2 , NH_3 , glycine & aspartic acid
PRPP synthetase
ribose 5-P
de novo purine synthesis
1-(5' phosphoribosyl) -5-aminoimidazole
thiamine metabolism
AICAR
histidine metabolism
purine metabolism
ZMP
PRPP synthetase
ZTP

RNA dependent RNA polymerization

(Poliovirus 3d polymerase)

DNA dependent RNA polymerization

(T7 RNA polymerase)

RNA dependent DNA polymerization

(Reverse transcriptase)

DNA dependent DNA polymerization

(DNA polymerase I)

Palm domain

(~124 residues)

Adenylate and Guanylate cyclase activity

(AC class III)

Ancient RNA-binding domains: relics from early protein evolution

Luis Delaye[1*], Federico Abascal[2], José Maria Fernández[2], Alfonso Valencia[2] and Antonio Lazcano[1]

[1]Laboratorio de Microbiología
Departamento de Biología Evolutiva
Facultad de Ciencias, UNAM
México, D.F. MEXICO
Phone (55)56224823
Fax (55) 56224828

[2]Protein Design Group
Centro Nacional de Biotecnología
Universidad Autónoma de Madrid
Madrid, SPAIN
Phone +34.91.585.45.70
Fax +34.91.585.45.06

Luis Delaye: infinito@pine.servidores.unam.mx, *Corresponding author
Federico Abascal, fabascal@cnb.uam.es
José Maria Fernández, jmfernandez@cnb.uam.es
Alfonso Valencia, valencia@cnb.uam.es
Antonio Lazcano, alar@correo.unam.mx

Running head: Ancient RNA-binding domains

Abstract

RNA-binding protein domains are some of the oldest polypeptides we can recognize. Among them, there is evidence of some of the oldest process that shaped early cell evolution. These includes gene duplications, gene fusions, *patchwork* evolution in ribosome assembly and an ancient mechanism of gene regulation.

Lists of abbreviations:
last common ancestor (LCA)

RNA first, DNA last and the RNA/protein world

The theory of the RNA-world suggest that this biomolecule played a prominent roll during the first steps of the evolution of life on Earth, functioning as both an information carrier polymer and a catalytic ribozyme [1]. Based on the following functional and historical arguments [2] it has been argued that proteins and not DNA evolved directly from the RNA world: a) the sulfur-based biochemistry of ribonucleotide reduction that involves a free-radical may be out of the scope of RNA biochemistry (although it has been proposed that RNA could have used another mechanism of ribonucleotide reduction, such as an attached purine followed by acid-catalysed elimination of the ribose 2′ -hydroxyl [3]); b) despite the high number of divergences between families of ribonucleotide reductases, all of them share the same basic biochemistry, suggesting that this is the only mechanism nature has found for ribonucleotide reduction; c) proteins are synthesized mainly by RNA (the cell machinery responsible of synthesizing proteins is basically made of functional RNA molecules such as the ribosome and the tRNAs), while DNA is synthesized basically by proteins (ribonucleotide reductases, thymidilate synthases, and DNA polymerases, to name a few). A pattern that strongly suggest that proteins are the direct invention of the RNA world, while DNA evolved from protein biochemistry.

The RNA-protein world stage of cellular evolution began with the origin of the genetic code and the first ribosome synthesized polypeptides, and ended with the synthesis of the first DNA genomes. Several important aspects of extant cells, like the structure of the genetic code, the topologies of the first proteins, or very ancient biochemical pathways with universal distribution, presumably evolved during this stage of evolution. The first cells capable of synthesizing ribosome mediated proteins had an enormous selective advantage over those cells which lacked that biochemical ability mainly for: i) their superior catalytic capacities as compared to those of RNA, ii) their capacity to function as very simple transmembranal transporters of ancient cell membranes, iii) or their capacity to bind to the RNA molecule in order to stabilize its structure, as still happens with ribosomal proteins in the ribosome. Although it is not clear at all which of the previous functions the first coded proteins had, it is likely that the function of the translation machinery improved as more RNA-binding protein domains were added to it, thus extending the evolutionary capabilities of cells.

It has been previously suggested that extant RNA-binding protein domains widely conserved in the three main cellular lineages are among the ancient polypeptides still recognizable [4],[5]. Here we discuss some recently discovered aspects of the early evolution of this protein domains.

Ancient RNA-binding protein domains and their *mode* of evolution

Protein domain boundaries and phylogenetic relationships between domains are best identified at the third structure level. In Table 1 we show the structural classification at the family level according to **Structural Classification of Proteins (SCOP) database** [http://scop.mrc-lmb.cam.ac.uk/scop/] of 68 different RNA-binding protein domains with known structure, along with its phylogenetic distribution. Clearly, RNA binding activity has evolved several times, and different protein topologies are suitable for this function (i.e., secondary structure composition can be all beta, all alpha as well as alpha/beta). According to its phylogenetic distribution [5] 35 of this RNA-binding domains were inherited from the last common ancestor (Figure 1). Very likely, some of them may predate from the RNA-protein world. This set is composed by ribosomal proteins, anticodon binding domains from aminoacyl-tRNA synthetases, domains from EF-G and EF-Tu, protein secretion domains (Ffh), tRNA modification domains, regulation of transcription, and general RNA-binding domains like the KH domain. There are some likely cases of domains present in representatives from Archaea, Bacteria and Eucarya because of horizontal gene transfer. RNA-binding protein domains widely distributed in one or two lineages but present in small number in the other(s) lineage(s) likely represent cases of horizontal gene transfer events. For example: cold shock protein B that is widely represented in Bacteria, but only present in *Halobacterium* sp. within the Archaea.

According to SCOP, all the protein domains classified in the same *Family* have a clear evolutionary relationship, while protein domains classified in the same *Superfamily* have a very probable common evolutionary origin. As long as a similar fold at the domain structure level reflects common ancestry between proteins, a clear pattern is the origin of new functions driven by the fusion of pre-existing protein domains (Figure 2). This is clearly the case of ribosomal protein S5 (RpS5). Composed of two different domains, one similar to the double stranded RNA-binding domain from Staufen protein, and the other similar to domain IV from Elongation factor G. A similar pattern is shown for EF-G a multi-domain protein where two of its domains are homologous to ribosomal proteins (ribosomal protein L3 and ribosomal protein S5 for domains II and IV of EF-G respectively). Also, the anticodon nucleic acid binding domain of

Aspartyl tRNA synthetase (N0B) belongs to the superfamily of Nucleic acid-binding proteins from SCOP that includes ribosomal proteins S12 and S17 and are likely homologous. The fact that all these domains have an universal distribution suggest that these domains originated by duplications of pre-existing domains prior the existence of the last *common ancestor (LCA). According to the concept of evolution by tinkering [6], evolution proceeds by the use and* recombination of pre-existing elements in order to generate evolutionary innovations. The outcome of this process is thought to be highly dependent on historical contingences. The pattern of domain fusion here described suggest that evolution by tinkering was active during the very early steps of life on Earth. This evolutionary mechanism is clearly consistent with the notion of domains as the structural and evolutionary units of proteins [7].

According to the *patchwork* hypothesis, primordial biosynthetic pathways were assembled by the recruitment of slow, inefficient enzymes of broad substrate specificity [8]. The *patchwork* hypothesis pretends to explain not the origin of protein structure by the fusion of different domains, but the origin of biochemical pathways by recruitment of enzymes of different evolutionary origins. Ribosomal proteins S12p and S17 that belong to the SCOP family of cold shock DNA-binding domain like, and ribosomal proteins S5 and S9 that belong to the SCOP family of translation machinery components are an example of this mode of evolution. Although those proteins do not participate directly in a metabolic pathway, they do participate in the assembly of the SSU rRNA (Figure 3). Specifically, ribosomal protein S17 is one of the six primary rRNA-binding proteins necessary to start the assembly of the SSU rRNA. Ribosomal proteins S5 and S9 are both secondary proteins in the assembly process, but are still important ones as can be deduced from its universal phylogenetic distribution.

It is known that some of the ribosomal proteins adopt strategies similar for RNA-binding to other non-ribosomal RNA-binding proteins [9]. It has been argued that once the 3D structures of several RNA-binding domains including ribosomal and non-ribosomal proteins became known, it would be possible to know if some of extant RNA-binding domains evolved from ribosomal RNA-binding proteins [10]

This is perhaps the case for some of the domains of the SCOP Superfamily of "Nucleic acid binding proteins". According to SCOP database, ribosomal proteins S12 and S17, the C-terminal domain of eIF-5a, and the cold shock protein B (cspB), belong all to the family of "Cold shock DNA-binding domain-like" and are likely homologous to the anticodon binding domain of aspartyl-tRNA synthetase (all belong to the same superfamily). While the ribosomal proteins and the anticodon binding domain have a universal distribution, the eIF-5a C-terminal domain and the cspB are restricted to the Archaea-Eucarya and Bacteria clades respectively. This suggest that the last two domains were derived from the ribosomal proteins by duplication after the separation of cellular lineages.

We observe a similar pattern in the N-terminal domain of ribosomal protein S5 that has universal distribution, while the double stranded RNA binding domain (dsRBD) is restricted to Bacteria and Eucarya. Perhaps this domain originated in Bacteria by duplication from the ribosomal protein, and then was transferred to Eucarya through endosymbiosis. Also, RNaase P protein may have originated from translation machinery after the divergence of Bacteria [11]. This protein has a distribution restricted to Bacteria while its homologous, the domain IV of EF-G, the C-terminal domain of ribosomal protein S5, and the ribosomal protein S9 have an universal distribution.

The KOW domain (named after Kyrpides, Woese and Ouzounis, [12]) may be another case of a domain that has its origin in ribosomal proteins. This domain has an universal distribution (Figure 1) and is present in a wide number of different proteins. This domain is present in ribosomal protein L24p, ribosomal protein L27e [13], bacterial transcription elongation factor NusG [12], and in the eucaryotic initiation factor 5a N-terminal domain. Because ribosomal protein L24p is the only with an universal distribution among the previous proteins, and because it has an essential function in ribosome assembly process of the LSU rRNA [14] it is very likely that this protein is the anciest from which the other members of this family evolved by duplications after the divergence of the main cellular lineages from the LCA.

A model for early regulation of protein expression

Ribosomal protein S17p performs several functions in *E. coli*. First as a translational repressor protein by regulating the expression of the *str* operon by binding to its own mRNA. Second, it is one of the primary rRNA binding proteins binding directly to 16S rRNA where it nucleates the assembly of the head domain of the 30S subunit. Third, it probably blocks the exit of the E-site tRNA [15]. This ribosomal protein is also universally conserved (Figure 1), thus it may predate from the RNA-protein world. It is likely that its mode of gene regulation represent the mode of gene expression of some of the first coded proteins. Such relatively multifunctional enzymes might represent a mechanism by which primitive cells with small RNA genomes could overcome their limited coding abilities. This is consistent with the notion of a less accurate, error prone primordial primitive translational apparatus synthesizing small "statistical" enzymes [16].

A chimerical nature for the origin of Eucaryots

It is interesting to see from Figure 1 that while the intersections of Bacteria and Eucarya (B-E), and Archaea and Eucarya (A-E) share RNA-binding domains, the intersection of Bacteria and Archaea is empty. Its has been suggested

that Eucaryots evolved via symbiogenesis by an Archaea (*Thermoplasma*-like) and a Bacteria (*Spirochaeta*-like) [17]. The pattern of conservation at the fundamental level of RNA-binding proteins domains is what we may expect if Eucaryots had a chimeric origin between an Archaea and a Bacteria. It would be interesting to see if this pattern extends to other(s) functional class(es) of proteins.

An analysis of the phylogenetic distribution using BLAST [18] and ORFandDB (Falta referencia ORFandDB) [19] of RNA-binding domains together with domains definitions can be downloaded from [http://bacteria.fciencias.unam.mx/RNAbinding/].

Acknowledgments
This paper was completed during an internship of one of us (Luis Delaye) in the Protein Design Group, CNB. Support offered is kindly appreciated.

References

1. Gilbert W: **The RNA world.** *Nature* 1986, **319**:618.

2. Freeland SJ, Knight RD, Landweber LF, **Do proteins predate DNA?** *Science* 1999, **286**:690-692 .

3. Joyce GF: **The antiquity of RNA-based evolution.** *Nature* 2002, **418**:214-221.

4. Delaye L, Lazcano A: **RNA-binding peptides as molecular fossils.** In *Astrobiology: Origins from the Big-Bang to Civilisation.* Edited by J. Chela-Flores, G. Lemerchand, and J. Oró. Dordrecht: Kluwer Academic Publishers; 2000: 285-288.

5. Anantharaman V, Koonin EV, Aravind L: **Comparative genomics and evolution of proteins involved in RNA metabolism.** *Nucleic Acid Res* 2002, **30**:1427-1464.

6. Jacob F: **Evolution and tinkering.** *Science* 1977, 196:1161-1166..

7. Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L: **Gene families: the taxonomy of protein paralogs and chimeras.** *Science* 1997, **278**:609-614.

8. Jensen RA: **Enzyme recruitment in evolution of new function.** *Annu Rev Microbiol* 1976, **30**:409-425.

9. Draper DE, Reynaldo LP: **RNA binding strategies of ribosomal proteins.** *Nucleic Acids Res* 1999, **27**:381-388.

10. Draper D: **RNA-protein interactions in ribosomes**. In *RNA-Protein Interactions.* Edited by Kiyoshi Nagai and Iain E. Mattaj. USA: IRL Press; 1994

11. *Stams T, Niranjanakumari S, Fierke CA, Christianson DW:* **Ribonuclease P protein structure: evolutionary origins in the translational apparatus.** *Science* 1998, **280**:752-755.

12. Kyrpides NC, Woese CR, Ouzounis CA: **KOW: a novel motif linking a bacterial transcription factor with ribosomal proteins.** *Trends Biochem Sci* 1996, **21**:425-426.

13. Lecompte O, Ripp R, Thierry JC, Moras D, Poch O: **Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale.** *Nucleic Acids Res* 2002, **30**:5382-5390.

14. Noller HF, Nomura M: *Escherichia coli and Salmonella cellular and molecular biology.* Second Edition, Frederick C. Neidhardt, Editor in Chief. ASM Press Washington, DC; 1996.

15. O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R:
High-quality protein knowledge resource: SWISS-PROT and TrEMBL.
Brief Bioinform 2002, **3**:275-284.

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

17. Margulis L, Doland MF, Guerrero R: **The chimeric Eukaryote: origin of the nucleus from the karyomastigont in amitochondriate protist.** In *Variation and Evolution in Plants and Microorganisms.* Edited by Fransisco J. Ayala, Walter M. Fitch and Michael T. Clegg. USA: National Academy Press; 2000:21-34.

18. Altschul SF, Madden TL, Schaffer AA, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acid Res* 1997, **25**:3389-3402.

19. Falta referencia de ORFandDB

20. Markus MA, Hinck AP, Huang S, Draper DE, Torchia DA: **High resolution solution structure of ribosomal protein L11-C76, a helical protein with a flexible loop that becomes structured upon binding to RNA.** *Nat Struct Biol* 1997, **4**:70-77.

21. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA: **The complete atomic structure of the large ribosomal subunit at 2.4 A resolution.** *Science* 2000, **289**:905-920.

22. Davies C, White SW, Ramakrishnan V: **The crystal structure of ribosomal protein L14 reveals an important organizational component of the translational apparatus.** *Structure* 1996, 4:55-66.

23. Unge J, Berg A, Al-Kharadaghi S, Nikulin A, Nikonov S, Davydova N, Nevskaya N, Garber M, Liljas A: **The crystal structure of ribosomal protein L22 from Thermus thermophilus: insights into the mechanism of erythromycin resistance.** *Structure* 1998, 6:1577-1586.

24. Stoldt M, Wohnert J, Ohlenschlager O, Gorlach M, Brown LR: **The NMR structure of the 5S rRNA E-domain-protein L25 complex shows preformed and induced recognition.** *Embo J* 1999, **18**:6508-6521.

25. Mao H, White SA, Williamson JR: ***A novel loop-loop recognition motif in the yeast ribosomal protein L30* autoregulatory RNA complex.** *Nat Struct Biol* 1999, 6:1139-1147.

26. Nikonov SV, Nevskaya NA, Fedorov RV, Khairullina AR, Tishchenko SV, Nikulin AD, Garber MB: **Structural studies of ribosomal proteins.** *Biol Chem* 1998, **379**:795-805.

27. Hard T, Rak A, Allard P, Kloo L, Garber M: **The solution structure of ribosomal protein L36 from *Thermus thermophilus* reveals a zinc-ribbon-like fold.** *J Mol Biol* 2000, **296**:169-180.

28. Wahl MC, Bourenkov GP, Bartunik HD, Huber R: **Flexibility, conformational diversity, and two dimerization modes in complexes of ribosomal protein L12.** *Embo J* 2000, **19**:174-186.

29. Wimberly BT, Brodersen DE, Clemons WM Jr, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V: **Structure of the 30S ribosomal subunit.** *Nature*. 2000, **407**:327-339.

30. Tocilj A, Schluenzen F, Janell D, Gluhmann M, Hansen H, Harms J, Bashan A, Bartels H, Agmon I, Franceschi F, Yonath A: **The small ribosomal subunit from *Thermus thermophilus* at 4.5 angstrom resolution: pattern fittings and the identification of a functional site.** *Proc Nat Acad Sci USA* 1999, **96**:14252-14257.

31. Hosaka H, Nakagawa A, Tanaka I, Harada N, Sano K, Kimura M, Yao M, Wakatsuki S: **Ribosomal protein S7: a new RNA-binding motif with structural similarities to a DNA architectural factor.** *Structure* 1007, **5**:1199-1208.

32. Clemons Jr, WM, May JLC, Wimberly BT, Mccutcheon JP, Capel MS, Ramakrishnan V: ***Partial model for 30S* ribosomal subunit.** *Nature* 1999, **400**:833-840.

33. Arnez JG, Augustine JG, Moras D, Francklyn CS: **The first step of aminoacylation at the atomic level in histidyl-tRNA synthetase.** *Proc Natl Acad Sci U S A* 1997, **94**:7144-7149.

34. Cavarelli J, Delagoutte B, Eriani G, Ganglotf J, Moras D: **L-arginine recognition by yeast arginyl-tRNA *synthetase*.** *Embo J* 1998, **17**:5438-5448.

35. Schmitt E, Moulinier L, Fujiwara S, Imanaka T, Thierry JC, Moras D: **Crystal structure of aspartyl-tRNA synthetase from *Pyrococcus kodakaraensis* KOD: archaeon specificity and catalytic mechanism of adenylate formation.** *EMBO J* 1998, **17**:5227-5237.

36. Cahuzac B, Berthonneau E, Birlirakis N, Guittet E, Mirande M: **A recurrent RNA-binding domain is appended *to eukaryotic aminoacyl-tRNA synthetases*.** *Embo J* 2000, **19**:445-452.

37. al-Karadaghi S, Aevarsson A, Garber M, Zheltonosova J, Liljas A: **The structure of elongation factor G in complex with GDP: conformational flexibility and nucleotide exchange.** *Structure* 1996, 4:555-565.

38. Nissen P, Kjeldgaard M, Thirup S, Polekhina G, Reshetnikova L, Clark BF, Nyborg J: **Crystal structure of the ternary complex of Phe-tRNAPhe, EF-Tu, and a GTP analog.** *Science* 1995, **270**:1464-1472.

39. Kim KK, Hung LW, Yokota H, Kim R, Kim SH: ***Crystal structures of eukaryotic translation initiation factor* 5A from *Methanococcus jannaschii* at 1.8 A resolution.** *Proc Natl Acad Sci U S A.* 1998, **95**:10419-10424.

40. Birse DE, Kapp U, Strub K, Cusack S, Aberg A: **The crystal structure of the signal recognition particle Alu RNA binding heterodimer, SRP9/14.** *EMBO J* 1997, **16**:3757-3766.

41. Batey RT, Sagar MB, Doudna JA: **Structural and energetic analysis of RNA recognition by a universally *conserved protein from the signal recognition particle*.** *J MolBiol* 2001, **307**:229-246.

42. Schmitt E, Blanquet S, Mechulam Y: **Structure of crystalline *Escherichia coli* methionyl-tRNA(f)Met formyltransferase: comparison with glycinamide ribonucleotide formyltransferase.** *EMBO J* 1996. **15**:4749-4758.

43. *Graedler U, Gerber H-D, Goodenough-Lashua DM, Garcia GAG, Ficner R, Reuter K, Stubbs MT, Klebe G:* **A new target for shigellosis: rational design and crystallographic studies of inhibitors of tRNA-guanine transglycosylase.** *J Mol Biol* 2001, **306**:455-467.

44. Li H, Trotta CR, Abelson J: **Crystal structure and evolution of a transfer RNA splicing enzyme.** *Science* 1998, **280**:279-284.

45. *Manival X, Yang Y, Strub MP, Kochoyan M, Steinmetz M, Aymerich S:* ***From genetic to structural*** **characterization of a new class of RNA-binding domain within the SacY/BglG family of antiterminator proteins.** *EMBO J* 1997, **16**:5019-5029.

46. Schindelin H, Marahiel MA, Heinemann U: **Universal nucleic acid-binding domain revealed by crystal structure of the *B. subtilis* major cold-shock protein.** *Nature* 1993, **364**:164-168.

47. *Gopal B, Haire LF, Cox RA, Colston MJ, Major S, Brannigan JA, Smerdon SJ, Dodson GG:* **The *crystal structure* of Nusb from *Mycobacterium tuberculosis*.** *Nat Struct Biol* 2000, 7:475-478.

48. Chen Xp, Antson AA, Yang M, Li P, Baumann C, Dodson EJ, Dodson GG, Gollnick P: **Regulatory features of the *trp* operon and the crystal structure of the trp RNA-binding attenuation protein from *Bacillus stearothermophilus*.** *J Mol Biol* 1999. **289**:1003-1016.

49. *Teplova M, Tereshko V, Sanishvili R, Joachimiak A, Bushueva T, Anderson WF, Egli M:* **The *structure of the yrdC* gene product from *E. coli* reveals a new fold and suggests a role in RNA-binding.** *Protein Sci* 2000, 9:2557-2566.

50. Avis JM, Allain FH, Howe PW, Varani G, Nagai K, Neuhaus D: **Solution structure of the N-terminal RNP domain of U1A protein: the role of C-terminal residues in structure stability and RNA binding.** *J Mol Biol* 1996, **257**:398-411.

51. Lewis HA, Chen H, Edo C, Buckanovich RJ, Yang YYL, Musunuru K, Zhong R, Darnell RB, Burley SK: **Crystal Structures of Nova-1 and Nova-2 K-Homology RNA-Binding Domains.** *Structure (London)* 1999, 7:191-203.

52. Wang H, Boisvert D, Kim KK, Kim R, Kim S-H: **Crystal structure of a fibrillarin homologue from *Methanococcus jannaschii*, a hyperthermophile, at 1.6 A resolution.** *Embo J* 2000, **19**: 317-323.

53. Katayanagi K, Okumura M, Morikawa K: **Crystal structure of *Escherichia coli* RNase HI in complex with Mg2+ at 2.8 A resolution: proof for a single Mg(2+)-binding site.** *Proteins* 1993, 17:337-346.

54. Evans SP, Bycroft M: **NMR structure of the N-terminal domain of *Saccharomyces cerevisiae* RNase Hi reveals a fold with a strong resemblance to the N-terminal domain of ribosomal protein L9.** *J Mol Biol* 1999, **291**:661-669.

55. Banner DW, Kokkinidis M, Tsernoglou D: **Structure of the ColE1 rop protein at 1.7 A resolution.** *J Mol Biol* 1987, **196**:657-675.

56. Sevcik J, Urbanikova L, Dauter Z, Wilson KS: **Recognition of RNase Sa by the inhibitor barstar: structure of the complex at 1.7 A resolution.** *Acta Crystallogr D Biol Crystallogr* 1998, **54**:954-963.

57. Bell JA: **X-Ray crystal structures of a severely desiccated protein.** *Protein Sci* 1999, **8**:2033-2040.

58. Bycroft M, Grunert S, Murzin AG, Proctor M, St Johnston D: **NMR solution structure of a dsRNA binding domain from *Drosophila* staufen protein reveals homology to the N-terminal domain of ribosomal protein S5.** *EMBO J* 1995. **14**:3563-3571.

Table 1. Structural classification of 68 RNA-binding proteins domains. Domains are classified according to cellular function. Domain definition is according to SCOP database. Secondary structure composition is given according to the class level in SCOP classification.. Gray colored cells in "Phylogenetic distribution" column with LCA represents domains widely distributed in Archaea, Bacteria and Eucarya, thus inferred to be present in the Last Common Ancestor.
*SWISSPROT source: http://ca.expasy.org/sprot/ [15].

Figure 1. Overview of the phylogenetic distribution of 68 RNA-binding domains Protein domains widely conserved in cellular proteomes from Bacteria, Archaea and Eucarya are assumed to be present in the LCA. HGT (LCA) are domains that are suspected to be present in the three lineages because of horizontal gene transfer. *Eucaryal ribosomal proteins shared with Bacteria and not with Archaea are of organelle origin

Figure 2. Evidence of evolution of domain fusion in some of the universally conserved RNA-binding proteins Aspartyl tRNA synthetase share an homologous RNA-binding domain with ribosomal proteins S12 and S17. EF-G share RNA-binding domains with ribosomal protein S5 and L3, and also has an internal duplication.

Figure 3. Homologous proteins in the assembly map of the 30S ribosomal subunit from *E. coli*. Arrows between proteins indicate the facilitating effect of one protein on the binding of another A thick arrow indicates a major facilitating effect. The thick arrows from 16S RNA to S4, S8, S15, S17, and S20 indicates that each of these proteins binds directly to rRNA in the absence of other proteins. Ribosomal proteins in gray are conserved in Bacteria, Archaea and Eucarya. Proteins S12 and S17 belong to the SCOP family of Cold shock DNA-binding domain like, and proteins S5 and S9 to the SCOP family Translation machinery components. Modified from Noller and Nomura [14].

Large subunit ribosomal protein domains

RNA-binding domain	Domain definition	Secondary structure composition**	Information/function of the protein	Phylogenetic distribution	Crystal structure reference
Ribosomal protein L11, C-terminal domain (L11p family)	Ribosomal protein L11, C-terminal domain (SCOP family)	All alpha proteins	Recognizes and binds tightly to a highly conserved 58 nucleotide domain of 23 S ribosomal RNA	LCA	[20]
Ribosomal protein L13 (L13p family)	Ribosomal protein L13 (SCOP family)	Alpha and beta proteins (a/b)	Interacts substantially with domain VI of 23S rRNA and participates in the protein cluster composed of L3, L6, L13, L14 and L24e that is found close to the factor binding site	LCA	[21]
Ribosomal protein L14 (L14p family)	Ribosomal protein L14 (SCOP family)	All beta proteins	Occupies a central location between the peptidyl transferase and GTPase regions of the large ribosomal subunit	LCA	[22]
Ribosomal protein L19 (L19e family)	Ribosomal protein L19 (L19e) (SCOP family)	All alpha proteins	Interacts substantially with domain II, III and IV of 23S rRNA	Archaea and Eucarya	[21]
Ribosomal protein L21e (L21e family, SH3-like β-barrel)	Ribosomal proteins L24p and L21e (SCOP family)	All beta proteins	Attaches to helix 1 and helix 2/3 of 5S rRNA and domains II and V of 23S rRNA	Archaea and Eucarya	[21]
Ribosomal protein L22 (L22p family)	Ribosomal protein L22 (SCOP family)	Alpha and beta proteins (a+b)	Is one of five proteins necessary for the formation of an early folding intermediate of the 23S rRNA and interacts with all six domains of 23S rRNA	LCA	[23]
Ribosomal protein L23 (L23P family)	L23p (SCOP family)	Alpha and beta proteins (a+b)	Interacts substantially with domain III of 23S rRNA	LCA	[21]
Ribosomal protein L24e (L24e family)	Ribosomal protein L24e (SCOP family)	Small proteins	Interacts substantially with domain VI of 23S rRNA and participates in the protein cluster composed of L3, L6, L13, L14 and L24e that is found close to the factor binding site	Archaea and Eucarya	[21]
Ribosomal protein L25 (L25p family)	Ribosomal protein L25-like (SCOP family)	All beta proteins	Binds to the 5S rRNA*	Bacteria (HGT to Eucarya?)	[24]
Ribosomal protein L29 (L29p family)	Ribosomal protein L29 (L29p) (SCOP family)	All alpha proteins	Interacts substantially with domain I of 23S rRNA	LCA	[21]
Eukaryotic ribosomal protein L30 (L30e family pefota domain)	L30e/L7ae ribosomal proteins (SCOP family)	Alpha and beta proteins (a+b)	Negatively autoregulates its production by binding to a helix-loop-helix structure formed in its pre-mRNA and its mRNA	Archaea and Eucarya (HGT to Bacteria?)	[25]
Prokaryotic ribosomal protein L30 (L30p family)	Ribosomal protein L30p/L7e (SCOP family)	Alpha and beta proteins (a+b)	In *Haloarcula marismortui* interacts substantially with domains II of 23S rRNA and to the 5S rRNA	LCA	[26]
Ribosomal protein L31e (L31e family)	Ribosomal protein L31e (SCOP family)	Alpha and beta proteins (a+b)	Interacts substantially with domains III, IV and VI of 23S rRNA	Archaea and Eucarya	[21]
Ribosomal protein L36 (L36p family)	Ribosomal protein L36 (SCOP family)	Small proteins	Contains a zinc-ribbon like fold	Bacteria and Eucaryotic organelles	[27]
Ribosomal protein L37e (L37e family)	Ribosomal protein L37e (SCOP family)	Small proteins	Interacts substantially with domains I, II and III of 23S rRNA	Archaea and Eucarya	[21]
Ribosomal protein L39e (L39e family)	Ribosomal protein L39e (SCOP family)	All alpha proteins	Interacts substantially with domains I and III of 23S rRNA	Archaea and Eucarya	[21]
Ribosomal protein L3 (L3p family)	Ribosomal protein L3 (SCOP family)	All beta proteins	Interacts substantially with domains IV, V and VI of 23S rRNA and participates in the protein cluster composed of L3, L6, L13, L14 and L24e that is found close to the factor binding site	LCA	[21]
Ribosomal protein L44e (L44e family)	Ribosomal protein L44e (SCOP family)	Small proteins	Interacts substantially with domains I and V of 23S rRNA	Archaea and Eucarya	[21]
Ribosomal protein L5 C-terminal domain (L5p family)	Ribosomal protein L5 C-terminal domain (Pfam family Ribosomal_L5_C)	Alpha and beta proteins (a+b)	Interacts substantially with domain V of 23S rRNA	LCA	[21]
Ribosomal protein L5 N-terminal domain (L5p family)	Ribosomal protein L5 N-terminal domain (Pfam family Ribosomal_L5)	Alpha and beta proteins (a+b)*	Interacts substantially with domain V of 23S rRNA	LCA	[21]
Ribosomal protein L7/12, C-terminal domain (L12p family)	Ribosomal protein L7/12, C-terminal domain (SCOP family)	Alpha and beta proteins (a+b)	Presumed to be involved in the binding of translation factors, stimulating factor-dependent GTP hydrolysis	Bacteria and Eucaryotic organelles	[28]

Small subunit ribosomal protein domains

RNA-binding domain	Domain definition	Secondary structure composition**	Information/function of the protein	Phylogenetic distribution	Crystal structure reference
Ribosomal protein S10 (S10p family)	Ribosomal protein S10 (SCOP family)	Alpha and beta proteins (a+b)	Involved in the binding of tRNA to the ribosomes*	LCA	[29]
Ribosomal protein S11 (S11p (S14e) family)	Ribosomal protein L18 and S11 (SCOP family)	Alpha and beta proteins (a/b)	Packs flat against the minor groove of rRNA	LCA	[29]
Ribosomal protein S12 (S12p family)	Cold shock DNA-binding domain-like (SCOP family)	All beta proteins	Is involved in the translation initiation step*	LCA	[29]
Ribosomal protein S14 (S14p family)	Ribosomal protein S14 (SCOP family)	Small proteins	Known to be required for the assembly of the 30S particles and may also be responsible for determining the conformation of the 16S rRNA at the A site*	LCA	[29]
Ribosomal protein S15 (S15p family)	Ribosomal protein S15 (SCOP family)	All alpha proteins	Is one of the 16S ribosomal RNA binding proteins*	LCA	[29]
Ribosomal protein S16 (S16p family)	Ribosomal protein S16 (SCOP family)	Alpha and beta proteins (a+b)	Belongs to the S16P family of ribosomal proteins*	Bacteria and Eucaryotic organelles	[29]
Ribosomal protein S18 (S18p family)	Ribosomal protein S18 (SCOP family)	All alpha proteins	This protein has been implicated in aminoacyl-transfer RNA binding. It appears to be situated at the decoding site of messenger RNA*	Bacteria and Eucaryotic organelles	[29]
Ribosomal protein S19 (S19p family)	Ribosomal protein S19 (SCOP family)	Alpha and beta proteins (a+b)	Forms a complex with S13 that binds strongly to the 16S rRNA*	LCA	[29]
Ribosomal protein S2 (S2p family)	Ribosomal protein S2 (SCOP family)	Alpha and beta proteins (a/b)	Contains a long helical α-hairpin extension	LCA	[29]
Ribosomal protein S5 C-terminal domain (S5p family)	Translational machinery components (SCOP family)	Alpha and beta proteins (a+b)	Is important in the assembly and function of the 30S ribosomal subunit*	LCA	[30]
Ribosomal protein S5 N-terminal domain (S5p family)	Ribosomal S5 protein, N-terminal domain (SCOP family)	Alpha and beta proteins (a+b)	Is important in the assembly and function of the 30S ribosomal subunit*	LCA	[30]
Ribosomal protein S6 (S6p family)	Ribosomal protein S6 (SCOP family)	Alpha and beta proteins (a+b)	Binds together with S18 to 16S rRNA*	Bacteria and Eucaryotic organelles	[29]
Ribosomal protein S7 (S7p family)	Ribosomal protein S7 (SCOP family)	All alpha proteins	Is one of the primary 16S rRNA-binding proteins responsible for initiating the assembly of the head of the 30S subunit and has been shown to be the major protein component to cross-link with tRNA molecules bound at both the (A) and (P) sites of the ribosome	LCA	[31]
Ribosomal protein S9 (S9P family)	Translational machinery components (SCOP family)	Alpha and beta proteins (a+b)	Belongs to the S9P family of ribosomal proteins*	LCA	[29]
Ribosomal protein S17 (S17p family)	Cold shock DNA-binding domain-like (SCOP family)	All beta proteins	Protein S17 binds specifically to the 5' end of 16S rRNA*	LCA	[32]

Aminoacyl-tRNA synthetases related domains

RNA-binding domain	Domain definition	Secondary structure composition**	Information/function of the protein	Phylogenetic distribution	Crystal structure reference
Histidyl-tRNA synthetase (HisRS), C-terminal domain (α/β) ACB HXGXTP)	Anticodon-binding domain of Class II aaRS (SCOP family)	Alpha and beta proteins (a/b)	tRNA recognition and aminoacylation	LCA	[33]
Arginyl-tRNA synthetase (ArgRS), N-terminal 'additional' domain N-Arg	Arginyl-tRNA synthetase (ArgRS), N-terminal 'additional' domain (SCOP family)	Alpha and beta proteins (a+b)	tRNA recognition and aminoacylation	LCA	[34]
Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases (DALR)	Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases (SCOP family)	All alpha proteins	tRNA recognition and aminoacylation	LCA	[34]
Aspartyl-tRNA synthetase (AspRS) Anticodon-binding domain (NOB)	Anticodon-binding domain (SCOP fold: OB fold)	All beta proteins	tRNA recognition and aminoacylation	LCA	[35]
Multifunctional Glu-Pro-tRNA synthase (EPRS) second repeated element (Pfam: WHEP-TRS)	a tRNA synthase domain (SCOP family)	All alpha proteins	This repeated motif may represent a novel type of general RNA-binding domain appended to Eucaryotic aminoacyl-tRNA synthetases to serve as a cis-acting tRNA-binding cofactor	Eucarya	[36]

Elongation and initiation factors related domains

RNA-binding domain	Domain definition	Secondary structure composition**	Information/function of the protein	Phylogenetic distribution	Crystal structure reference
Elongation factor G (EF-G), domain II	Elongation factor Tu domain 2 (Pfam family GTP_EFTU_D2)	All beta proteins	Catalyzes the translocation step of translation	LCA	[37]
Elongation factor G (EF-G), domain IV	Translational machinery components (SCOP family)	Alpha and beta proteins (a+b)	Catalyzes the translocation step of translation	LCA	[37]
Elongation factor G (EF-G), domain V	Elongation factor G (EF-G), domains III and V (SCOP family)	Alpha and beta proteins (a+b)	Catalyzes the translocation step of translation	LCA	[37]
Elongation factor Tu (EF-Tu) C-terminal domain or domain III	EF-Tu/eEF-1alpha C-terminal domain (SCOP family)	All beta proteins	Placing the amino acids in their correct order when messenger RNA is translated into a protein sequence on the ribosome	LCA	[38]
C-terminal domain of eukaryotic initiation translation factor 5a	Cold shock DNA-binding domain-like (SCOP family)	All beta proteins	Initiation of protein biosynthesis	Archaea and Eucarya	[39]
N-terminal domain of eukaryotic initiation translation factor 5a (Pfam KOW domain)	N-terminal domain of eukaryotic initiation translation factor 5a (SCOP family)	All beta proteins	Initiation of protein biosynthesis	LCA	[39]

Secretion protein related domains

RNA-binding domain	Domain definition	Secondary structure composition**	Information/function of the protein	Phylogenetic distribution	Crystal structure reference
Signal recognition particle 14 kDa protein (srp9/14)	Signal recognition particle alu RNA binding heterodimer, SRP9/14 (SCOP family)	Alpha and beta proteins (a+b)	Recognizes the signal sequence of the nascent polypeptide chain emerging from the ribosome, and targets the ribosome-nascent chain-SRP complex to the rough endoplasmic reticulum	Eucarya	[40]
Signal sequence binding protein Ffh	Signal peptide-binding domain (SCOP family)	All beta proteins	Responsible for targeting proteins to the inner membrane in Prokarya	LCA	[41]

tRNA modification related domains

RNA-binding domain	Domain definition	Secondary structure composition**	Information/function of the protein	Phylogenetic distribution	Crystal structure reference
Methionyl-tRNAfmet formyltransferase, C-terminal domain (Met-tRNA-f)	Methionyl-tRNAfmet formyltransferase, C-terminal domain (SCOP family)	All beta proteins	Formylation of the methionyl moiety esterified to the 3' end of tRNA(f)Met	Bacteria and Eucarya	[42]
tRNA-guanine transglycosylase (TGT)	tRNA-guanine transglycosylase (SCOP family)	Alpha and beta proteins (a/b)	Involved in the hyper-modification of cognate tRNAs leading to the exchange of G34 at the wobble position in the anticodon loop by preQ1	LCA	[43]
RNase P protein	RNase P protein (SCOP family)	Alpha and beta proteins (a+b)	5' maturation of pre-tRNA substrates	Bacteria	[11]
tRNA splicing endonuclease, C-terminal domain (tRNA-int endo)	tRNA splicing endonuclease, C-terminal domain (SCOP family)	Alpha and beta proteins (a/b)	Cleaves pre-tRNA at the 5' and 3' splice sites to release the intron	Archaea and Eucarya	[44]
tRNA splicing endonuclease EndA, N-terminal domain (tRNA-int endo)	tRNA splicing endonuclease EndA, N-terminal domain (SCOP family)	Alpha and beta proteins (a+b)	Cleaves pre-tRNA at the 5' and 3' splice sites to release the intron	Archaea	[44]

Regulation of transcription related domains

RNA-binding domain	Domain definition	Secondary structure composition**	Information/function of the protein	Phylogenetic distribution	Crystal structure reference
SacY RNA-binding domain	BglG-like antiterminator proteins (SCOP family)	All beta proteins	Antiterminator protein	Bacteria	[45]
Cold shock protein cspB	Cold shock DNA-binding domain-like (SCOP family)	All beta proteins	Involved in cold-shock response (transcriptional activator)	Bacteria and Eucarya (HGT to Archaea?)	[46]
Antitermination factor NusB	Antitermination factor NusB (SCOP family)	All alpha proteins	Mediates the process of transcriptional antitermination	Bacteria (HGT to Archaea and Eucarya?)	[47]
Trp RNA-binding attenuation protein (TRAP) (Pfam: TrpBP)	Trp RNA-binding attenuation protein (TRAP) (SCOP family)	All beta proteins	Required for transcription attenuation control in the trp operon	Bacteria	[48]
Hypothetical protein YrdC (Sua5)	Hypothetical protein YrdC (SCOP family)	Alpha and beta proteins (a+b)	The yrdC family of genes codes for proteins that occur both independently and as a domain in proteins that have been implicated in regulation	LCA	[49]

Splicing and RNA processing other than tRNA related domains

RNA-binding domain	Domain definition	Secondary structure composition**	Information/function of the protein	Phylogenetic distribution	Crystal structure reference
Spliceosomal U1A protein RRM	Canonical RBD (SCOP family)	Alpha and beta proteins (a+b)	Binds stem loop II of U1 snRNA and it is the first sn-RNP to interact with pre-mRNA	Bacteria and Eucarya	[50]
Neuro-oncological ventral antigen 2 nova-2 KH3 (KH domain)	Eukaryotic type KH-domain (eKH-domain) (SCOP family)	Alpha and beta proteins (a+b)	May regulate RNA splicing or metabolism in a specific subset of developing neurons	LCA	[51]
Fibrillarin homologue	Fibrillarin homologue (SCOP family)	Alpha and beta proteins (a/b)	Processing of pre-rRNA	Archaea and Eucarya	[52]

DNA replication related domains

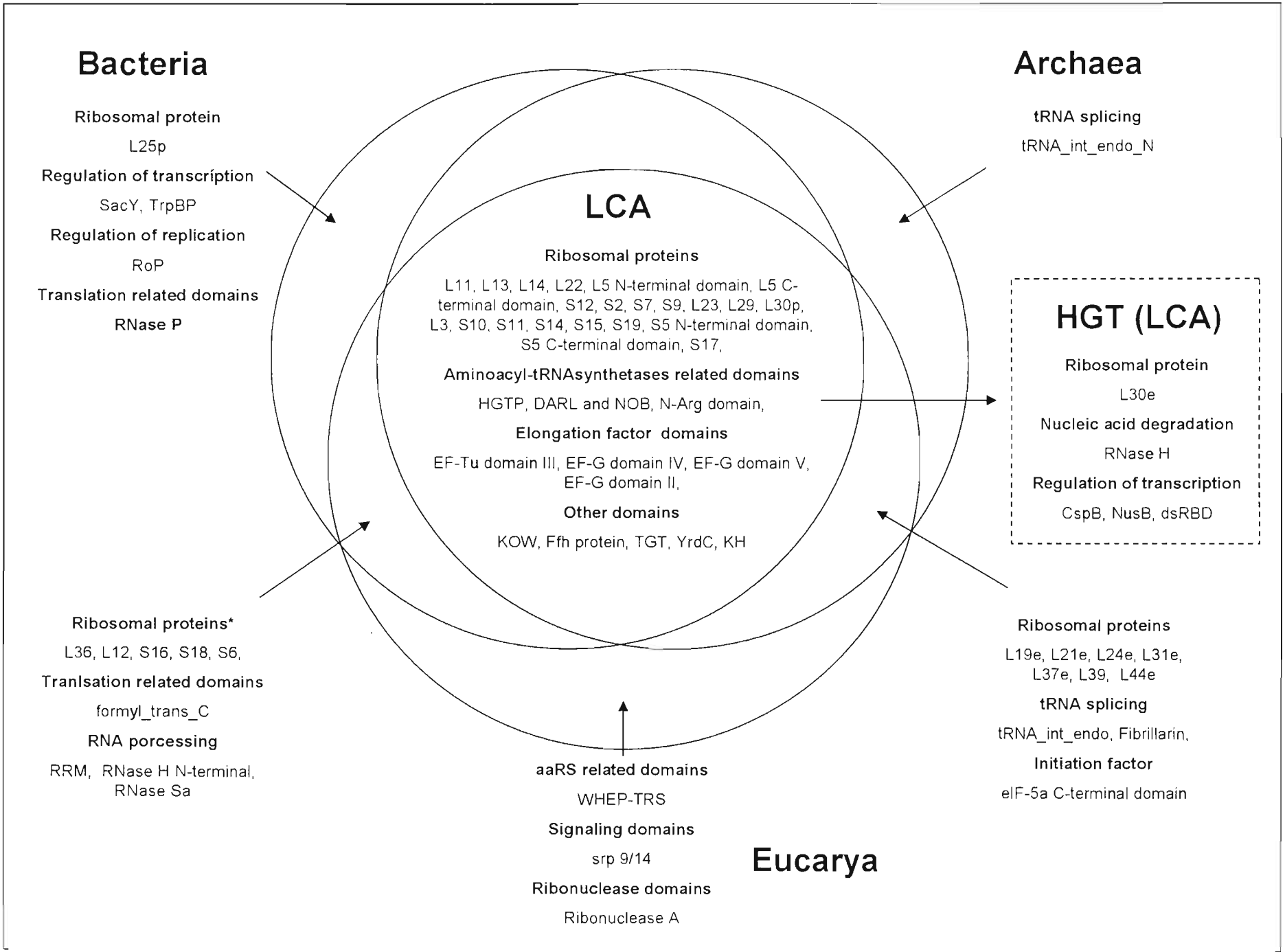
RNA-binding domain	Domain definition	Secondary structure composition**	Information/function of the protein	Phylogenetic distribution	Crystal structure reference
RNase H (RNase HI)	RnaseH (Pfam family)	Alpha and beta proteins (a/b)	Degrades the RNA of RNA-DNA hybrids specifically during DNA biosynthesis, it helps to specify the origin of genomic replication by suppressing initiation at origins other than the locus one	Bacteria and Eucarya (HGT to Archaea?)	[53]
N-terminal domain of RNase HI	N-terminal domain of RNase HI (SCOP family)	Alpha and beta proteins (a+b)	Endonuclease that degrades the RNA of RNA-DNA hybrids specifically	Bacteria and Eucarya	[54]
ROP protein	ROP protein (SCOP family)	All alpha proteins	Acts in the control of plasmid replication via regulation of an RNA-RNA interaction	Bacteria	[55]

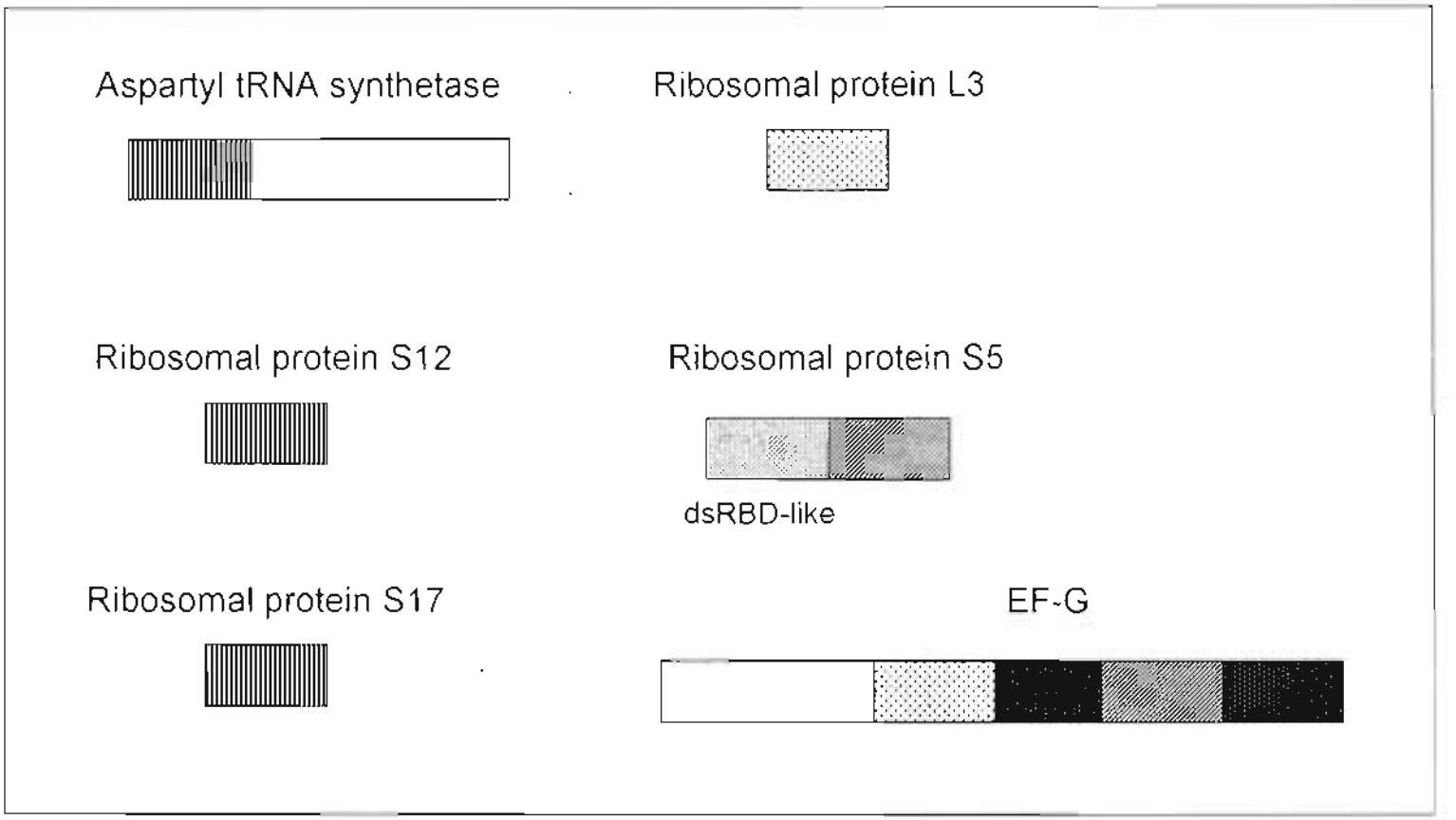
Other RNAses domains

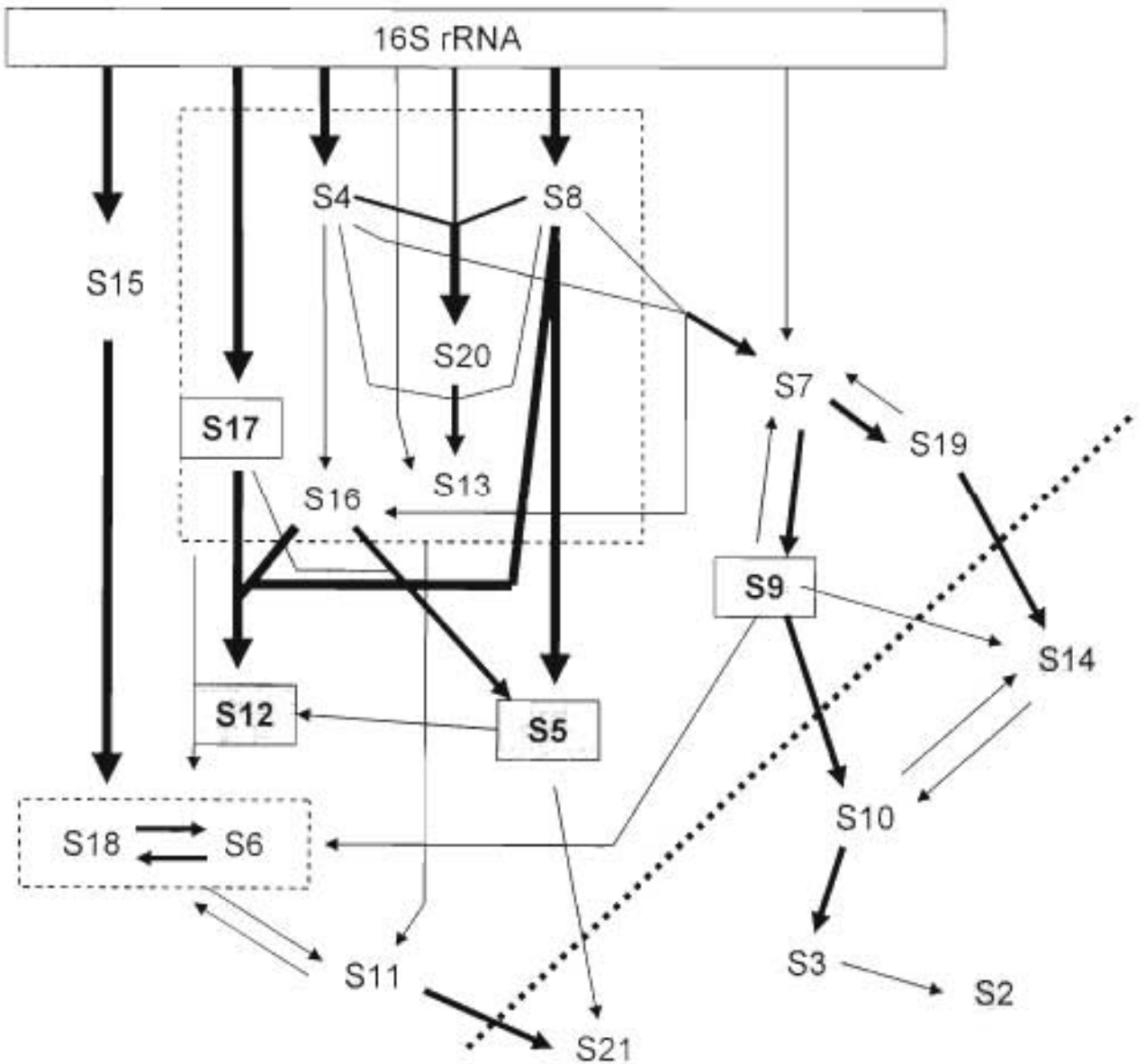
RNA-binding domain	Domain definition	Secondary structure composition**	Information/function of the protein	Phylogenetic distribution	Crystal structure reference
RNase Sa (Barnase)	Microbial ribonucleases (SCOP family)	Alpha and beta proteins (a+b)	Extracellular microbial ribonuclease	Bacteria and Eucarya	[56]
Ribonuclease A	Ribonuclease A-like (SCOP family)	Alpha and beta proteins (a+b)	Endonucleolytic cleavage to nucleoside 3'-phosphates and 3'-phosphooligonucleotides ending in C-P or U-P with 2',3'-cyclic phosphate intermediates	Eucarya (HGT to Bacteria?)	[57]

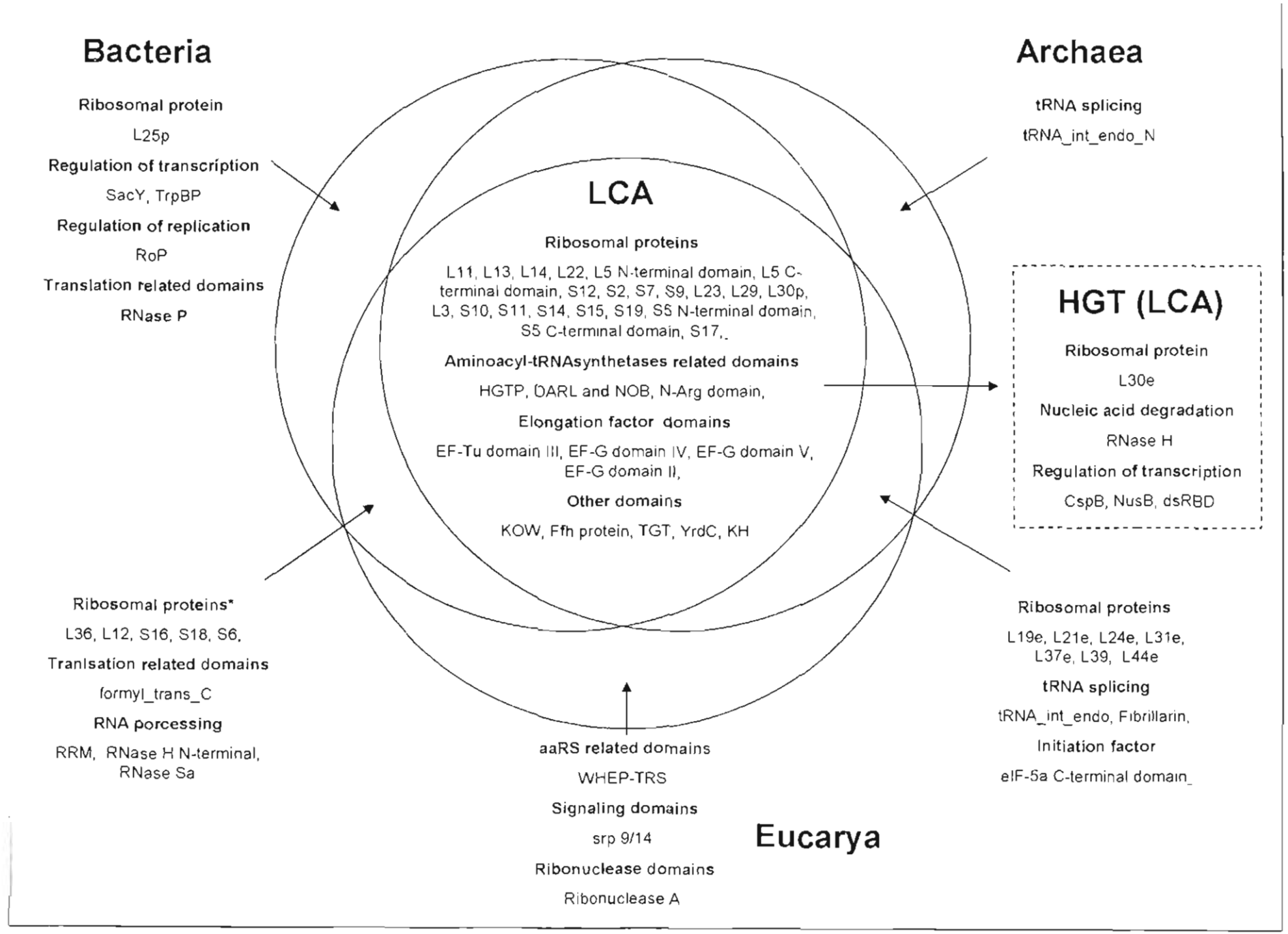
Double stranded RNA-binding domains

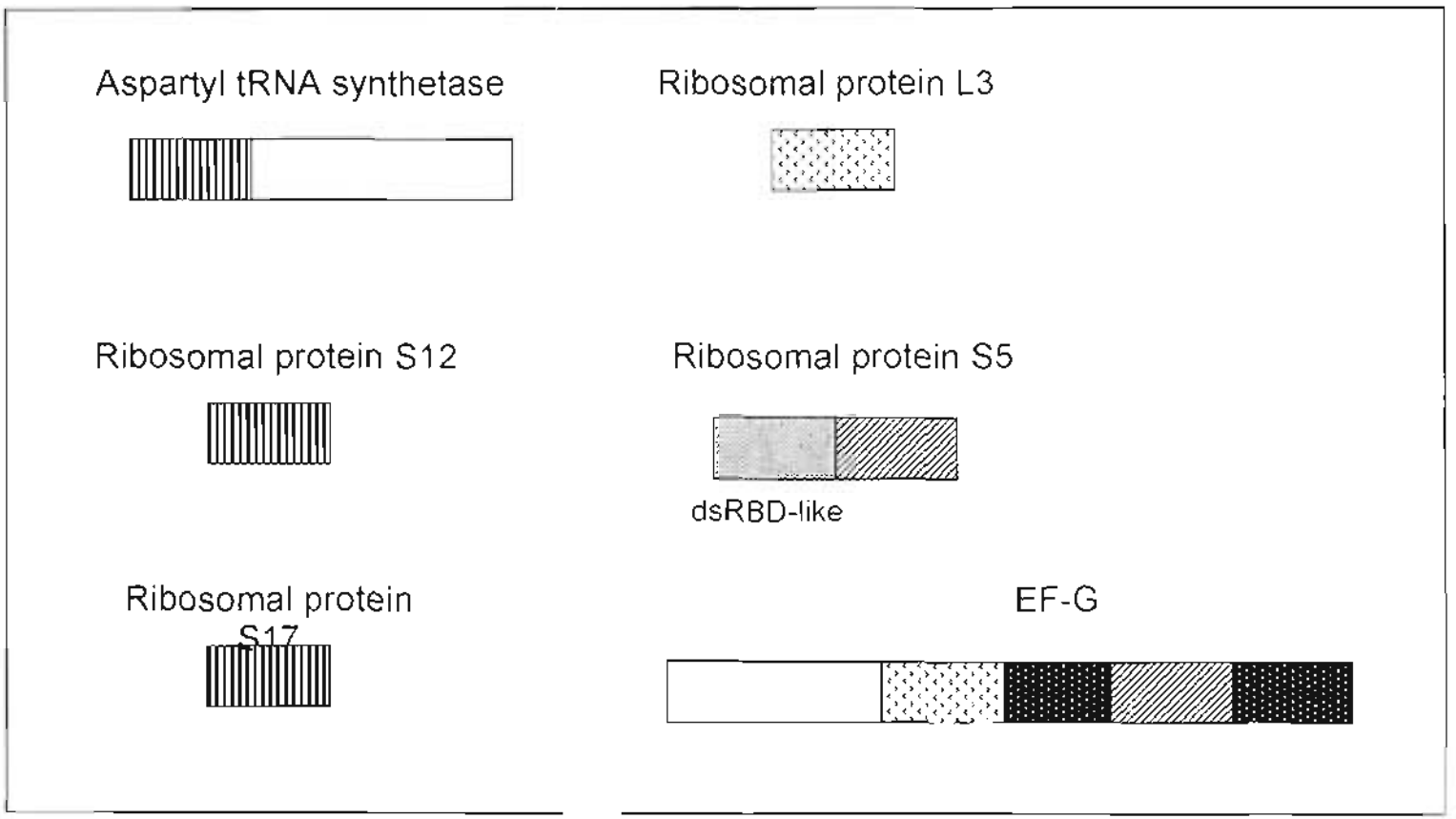
RNA-binding domain	Domain definition	Secondary structure composition**	Information/function of the protein	Phylogenetic distribution	Crystal structure reference
Double-stranded RNA-binding domain (dsRBD) of Staufen protein	Double-stranded RNA-binding domain (dsRBD) (SCOP family)	Alpha and beta proteins (a+b)	Associates with mRNA during oogenesis in *Drosphila*	Bacteria and Eucarya (HGT to Archaea?)	[58]

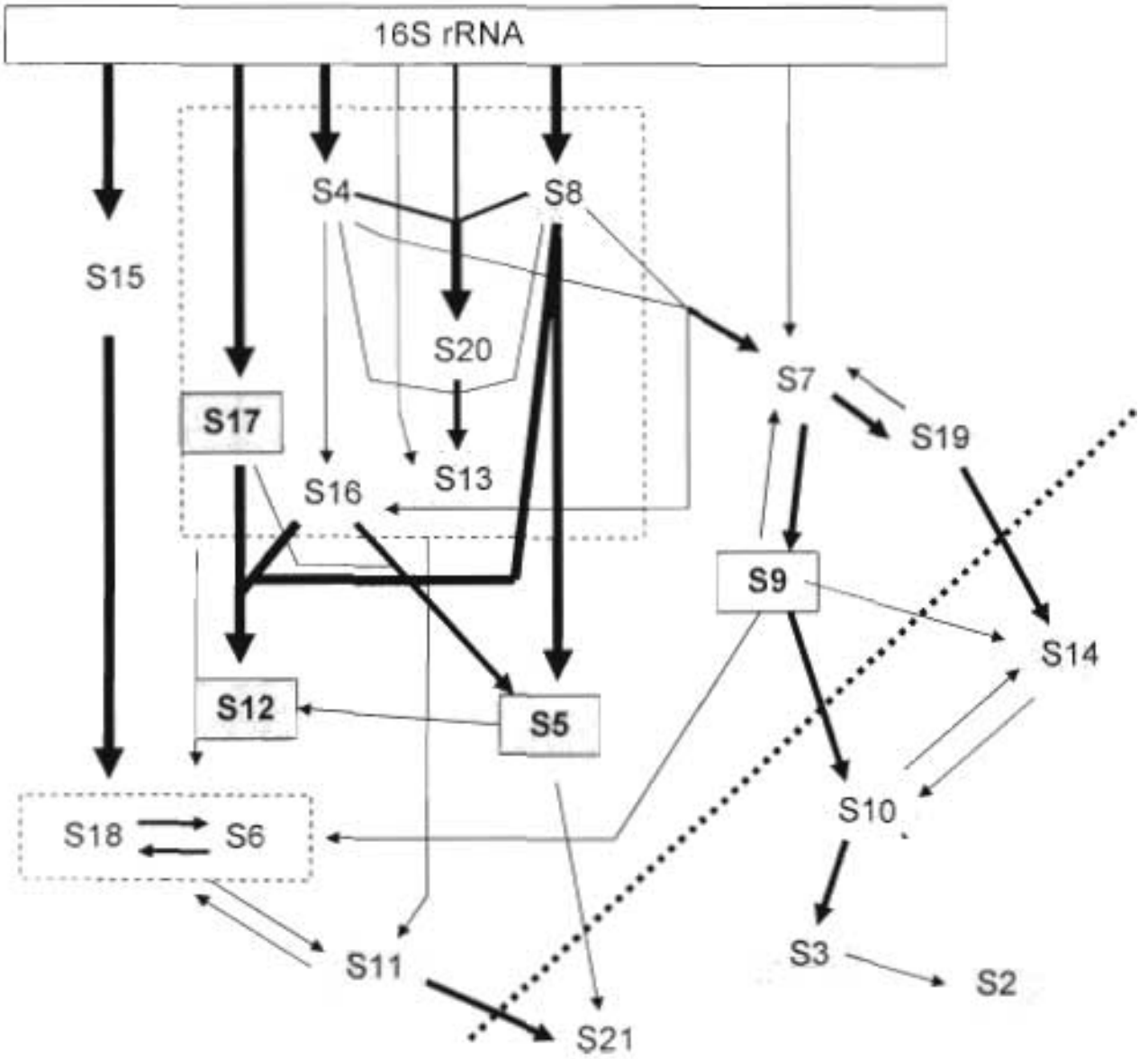

Bacteria
Ribosomal protein
L25p
Regulation of transcription
SacY, TrpBP
Regulation of replication
RoP
Translation related domains
RNase P
Archaea
tRNA splicing
tRNA_int_endo_N
LCA
Ribosomal proteins
L11, L13, L14, L22, L5 N-terminal domain, L5 C-terminal domain, S12, S2, S7, S9, L23, L29, L30p, L3, S10, S11, S14, S15, S19, S5 N-terminal domain, S5 C-terminal domain, S17,
Aminoacyl-tRNAsynthetases related domains
HGTP, DARL and NOB, N-Arg domain,
Elongation factor domains
EF-Tu domain III, EF-G domain IV, EF-G domain V, EF-G domain II,
Other domains
KOW, Ffh protein, TGT, YrdC, KH
HGT (LCA)
Ribosomal protein
L30e
Nucleic acid degradation
RNase H
Regulation of transcription
CspB, NusB, dsRBD
Ribosomal proteins*
L36, L12, S16, S18, S6,
Tranlsation related domains
formyl_trans_C
RNA porcessing
RRM, RNase H N-terminal, RNase Sa
aaRS related domains
WHEP-TRS
Signaling domains
srp 9/14
Ribonuclease domains
Ribonuclease A
Ribosomal proteins
L19e, L21e, L24e, L31e, L37e, L39, L44e
tRNA splicing
tRNA_int_endo, Fibrillarin,
Initiation factor
eIF-5a C-terminal domain
Eucarya


Aspartyl tRNA synthetase
Ribosomal protein L3
Ribosomal protein S12
Ribosomal protein S5
dsRBD-like
Ribosomal protein S17
EF-G


16S rRNA
S4
S8
S15
S20
S7
S19
S17
S16
S13
S9
S14
S12
S5
S10
S18
S6
S11
S3
S2
S21


Bacteria
Ribosomal protein
L25p
Regulation of transcription
SacY, TrpBP
Regulation of replication
RoP
Translation related domains
RNase P
Archaea
tRNA splicing
tRNA_int_endo_N
LCA
Ribosomal proteins
L11, L13, L14, L22, L5 N-terminal domain, L5 C-terminal domain, S12, S2, S7, S9, L23, L29, L30p, L3, S10, S11, S14, S15, S19, S5 N-terminal domain, S5 C-terminal domain, S17,.
Aminoacyl-tRNAsynthetases related domains
HGTP, DARL and NOB, N-Arg domain,
Elongation factor domains
EF-Tu domain III, EF-G domain IV, EF-G domain V, EF-G domain II,
Other domains
KOW, Ffh protein, TGT, YrdC, KH
HGT (LCA)
Ribosomal protein
L30e
Nucleic acid degradation
RNase H
Regulation of transcription
CspB, NusB, dsRBD
Ribosomal proteins*
L36, L12, S16, S18, S6,
Translsation related domains
formyl_trans_C
RNA porcessing
RRM, RNase H N-terminal, RNase Sa
aaRS related domains
WHEP-TRS
Signaling domains
srp 9/14
Ribonuclease domains
Ribonuclease A
Ribosomal proteins
L19e, L21e, L24e, L31e, L37e, L39, L44e
tRNA splicing
tRNA_int_endo, Fibrillarin,
Initiation factor
eIF-5a C-terminal domain
Eucarya


Aspartyl tRNA synthetase
Ribosomal protein L3
Ribosomal protein S12
Ribosomal protein S5
dsRBD-like
Ribosomal protein S17
EF-G


16S rRNA
S4
S8
S15
S20
S7
S19
S17
S16
S13
S9
S14
S12
S5
S10
S18
S6
S3
S11
S2
S21

On the early evolution of polymerase function

Luis Delaye, Arturo Becerra, and Antonio Lazcano

[1]Laboratorio de Microbiología
Departamento de Biología Evolutiva
Facultad de Ciencias. UNAM
México. D.F. MÉXICO

Abstract

Nucleic acid polymerization is one of the most ancient functions and is central to all forms of life on Earth. In extant cells, polymerization is performed by a plethora of enzymes from different phylogenetic origins. Although all present day organisms have DNA based genomes, there has been substantial discussion regarding the chemical nature of the genome of the Last Common Ancestor (LCA), (whether it had DNA or RNA), mainly because of the lack of conservation of the central replicative DNA polymerase in the tree cellular lineages (Archaea, Bacteria and Eucarya). Here we review the phylogenetic distribution, the complex domain structure and some aspects of the general pattern of the evolution of nucleic acid polymerases. We show that the evolution of nucleic acid polymerization function has been shaped in several cases by convergences and non-orthologous gene displacements. We suggest that this mode of evolution could explain that the lack of conservation of the central DNA replicative enzyme between present day organisms, a proposal that is in concordance with the most parsimonious scenario, a LCA with a DNA based genome. We also show that the catalytic domain, the so called "palm" domain, is associated to polymerases from different families having all the functions necessaries to understand the transition from an RNA-protein world to the present day DNA-RNA-protein world (these are RNA polymerases, Reverse transcriptases, and DNA polymerases), thus, it is likely that this domain can be traced back to the ancestral replicase of the RNA-protein stage of evolution.

Introduction.

Polymerization of genetic material is one of the central process to all life on Earth. Its origin is intimately coupled with the appearance of life and the beginning of Darwinian evolution. The evolution of genetic material itself, is somehow coupled to the evolution of the molecules implicated on nucleic acid metabolism, being polymerases one of the central enzymes.

In extant biological systems, nucleic acids are polymerized by a plethora of enzymes. This enzymes are in some cases clearly homologous, or related because they share homologous domains, mainly the catalytic ones. But in another cases, the polymerases are definitively unrelated, as long as enzymes having different folds have different evolutionary origins. The functions that nucleic acid polymerases perform in cells are also extremely varied, going from template dependent replication of genetic material during replication, DNA repair, transcription of different kinds of RNA, retrotranscription (by retroviruses), template independent polymerization of CCA onto the 3' terminus of immature tRNA, polyadenylation of pre-mRNA, generation of diversity in the vertebrate immune system, transfer of a nucleotide to a non-nucleic acid molecules involved in several metabolic pathways, or antibiotic resistance in bacteria, between others. Clearly, some of these functions are more ancient than others.

Due to the antiquity and relevance of the process of DNA replication, it was reasonable to expect that the enzymes responsible of such nucleic acid polymerization, had to be well conserved across the three main cellular lineages, Archaea, Bacteria and Eucarya, as is in fact the case of transcription and translation cell machinery. But quite surprisingly this happens not to be the case, cellular lineages differ in the kind of enzymes used to polymerize nucleic acids, in such degree that, excluding the main cellular RNA transcriptase β and β' there is no nucleic acid polymerase that is conserved across the broad phylogenetic spectrum of the three main cellular lineages (Archaea, Bacteria and Eucarya). As reviewed by Olsen and Woese (1997), Bacteria use one kind of polymerase known as DNA polymerase III, while Archaea and Eucarya use DNA polymerases from family II as their main replicative enzymes, (although it is possible that Euryarchaea uses another kind of DNA polymerase called polymerase D as its main replicative enzyme (Cann et al, 1998; Bohlke, et al 2002)). Several hypothesis has been proposed to explain this pattern. Edgell and Doolittle (1997) suggested that the LCA had a DNA genome and proposed three (not mutually exclusive) possibilities to explain the lack of conservation of the central replicative enzymes: (i) different replicating proteins are in fact homologous but we are unable to recognize the phylogenetic relationship due to extreme sequence divergence between extant enzymes, (ii) the last common ancestor (LCA) or *cenancestor* contained both systems (Bacterial and Archaea-Eucaryotic, perhaps one for reparation and the other for replication) and different components were lost after divergence, (iii) non homologous proteins were recruited into replication function in one or the other lineages, replacing cenancestral components.

An other hypothesis based on the idea of a higher rate of horizontal gene transfer during the early evolution of life and the observation of the different degrees of conservation of different components of the cell machinery (i.e., transcription and translation versus replication machinery), was proposed by Woese (1998) who has suggested that during primeval stages of the evolution of life, the notion of cellular ancestor had no physical meaning, because biological entities, at that time, behaved more like an evolving community, coupled by intense horizontal transfer, in such a way that we can not reconstruct the gene complement of a single cell or population of cells because they didn't exist as such. He suggested that the most conserved proteins in the cellular lineages are the ones that belong to systems that 'crystallized' first (their individual components began to be vertically inherited rather that horizontally transferred

more often). In this way, translation function was one of the firsts systems to 'crystallize' while replication function 'crystallized' later on, so generating the pattern that we see today. But, it is also likely that by the time of the *cenancestor* this model of *annealing* massive horizontal gene transfer (if it existed at all) wasn't as intense as suggested by Woese (1998), and the differences in the patterns of conservation between different parts of the cell machinery (replication versus transcription) are due to a plethora of different reasons.

A different proposal, based on an extensive sequence analysis of the enzymes involved in DNA replication (Leipe et al, 1999), concluded that the most parsimonious explanation was that the *cenancestor* had a retrovirus-like genetic system, and modern DNA replication was invented twice, once in the Bacterial lineage, and once in the Archaeal-Eucaryal lineage. In this cellular ancestor, DNA was retrotranscribed from RNA by a reverse transcriptase. According to Leipe et al, (1999), once DNA polymerases DNA dependents evolved, selection favored elimination of reversetranscriptase enzymes to prevent 'backward' propagation of damage to RNA onto DNA. The latter model may be important because at some stage in the evolution of life, cellular genetic systems based on RNA had to be somehow retrotranscribed onto DNA without loosing coded genetic information, although it is also likely that due to replicases with low discriminatory power between ribonucleotides and desoxyribonucleotides a cellular stage in which a reverse transcriptase played a central role in genome replication, may not be necessary. This means, that it is likely that some of the extant genes predate from an evolutionary stage where the sole informative biomolecules were RNA and protein and there was no DNA. Anyway, the lack of conservation of the enzymes participating in a particular cellular function (in this case, DNA polymerization) can still be explained by one of the proposals made by Edgell and Doolittle (1997), specially if we find the biological reasons for the extreme divergence of the proteins involved in the function, or the process that had replaced DNA replicating proteins so frequently and intensely, specifically in the divergence between Bacterial and Archaea-Eucaryal lineages.

Recently, Forterre (1999) has suggested a mechanism that could explain the lack of conservation of DNA replication related enzymes. He proposed that non-orthologous gene displacement (in which viruses and phages played are a bountiful source of displacing proteins), could account for the lack of conservation of DNA polymerases and other proteins of the replication apparatus and the puzzling phylogenetic patterns. Viruses and phages could have played either, as vehicles for transportation and modification of proteins from one cellular lineage to another, or as a source of proteins from viral origin. He also suggested that proteins related to the translation apparatus haven't suffered from this process because viruses can not be a source of new genes involved in protein synthesis (except for a few regulatory proteins). Latter on, he went further (Forterre, 2002), and proposed that DNA was indeed invented in viruses or phages, latter, DNA and DNA replication proteins were transferred to RNA cells to give a DNA cell.

Anyway, besides the fact that DNA polymerases are not conserved between the main cellular lineages, DNA replication, as have been noted by Leipe et al, (1999), is basically achieved in the same way in all extant cells: (i) replication is semiconservative; (ii) replication always initiate at defined origins with the participation of a origin recognition system; (iii) replication fork movement is typically bi-directional; (iv) replication is continuous in the leading strand and discontinuous in the lagging strand; (v) in cells, RNA primers are needed to start DNA replication; (vi) nucleases, polymerases and ligases replace the RNA primers with DNA and seal the remaining nicks. It is possible that several of this common characteristics of DNA replication are the result of functional constraints rather than the result of common ancestry (perhaps it is not possible to have two leading strands). Or perhaps, some of them are necessary convergences due to evolution from a common ground (DNA already existed, and primases were invented twice in different lineages in order to replicate the molecule). Anyway, it is not for granted that all cells uses DNA as their genetic material. As Forterre (1999) has pointed out, many types of nucleic acids could have been produced from RNA modification, so it seems unlikely that the LCA had an RNA genome as originally suggested by Mushegian and Koonin, (1996), because DNA had to be invented at least twice, or invented in one lineage and then, horizontally transfer the molecular machinery to other lineage. Anyway, there are some proteins for DNA replication that are homologous between the tree cellular lineages (Leipe, et al, 1999). Maybe after all, it is possible that the most parsimonious scenario is an universal ancestor with a DNA genome, a possibility that in the light of present knowledge can not be completely ruled out. If it was in fact the case, why are DNA polymerases so different between extant cellular lineages?

In order to get insight about the early evolution of replication, (i.e., whether or not the LCA could had a DNA polymerase), we review the phylogenetic distribution, polymerases, crystal structures and classification schemes of nucleic acid polymerases. From this point of view, it is clear that the evolutionary history of nucleic acid polymerase function seems to be really dynamic, including non-orthologous gene displacement, convergent evolution and domain shuffling. Our aim is to provide a general view on the evolution of this fascinating enzymes to try to understand why this molecules are so different in character between the three cellular lineages, a question that is clearly related to the nature of the LCA or *cenancestor*. Here we provide evidence relating the DNA-RNA-protein world with the previous RNA-protein world through the central replicase.

Material and methods

The following crystal structures from representatives of each family of nucleic acids polymerases were downloaded from the Brookhaven Protein Data Bank (www.rcsb.org/pdb/): DNA polymerase I from *Escherichia coli* (1kfs) (Brautigam, et al, 1998); DNA dependent DNA polymerase from bacteriophage T7 (1t7p) (Doublie, et al, 1998); DNA dependent RNA polymerase from bacteriophage T7 (1cez) (Cheetham, et al, 1999); DNA polymerase II from *Desulfurococcus sp.* (1d5a) (Zhao, et al, 1999); Bacteriophage RB69 DNA polymerase (1waj) (Wang, et al 1997); Reverse transcriptase from HIV-1 (Sarafianos, et al, 1999); Poliovirus 3d polymerase RNA-dependent RNA polymerase (1rdr) (Hansen, et al, 1997); DinB Lesion Bypass DNA polymerase *Sulfolobus solfataricus* (1im4) (Zhou, et al, 2001); DNA polymerase β from *Rattus norvegicus* (2bpf) (Pelletier, et al 1994); Terminal deoxynucleotidyltransferase (TdT) from *Mus musculus* (1jms) (Delarue, et al, 2002); Kanamycin nucleotidyltransferase (KNT) from *Staphylococcus aureus* (1kan) (Sakon, et al, 1993); RNA poly(A) polymerase from *Bos taurus* (1f5a) (Martin, et al, 2000); CCA-adding enzyme from *Bacillus stearothermophilus* (1miw) (Li, et al, 2002); DNA primase from *Escherichia coli* (1dd9) (Keck, et al, 2000); DNA primase from *Pyrococcus furiosus* (1g71) (Agustin, et al, 2001); RNA polymerase DNA dependent from *Thermus aquaticus* (1iw7) (Vassylyev, et al, 2002).

Due to the lack of crystal structures the following sequences were downloaded from NCBI (National Center for Biotechnology Information at www.ncbi.nlm.nih.gov/) Mitochondrial DNA directet RNA polymerase from *Saccharomyces cerevisiae* ssRNP (RPOM_YEAST); RNA polymerase RNA dependent from *Petunia* x hybrida (gi| 4138343); DNA polymerase III alpha chain from *E. coli* (DP3A_ECOLI); Polymerase DP1 and DP2 from *Pirococcus furiosus* (DP2S_PYRFU, DP2L_PYRFU); Qβ replicase from Qβ Bacteriophage (RRPO_BPQBE).

For proteins with known 3D structure, domain identification was followed using original papers. Whenever possible, numbers of CATH classification were added to the domains. For proteins with undetermined 3D structure, domain identification was assigned according to Pfam database (www.sanger.ac.uk/Software/Pfam/), (Bateman, et al, 2002).

The phylogenetic distribution of the nucleic acid polymerases in our database was analyzed using PSI-BLAST algorithm (Altschul, et al 1997) with an e-value cutoff 0.0001 until convergence on complete proteomes: 15 Eucaryotic (some of them fragmentary), 91 Bacterial and 16 Archaeal. Proteomes were downloaded from KEGG database (www.genome.ad.jp/kegg/) (Kanehisa, et al, 2002). The list of species with complete proteomes or with partial complete ones used in this analysis is: Eucarya: *Homo sapiens* (fragment), *Mus musculus* (fragment), *Rattus norvegicus* (fragment), *Danio rerio* (fragment), *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Oryza sativa* (fragment), *Zea mays* (fragment), *Plasmodium falciparum* (fragment), *Dictyostelium discoideum* (fragment), *Candida albicans* (fragment), *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Encephalitozoon cuniculi*; Bacteria: *Escherichia coli* K-12 MG1655, *Escherichia coli* K-12 W3110, *Escherichia coli* O157 EDL933, *Escherichia coli* O157 Sakai, *Escherichia coli* CFT073, *Salmonella typhi*, *Salmonella typhimurium*, *Yersinia pestis* CO92, *Yersinia pestis* KIM, *Shigella flexneri*, *Haemophilus influenzae*, *Pasteurella multocida*, *Xylella fastidiosa* 9a5c, *Xylella fastidiosa* Temecula1, *Xanthomonas campestris*, *Xanthomonas axonopodis*, *Vibrio cholerae*, *Vibrio vulnificus*, *Pseudomonas aeruginosa*, *Pseudomonas putida*, *Shewanella oneidensis*, *Buchnera* sp. APS, *Buchnera aphidicola* (*Schizaphis graminum*), *Buchnera aphidicola* (*Baizongia pistaciae*), *Wigglesworthia brevipalpis*, *Neisseria meningitidis* MC58 (serogroup B), *Neisseria meningitidis* Z2491 (serogroup A), *Ralstonia solanacearum*, *Helicobacter pylori* 26695, *Helicobacter pylori* J99, *Campylobacter jejuni*, *Rickettsia prowazekii*, *Rickettsia conorii*, *Mesorhizobium loti*, *Sinorhizobium meliloti*, *Agrobacterium tumefaciens* C58 (UWash/Dupont), *Agrobacterium tumefaciens* C58 (Cereon), *Brucella melitensis*, *Brucella suis*, *Bradyrhizobium japonicum*, *Caulobacter crescentus*, *Bacillus subtilis*, *Bacillus halodurans*, *Oceanobacillus iheyensis*, *Staphylococcus aureus* N315 (MRSA), *Staphylococcus aureus* Mu50 (VRSA), *Staphylococcus aureus* MW2, *Staphylococcus epidermidis* , *Listeria monocytogenes*, *Listeria innocua*, *Lactococcus lactis*, *Streptococcus pyogenes* SF370 (serotype M1), *Streptococcus pyogenes* MGAS8232 (serotype M18), *Streptococcus pyogenes* MGAS315 (serotype M3), *Streptococcus pneumoniae* TIGR4, *Streptococcus pneumoniae* R6, *Streptococcus agalactiae* 2603, *Streptococcus agalactiae* NEM316, *Streptococcus mutans*, *Clostridium acetobutylicum*, *Clostridium perfringens*, *Clostridium tetani*, *Thermoanaerobacter tengcongensis*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Mycoplasma pulmonis*, *Mycoplasma penetrans*, *Ureaplasma urealyticum*, *Mycobacterium tuberculosis* H37Rv (lab strain), *Mycobacterium tuberculosis* CDC1551, *Mycobacterium leprae*, *Corynebacterium glutamicum*, *Corynebacterium efficiens*, *Streptomyces coelicolor*, *Bifidobacterium longum*, *Fusobacterium nucleatum*, *Chlamydia trachomatis*, *Chlamydia muridarum*, *Chlamydophila pneumoniae* CWL029, *Chlamydophila pneumoniae* AR39, *Chlamydophila pneumoniae* J138, *Borrelia burgdorferi*, *Treponema pallidum*, *Leptospira interrogans*, *Synechocystis* sp. PCC6803, *Thermosynechococcus elongatus*, *Anabaena* sp. PCC7120 (*Nostoc* sp. PCC7120), *Chlorobium tepidum*, *Deinococcus radiodurans*, *Aquifex aeolicus*, *Thermotoga maritima*; Archaea: *Methanococcus jannaschii*, *Methanosarcina acetivorans*, *Methanosarcina mazei*, *Methanobacterium thermoautotrophicum*, *Methanopyrus kandleri*, *Archaeoglobus fulgidus*, *Halobacterium* sp.NRC-1, *Thermoplasma acidophilum*, *Thermoplasma volcanium*, *Pyrococcus horikoshii*, *Pyrococcus abyssi*, *Pyrococcus furiosus*, *Aeropyrum pernix*, *Sulfolobus solfataricus*, *Sulfolobus tokodaii*, *Pyrobaculum aerophilum*.

Results.

Classification of polymerases

Nucleic acids polymerases has been classified in several occasions. One of us (Lazcano, et al 1988) had already classified some of the RNA polymerases discussed in this paper into two families (cellular ββ' transcriptases and viral RNA polymerases). Later on, the bases of the modern classification of DNA polymerases were settle down by Ito and Braithwaite (1991). They classified polymerases into four families, these are: family A (DNA polymerase I from *E. coli*), family B (*E. coli* DNA polymerase II and replicative Eucaryotic α, δ, and ε DNA polymerase as well as Archaeal B DNA polymerases), family C (Bacterial DNA polymerase) and family X (nucleotidyltransferases).

More recently, using amino acid comparasions as well as crystal structure analysis Joyce and Steitz (1994), and Steitz (1999), classified polymerases into five families, (four of them are the same as in Ito and Braithwaite, 1991). These are: DNA polymerase I family or simply family A that includes the Klenow fragments of *E. coli* and *Bacillus* DNA pol I, *Thermus aquaticus* DNA polymerase and the T7 RNA and DNA polymerases; DNA polymerase α family or family B, that includes all eucaryotic replicating polymerases (α, δ, and ε), polymerases from phages T4 and RB69, and archaeal B polymerases; Nucleotidyltransferases (family X) that include rat DNA polymerase β and a large variety of molecules; and on the basis of sequence comparisons but no crystal structure DNA polymerase III (family C) that shows no relatedness to the other families of polymerases; and finally Reverse transcriptases (RT), RNA dependent RNA polymerases and telomerases seems to conform another family of enzymes.

Recently, new polymerases have been found in Euryarchaea named D polymerase (Ishino, et al 1998) and in Bacteria and Eucarya (Goodman and Tippin, 2000). With this findings, a new classification of polymerases that include the previous classification of Ito and Braithwaite (1991) plus the new Euryarchaeal family D of polymerases and the family Y of polymerases that include members from the three domains (Bacteria: DinX, DinB and UmuD; Eucarya: Rad30, ι, *DinB and REV1*; Archaea: Dbh) has been proposed by Fileé et al (2002). As we can see, the classification of polymerases has grown up to accommodate the new molecules that have recently been discovered and described.

In this paper, we want to expand the classification of polymerases to include also other important enzymes that polymerize nucleic acids, not included in previous classifications, and that generically conform a group of enzymes that polymerize nucleic acids. Accordingly, cellular polymerases (and some phage and viral homologs) can be classified at least on 10 different families (Table 1). There are crystal structures for representatives of 8 of the families (DNA pol I, DNA pol II, Reverse transcriptase and RNA polymerase, family Y, Nucleotidyltransferases, Bacterial primases, Archaeal-Eucaryal primases and the cellular ββ' transcriptase), there are not structures available for the remaining two families (Bacterial DNA pol III and the Euryarchaeal DNA pol D).

Phylogenetic distribution of polymerases

In order to review as much as possible the phylogenetic distribution of the members from different families of polymerases, we will discuss the results of our analysis (see methods) as well as those found in other previously published reviews and analysis (Nakamura and Cech, 1998; Fileé, et al 2002; Bohlke et al, 2002; Aravind and Koonin, 1999) for the sake of comparison and completeness.

For instance DNA polymerases from family I has a wide distribution among Bacteria (it was found in almost all bacterial genomes analyzed here). In Eucaryots DNA pol I Mus308 and its homologues (also called DNA pol eta in human and theta in mouse) it is present in some plants and animals. For instance, a BLAST search of human Mus308 (gi|16418479) against non-redundant database from NCBI, retrieve sequences homologues from *Mus musculus*, *Rattus norvergicus*, *Anopheles gambiae*, *Drosophila melanogaster* and *Arabidopsis thaliana*. On the other hand, mitochondria also harbors DNA polymerases and DNA dependent RNA polymerases from family I closely related to the DNA dependent DNA polymerase from Bacteriophage T7 (1t7p) and the DNA dependent RNA polymerase from Bacteriophage T7 (1cez). Phylogenetic analysis of family I DNA polymerases suggest that this family originated in Bacteria and the was horizontally transferred from Bacteria to originate the Mus308 and its homologues, and the mitochondrial polymerases originated from a non orthologous gene substitution from Bacteriophages T3/T7 (Fileé, et al 2002; Gray and Lang 1998). Although this conclusion should be taken with caution because the DNA pol I family tree is unrooted (Fileé, et al 2002).

Although members from DNA polymerase family II are present in the three cellular lineages (with a wide distribution among Eucarya and Archaea) it is very likely that its presence in the bacterial ur-kingdom is due to horizontal gene transfer. For instance, we found members of this family only in the proteomes of the proteobacterias *E. coli*, *Salmonella sp.*, *Yersinia pestis*, *Shigella flexneri*, *Vibrio sp.*, *Pseudomonas sp.*, *Shewanella oneidensis*, the cyanobacteria *Anabaena sp.*, and the green-sulfur bacteria *Chlorobium tepidum*. In Bacteria this enzyme has been implicated in the SOS repair response (Kornberg and Baker, 1992). On the other Polymerases α, δ, and ε from this family have been involved in chromosomal DNA replication in Eucaryonts (Hubscher et al, 2000). And as expected,

members of this family were found in all complete Eucaryal proteomes analyzed here. Also, members of this family were found in all Archaeal proteomes. In Crenarchaea, it has been suggested that they are the main replicative enzymes (reviewed in Bohlke et al, 2002) while in Euryarchaea they may play another role (perhaps reparation). In Archaea DNA polymerases II are further subdivided in three subfamilies (B1, B2 and B3). Chrenarchaeal species posses two (B1 and B3) or three (B1, B2 and B3) family II DNA polymerases and Euryarchaea posses only DNA polymerase type B3, with the exception of *A. fulgidus* and *Halobacterium* NRC1 which also have a B2 type DNA polymerase (reviewed in Bohlke et al, 2002).

As reviewed by Steitz (1999) Telomerases, Reverse Transcriptases and RNA dependent RNA polymerase appear to show some common structural similarities. Attempts to understand their phylogenetic relationships have been performed by Nakamura and Cech (1998). This family comprises a diverse set of enzymes that makes it difficult to understand its phylogenetic distribution. Using Reverse transcriptase from HIV-1 (1qe1) as a query against our dataset of complete proteomes we found homologous enzymes only in the Eucaryal proteomes of *D. melanogaster*, *C. elegans*, *A. thaliana*, *S. cerevisiae*, and *S. pombe*. And we were unable to detect any homologous sequence to the RNA dependent RNA polymerase from Poliovirus (1rdr). Anyway, it has been reported that there are several genetic elements encoding homologous enzymes from this family from a wide phylogenetic spectrum encompassing Bacteria and Eucarya: retrovirus and hepadnavirus in mammals and birds, caulimoviurs in plants, LTR, non-LTR retrotransposons and Telomerases in Animals, Plants, Fungi and Protoctist, group II intron in Bacteria, Fungi and Plant mitochondria, and chloroplast algae plastid, and Retron (msDNA) in purple and other bacteria (Nakamura et al, 1998). According to Pfam database, reverses transcriptases and telomerases (rvt family, accession number: PF00078) are encoded by a variety of Eucaryots (Fungi, Animals, Protists and Plants), Bacteria (Cyanobacteria, Actinobacteria, Bacteroidetes, Proteobacteria, Fusobacteria and Firmicutes) and Euryarchaeotas (*Methanosarcina acetivorans* and *Methanosarcina mazei*) as well as viruses. On the other hand, RNA dependent RNA polymerases (RNA_dep_RNA_pol domain, accession number: PF00680) is encoded by a wide variety of viruses. Based on the nearly universal distribution of telomerases among Animals, Plants, Fungi and Protoctist, it has been suggested by Nakamura et al (1998) that such enzyme was already present in the first Eucaryots.

DinB Lesion Bypass DNA polymerase from *Sulfolobus solfataricus* (1im4) belong to the Y family of DNA polymerases. These polymerases are implicated in replace stalled replicative polymerases and synthesize DNA past the damaged site when DNA has been damaged and it hasn't been repaired previously. According to Zhou et al, (2001) this family of polymerases can be subdivided in several subfamilies: DinB subfamily is present in all three domains of life, UmuC subfamily has only been found in Bacteria, and the Rad30A/B and Rev1 subfamilies have only been found in Eucaryonts. Although the family of DNA polymerases Y is present in the three cellular lineages, the phylogenetic analysis done by Fileé et al (2002) didn't suggest any clear evolutionary hypothesis about its origins. This fact, together with its scarce phylogenetic distribution (i.e., they are present in some of the genomes of Animals, Protists, Fungi, Proteobacteria, Firmicutes, Actinobacteria, Fusobacteria, Spyrochete, Cyanobacteria, and in some of the Euryarchaeas and Crenarchaeas analyzed here), make it difficult to asses if this kind of enzymes were already part of the genome of the LCA.

Nucleotidyltransferases is a family that include a large variety of molecules that can be subdivided in at least nine different groups (Aravind and Koonin, 1999). DNA polymerase β from *Rattus norvegicus* (2bpf), Terminal deoxynucleotidyltransferase (TdT) from *Mus musculus* (1jms), Kanamycin nucleotidyltransferase (KNT) from *Staphylococcus aureus* (1kan), RNA poly(A) polymerase from *Bos taurus* (1f5a) and CCA-adding enzyme from *Bacillus stearothermophilus* (1nuw) belong all to the family of nucleotidyltransferases. Members of the Nucleotidyltransferase family of polymerases are found in the three cellular lineages, but only one of the groups orthologous of nucleotidyltransferases (DNA polymerase X group) has members on the three domains of life. Anyway, its phylogenetic distribution is patchy among the Bacteria and in Archaea is found only in *Methanobacterium thermoautotrophicum* (Aravind and Koonin, 1999), so, it seems that there is no a single group of orthologous proteins from this family that is present in a wide number of genomes and could be confidently traced back to the LCA, as is the case of the cellular transcriptase ββ'. Nucleotidyltransferases are characterized by sharing the "head" catalytic domain (Figure 3) Li, et al (2002). The catalytic motif in this domain was also described by Aravind and Koonin (1999) as definitory of the family of nucleotidyltransferases. It has also been suggested by Aravind and Koonin (1999) that a set of nucleotidyltransferases comprising the minimal domain harboring the definitory motif present in Archaea and in a few Bacteria may represent the ancestral sate of this family of proteins. If this is true, then the most parsimonious scenario would suggest that this family of enzymes arose in de ur-kingdom of Archaea. Anyway this speculation should be taken with caution because these molecules can also be the result of reductive evolution.

Polymerase DP2 from *Pirococcus furiosus* (DP2L_PYRFU) belongs to family D of DNA polymerases. DNA polymerases from the D family are restricted to Euryarchaea. It has been suggested that they are the main replicative enzyme in this subkingdom (Cann, et al, 1998). The holoenzyme is composed of two subunits, the large subunit (DP2) harbors the catalytic and the 3'-5' exonuclease activity and has no sequence similarity to other polymerases. Perhaps once the crystal structure became known it will show up similarity to the 3'-5' exonuclease domain from polymerases I, II and III, and to some of the catalytic polymerizing domains form those enzymes. The catalytic reaction occurs only in

the presence of the small subunit (Bohlke, et al, 2002). The small subunit (DP1) present similarities to the small subunit of δ polymerases from Eucaryonts (Cann, et al, 1998). Both, seems to mediate contacts between the polymerases and the accessory subunits (Bohlke, et al, 2002).

Each one of the two families of primases (the Bacterial and Archaeal-Eucaryal) are restricted to their phylogenetic lineages (Leipe, et al 1999). On the other hand, DNA dependent RNA polymerase from *Thermus aquaticus* (1tw7) and the RNA polymerase RNA dependent from *Petunia* x hybrida (gi|4138343) belong to the family of cellular transcriptases. While the cellular transcriptase ββ' is conserved in all cellular genomes and must have been present in the LCA, the RNA polymerase RNA dependent that is involved in the amplification of regulatory microRNAs during post-transcriptional gene silencing is restricted to Eucarya and bacteriophages (Iyer, et al, 2003). Both enzymes share the same catalytic double-psi β-barrel domain (Iyer, et al, 2003). As mentioned before, DNA polymerase III α chain from *Eschericia coli* (DP3A_ECOLI) belong to DNA polymerase III family. The phylogenetic distribution of this family is restricted to Bacteria. Because there is no crystal structure from this enzyme it is not known if it shares a catalytic domain with other DNA polymerases or if it has a unique domain architecture.

As we can see, the only polymerase that is distributed in all cellular genomes, and thus can be confidently traced back to very early steps during the early evolution of life is the ββ' transcriptase. The rest of the enzymes are restricted to one lineage (like is the case of the bacterial DNA polymerase III, the bacterial primase, and the euryarchaeal DNA *polymerase D), or two of the main cellular lineages (like DNA polymerase I, and the Archaeal-Eucaryal primases), or* present in the three lineages but with a scatter distribution in one or two of the lineages, which may be an indication of horizontal gene transfer (like is the case of DNA polymerase II and the family Y of DNA polymerases), or present in the three ur-kingdoms because of a likely combination of horizontal gene transfer and presence of different paralogs in different lineages (like maybe is the case of the telomerases, reverse transcriptases and RNA dependent RNA polymerases in one hand, and in the other, the family of nucleotidyltransferases).

Conservation of structural domains between cellular nucleic acid polymerases

The first polymerase to be crystallized was the Klenow fragment from *Escherichia coli* DNA polymerase I (Ollis et al, 1985). The core structure of polymerases has been described as a hand, consisting of subdomains: fingers, palm and thumb (Kohlstaedt, et al, 1992). The function of the palm domain appears to be the catalysis of the phosphoryl transfer reaction whereas that of the fingers domain includes important interactions with the incoming nucleoside thriphosphate as well as the template base to which it is paired, the thumb on the other hand may play a role in positioning the dupiex DNA and in processivity and translocation (Steitz, 1999). As shown in Table 1, some of the families of polymerases are interconnected through homologous domains. The palm subdomains of polymerases from families I, II, telomerase, reversetranscriptase, RNA dependent RNA polymerase and family Y DNA polymerase, share enough structural similarity to support the hypothesis that they descend from a common ancestral palm domain (Steitz, 1999; Hansen, et al, 1997; Zhou, et al, 2001) (Figure 1). Polymerases from family I (DNA Polymerase I from *Escherichia coli*, DNA polymerase from Bacteriophage T7) and polymerases from family II (DNA polymerase II from *Desulfurococcus* sp. and DNA polymerase from Bacteriophage RB69) and the ε subunit of *E. coli* DNA polymerase III share an homologous 3'-5' exonuclease domain (Hamdan, et al, 2002) (Figure 1). This last domain (from DNA polymerase III) *that in E. coli correspond to the ε subunit from DNA polymerase holoenzyme (Kornberg and Baker, 1992) is coded in* a different subunit from the catalytic subunit (α subunit), but in other organisms is fused in a single polypeptide together with the catalytic subunit, as is the case of the Bacteria *Ureaplasma parvum* DNA polymerase III (DPO3_UREPA). Although not in the catalytic subunit, the DP1 regulatory subunit of DNA polymerase D from *P. furiosus* shares an homologous domain with the catalytic (α subunit) of *E. coli* DNA polymerase III (data not shown). At the light of present knowledge, no other domains are shared between the different families of the catalytic subunits of nucleic acid polymerases. The fact that several different families of polymerases are interconnected through homologs domains that are only recognizable as such, at the level of tertiary structure, clearly suggest that there may be a shared ancient history for these families of polymerases, and opens the possibility that this common history may be extended to the other families of polymerases (with out known 3D structure) once their tertiary structures become available.

On the other hand, the structure of rat DNA polymerase β nucleotidyltransferase has also been described using the metaphor of a hand (Sawaya, et al, 1994). Although there has been some discussion regarding if the classical palm domain of *E. coli* DNA polymerase I and the palm domain from DNA polymerase β are homologous (Sawaya, et al, *1994) there are important differences between the two that clearly suggest that these domains are phylogenetically* unrelated (Steitz, et al, 1994). As more enzymes from the nucleotidyltrasnferase family were crystallized, it became clear that all of them share the same "palm" catalytic domain (unrelated to the palm domain from polymerases I, and related enzymes). These domain was named as "head" domain in the CCA adding enzyme from *Bacillus stearothermophilus* for the resemblance of this enzyme with a seahorse, and the position of the catalytic domain in the N-terminal region of the molecule (Li et al, 2002). Thus, in order to avoid confusion between both domains (from *E. coli* Klenow fragment and from rat DNA polymerase β), we decide in this paper to reference as head domain all

catalytic domains of nucleotidyltransferases, including the so named palm domain from rat DNA polymerase β (Table 1).

As the structure of more nucleic acid polymerases became known, it was also clear that not all of them had the classical architecture of a hand with the catalytic domain positioned in the domain equivalent to the palm region of the hand, as is the case of the CCA adding enzyme from *B. stearothermophilus*, the RNA poly(A) polymerase from *B. taurus*, the DNA primase from *E. coli*, the DNA primase from *P. furiosus* and the RNA polymerase DNA dependent from *T. aquaticus* (Keck, et al, 2000; Martin, et al 2000; Agustin, et al, 2001; Li, et al, 2002; Vassylyev, et al, 2002). This shows that different architectures can be used in order to polymerize nucleic acids. In some degree, each one of the architectures may represent particular solutions to perform different kinds of nucleic acid polymerization (i.e., for instance, the CAA adding enzyme from the family of nucleotidyltransferases is the only enzyme capable of sequence specific template independent polymerization). Also, the catalytic subunit of polymerases inside one family has unique topologies that do not share with polymerases from other families besides those polymerases that share the palm domain as previously mentioned.

Conservation of catalytic carboxylate motifs between nucleic acid polymerases

Basically, all polymerases have carboxylate residues that are used to bind metal ions in order to catalyze the polymerization reaction (Beese, et al, 1991; Steitz, et al, 1993). Previous alignment sequences had identified conserved motifs in nucleic acid polymerases (Delarue, et al 1990). Recently, there has been an update in the alignment of this motifs using information derived from crystal structures (Wang, et al, 1997). The motifs A and C of Delarue et al, (1990) correspond approximately with the carboxylate residues involved in binding of metal ions for polymerases containing the palm domain. We have searched visually for the equivalent of motif C in the enzymes analyzed here that weren't included in Delarue's (1990) paper. These motifs, are certainly the result of convergence in the cases when the catalytic topologies of the domains are different (Figure 2). These convergences suggest that the chemistry of nucleic acid polymerization imposes certain stereo-chemical rules to the different solutions that evolution has found in the different families of polymerases in order to polymerize nucleic acids.

Discussion

On the polyphiletic origin of catalytic domains from polymerases

One of the striking features about nucleic acids polymerases, is that clearly the polymerase function has been invented more than once over the history of life. As shown in Table 1, without counting the polymerases for which there is no structure available yet, there are at least five different catalytic domains that perform this function. The fact that several different catalytic domains can polymerize nucleic acids, suggest that the sequence space of proteins capable of nucleic acid polymerization is rather large, and that new proteins that would eventually evolve this function, likely, would done so from different domains extracted from the pools of the existing protein domains. In this vein we suggest that domains having different structures arose independently and that all palm domains are related by common ancestry, rather than have evolved by convergence.

On the other hand, all kinds of nucleic acid polymerization, whether DNA or RNA replication, DNA or RNA dependent, as well as template independent NTP or dNTP polymerization seems to be based in the same chemistry of metal ions coordinated by conserved caroxylated residues (Beese, et al, 1991). This chemistry was also proposed for ribozymes that cleavage nucleic acids like RNase P (Steitz, 1993), so it seems that it is a general theme when dealing with DNA or RNA cleavage or polymerization. In fact, the carboxilated residues are superimposable between the structures of DNA polymerase I from *E. coli*, Reverse transcriptase from HIV-1, T7 RNA polymerase and rat DNA polymerase β (Steitz, et al, 1994; Pelletier, 1994), and between rat DNA polymerase β and DNA primase from *Pyrococcus furiosus* (Agustin, et al, 2001). Although the protein domains are structurally unrelated (palm domain, head domain and prim domain respectively for families DNA pol I, Nucelotidyltransferases and Archaeal-Eucaryal primases) they converged to the same stereo-chemistry.

Two cases of non-orthologous gene displacements

The evolutionary history of nucleic acid polymerization function has at least two cases of events of non-orthologous gene displacements. One of the most clear cases is the one of mitochondria were the original replicative DNA polymerase III has been replaced by a DNA polymerase from family I (γ polymerase) (Foury, 1989), and the RNA transcriptase, of viral origin, is also from family I (Gray, et al, 1998). This case is specially interesting because the fact that a genome (in this case a mitochondrial genome) is replicated and transcribed only with polymerases from family I suggest how an ancient cell could have survived with only one class of polymerases for these functions.

Another pattern that may suggest an event of non-orthologous gene displacement, is that of *P. furiosus* DNA polymerase from family D. The polymerase is an heterodimer composed of DP1 and DP2 independent subunits (Cann, et al, 1998). The enzyme DP2 has the polymerase and the 3'-5' exonuclease activity, and has no sequence similarity to any other family of polymerases at primary structure level. On the other hand, DP2 regulates the level of DNA polymerase activity, and is homologous to the non catalytic Eucaryotic Pol δ small subunit. The catalytic subunit of

Eucaryotic Pol δ is a DNA polymerase from family II (Eucaryotic Pol δ large subunit) and *P. furiosus* has in fact a DNA polymerase from family II (Pol I) (Uemori, et al, 1993). Anyway, DP1 interacts with DP2 rather than with Pol I as would be expected (Cann, et al, 1998). Although we can not discard the possibility that DP2 is homologous to polymerases from family II and we are unable to detect the relationship at the level of primary structure, it is also likely that during the early evolution of Euryarchaea, DP2 had an independent origin and displaced an enzyme from DNA pol II family.

The key, the lock and evolution of substrate specificity, the weak connection

From an evolutionary point of view, it also seem to be easy for polymerases to change the specificity for one kind of nucleotides for other (NTP, dNTP). As we can see from Table 1, different kinds of polymerizations has evolved inside several of the families of polymerases (i.e., in family I there are enzymes that can polymerize both kinds of nucleic acids, DNA and RNA). In fact, it has been reported that a single mutation in Molone murine leukemia virus reverse transcriptase is sufficient to change the specificity of this enzyme from DNA to RNA (Gao, et al, 1997). Also, as reviewed by Lazcano et al, (1988) and Lazcano et al, (1991) it has been shown experimentally that substituting Mg++ by Mn++ increase the misincorporation of deoxyribonucleotides by DNA-dependent DNA polymerases and in DNA-dependent RNA polymerases and alters their substrate specificities. This is important because suggest that once dNTP evolved, existing RNA polymerases could have retrotranscribed RNA to DNA and then polymerized DNA without too many changes.

Two likely cases of modular evolution among polymerases

As happens with the majority of proteins (Henikoff, et al, 1997), the evolution of DNA polymerases is modular to some extent. This is, domains are the structural and evolutionary unit, and recruitment and shuffling of domains is an important factor in the evolution of this enzymes. This seems to be true for palm and head catalytic domains. In both cases, these domains are found associated with different domains in different families, (for nucleotidyltransferases the head domain is associated with different domains inside the same family of proteins, and the palm domain is associated with different domains in the different families of polymerases, Figure 3). The distribution of the palm domain goes further because this domain is found in the catalytic domain of adenylyl cyclase (Artymiuk, et al, 1997). Domains similar to the palm domain has also been reported for several proteins: U1A RRM, ribosomal proteins L7/L12 and S6, the anticodon binding domain of phenylalanyl-tRNA synthetase, the phosphocarrier protein Hpr, the enzyme acyl phosphatase, the signal transducing protein PII, the regulatory subunit of aspartate transcarbamylase, nucleotide diphosphate kinase and procarboxypeptidase (Hansen, et al, 1997, and references therein). Whether the similitude of these domains is due to convergent or divergent evolution is still unknown but if its due to common ancestry it suggest that this domain belong to an ancient lineage that goes well before the LCA.

On the early evolution of DNA polymerases

As we have seen, the evolution of polymerase function shows a wide spectrum of evolutionary patterns that goes from polyphyletic origin of enzymes and structural convergence of active sites, events of non-orthologous gene displacement and cases in which the enzyme is conserved across all cellular genomes.

On the other hand, although the central enzymes of DNA replication are not conserved between the main cellular lineages, as mentioned before, genome replication is a process that is performed in a similar way in all extant cells (Leipe, et al, 1999). Because all cells had DNA based genomes, in principle, the most parsimonious scenario is a common ancestor with a DNA based genome. How can we explain this pattern?

For instance, Steiz (1999) has suggested that perhaps a ribozyme DNA polymerase originating in the "RNA world" may have persisted beyond the divergence of Eukaryotes and Prokaryotes and was replaced domain by domain differently. Anyway, it is also possible that by the time of the LCA this ribozyme was already replaced by a protein enzyme and another process is responsible of the pattern that we see today. Also, Lazcano et al, (1988) and Lazcano et al, (1992) in concordance with the high degree of conservation of extant RNA transcriptase ββ' among cellular genomes, has proposed that this enzyme functioned as a replicase during the RNA-protein world. Anyway, we still need a theory of how and when extant replicative machineries evolved. Nakamura et al (1998) also proposed that according to the RNA-world hypothesis, the RNA polymerases RNA dependent virus-like (belonging to the same family of reverse transcriptase and telomerase) was the replicase and that reverse transcriptases arose in the transition from the RNA-protein to the present DNA-RNA-protein world by gene duplication.

Here, we would like to propose an hypothesis to explain the conservation of several aspects of DNA replication, and the differences between the different replicative polymerases is needed based on recognizable patterns of polymerase evolution, such as the polyphiletic origin of the different enzymes that polymerize nucleic acids, the cases of non-orthologous gene displacements and the degree of conservation of the palm domain.

In accordance with the RNA world hypothesis, and if DNA was preceded by proteins (Froeland, et al, 1999), we suggest that a polymerase having the palm domain was the first protein enzyme that replicated an RNA genome during the RNA-protein stage of evolution, previous to the LCA (Figure 4). This ancient enzyme may have resembled extant

viral RNA polymerases RNA dependents like the Poliovirus 3d polymerase RNA dependent RNA polymerase in the sense that it is a RNA polymerase with a palm domain. With the apparcance of DNA, genetic information coded on RNA has to be retrotranscribed into DNA. In this way, RNA polymerases RNA dependent could have originated Reversetrascriptases (in fact both kind of enzymes belong to the same family), although this step may not be totally indispensable because, as mentioned before, there is certain degree of lack of specificity of polymerases. Finaly, this enzymes (perhaps Reversetranscriptases or RNA polymerases) originated DNA polymerases DNA dependents, harboring a palm domain, in order to duplicate a DNA genome. As mentioned before, in mitochondria, polymerases from family I perform both functions, replication and transcription, suggesting how ancient cells could have survived only with this kind of polymerases. Also, enzymes having the palm domain are the main replicative enzymes in Archaea (at least in Chrenarchaca) and in Eucaryonts (polymerases from family II). If this scenaryo is true, then the replicative machinery of Archaeas and Eucaryonts was the replicative marchinery of the LCA and evolved prior the replicative machinery of Bacteria. The replicative machinery of Bacteria arrose in the split of this lineage from the Archacal-Eucaryal linage by a non-orthologous gene displacement event. The origin of Bacterial DNA polymerase as an event of non-orthologous gene displacement has been suggested before (Cavalier-Smith, 2002). Here, we suggest that extant Archaeal-Eucaryal replication machinery can be traced back as the replication machinery of the LCA. On the other hand, if extant Reverse Transcriptases and Viral RNA polymerases RNA dependents are direct descendants of the enzymes that preceded the proposed replication machinery of the LCA (i.e., extant viral RNA polymerases RNA dependents are direct descentants from the RNA-protein world) or if they arose after the divergence of the cellular lineages, is a difficult question that is out of the scope of this paper. The central point is that all the important functions necessary to understand the lineage of evolutionary transitions from an ancient cell with an RNA genome replicated by a protein enzyme to one of the extant replication machineries (the Archaeal-Eucaryal one) are harbored by four families of DNA polymerases sharing the palm catalytic domain (DNA polymerases from families I, II, Telomerase Reverse Transcriptases and RNA dependent RNA polymerases and family Y of DNA polymerases).

References.

Altschul, SF., Madden, TL., Schaffer, AA., Zhang, Z., Miller, W., Lipman, DJ. (1997) Gapped BLAST and PSI-BLAST: *a new generation of protein database search programs. Nucleic Acid Res* **25**: 3389-3402

Aravind, L., Koonin, E.V. (1999) DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history. *Nucleic Acids Res* **27**: 1609-1618

Artymiuk, P.J., Poirrette, A.R., Rice, D.W., Willett, P. (1997) A polymerase I palm in adenylyl cyclase? *Nature* **388**: 33-34

Augustin, M.A., Huber, R., Kaiser, J.T. (2001) Crystal structure of a DNA-dependent RNA polymerase (DNA primase). *Nat Struct Biol* **8**: 57-61

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R , Griffiths-Jones, S., Howe, K.L., Marshall, M., Sonnhammer, E.L. (2002) The Pfam protein families database. *Nucleic Acids Res* **30**: 276-280

Beese, L.S., Steitz, T.A. (1991) Structural basis for the 3'-5' exonuclease activity of Escherichia coli DNA polymerase I: *a two metal ion mechanism. EMBO J* **10**: 25-33

Brautigam, C.A., Steitz, T.A. (1998) Structural principles for the inhibition of the 3'-5' exonuclease activity of *Escherichia coli* DNA polymerase I by phosphorothioates. *J Mol Biol* **277**: 363-377

Bohlke, K., Pisani, F.M., Rossi, M. and Antranikian, G. (2002) Archaeal DNA replication: spotlight on a rapidly moving field. *Extremophiles* **6**: 1-14

Cann, I.K., Komori, K., Toh, H., Kanai, S., Ishino, Y. (1998) A heterodimeric DNA polymerase: *evidence that* members of Euryarchaeota possess a distinct DNA polymerase. *Proc Natl Acad Sci U S A* **95**: 14250-14255

Cavalier-Smith, T. (2002) The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int J Syst Evol Microbiol* **52**: 7-76

Cheetham, G.M., Jeruzalmi, D., Steitz, T.A. (1999) Structural basis for initiation of transcription from an RNA *polymerase-promoter complex. Nature* **399**: 80-83

Delarue, M., Poch, O., Tordo, N., Moras, D., Argos, P. (1990) An attempt to unify the structure of polymerases. *Protein Eng* **3**: 461-467

Delarue, M., Boule, J.B., Lescar, J., Expert-Bezancon, N., Jourdan, N., Sukumar, N., Rougeon, F., Papanicolaou, C. (2002) Crystal structures of a template-independent DNA polymerase: murine terminal deoxynucleotidyltransferase. *EMBO J* **21**: 427-439

Doublie, S., Tabor, S., Long, A.M., Richardson, C.C., Ellenberger, T. (1998) Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 A resolution. *Nature* **391**: 251-258

Edgell, D.R., Doolittle, W.F. (1997) Archaea and the origin(s) of DNA replication proteins. *Cell* **89**: 995-998

Filée, J., Forterre, P., Sen-Lin, T., Laurent, J. (2002) Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J Mol Evol* **54**: 763-773

Forterre, P. (1999) Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. *Mol Microbiol* **33**: 457-465

Forterre, P. (2002) The origin of DNA genomes and DNA replication proteins. *Curr Opin Microbiol* **5**: 525-32

Foury, F. (1989) Cloning and sequencing of the nuclear gene MIP1 encoding the catalytic subunit of the yeast mitochondrial DNA polymerase. *J Biol Chem* **264**: 20552-20560

Freeland, S.J., Knight, R.D., Landweber, L.F. (1999) Do proteins predate DNA? *Science* **286**: 690-692

Gao, G., Orlova, M., Georgiadis, M.M., Hendrickson, W.A., Goff, S.P. (1997) Conferring RNA polymerase activity to a DNA polymerase: a single residue in reverse transcriptase controls substrate selection. *Proc Natl Acad Sci U S A* **94**: 407-411

Goodman, M.F. and Tippin, B. (2000) The expanding polymerase universe. *Nat Rev Mol Cell Biol* **1**: 101-109

Gray, M.W., Lang, B.F. (1998) Transcription in chloroplasts and mitochondria: a tale of two polymerases. *Trends Microbiol* **6**: 1-3

Hamdan, S., Carr, P.D., Brown, S.E., Ollis, D.L., Dixon, N.E. (2002) Structural basis for proofreading during replication of the Escherichia coli chromosome. *Structure* (Camb). **10**: 535-546

Hansen, J.L., Long, A.M., Schultz, S.C. (1997) Structure of the RNA-dependent polymerase of polivirus. *Structure* **5**: 1109-1122

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K., Hood, L. (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**: 609-614

Hubscher, U., Nasheuer, H.P., Syvaoja, J.E. (2000) Eukaryotic DNA polymerases, a growing family. *Trends Biochem Sci* **25**: 143-147

Iyer, L.M., Koonin, E.V., Aravind, L. (2003) Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struct Biol* **3**: 1

Joyce, C.M. & Steitz, T.A. (1994) Function and structure relationships in DNA polymerases. *Annu Rev Biochem* **63**: 777-822

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res* **30**: 42-46

Keck, J.L., Roche, D.D., Lynch, A.S., Berger, J.M. (2000) Structure of the RNA polymerase domain of E. coli primase. *Science* **287**: 2482-2486

Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice, P.A., Steitz, T.A. (1992) Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Nature* **256**: 1783-1790

Kornberg, A. & Baker, T.A. DNA Replication, 2nd Ed. W.H. Freeman USA, pp. 931

Lazcano, A., Fastag, J., Gariglio, P., Ramirez, C., Oro, J. (1988) On the early evolution of RNA polymerase. *J Mol Evol* **27**: 365-376

Lazcano, A., Llaca, V., Cappello, R., Valverde, V. and Oro, J. (1992) the origin and early evolution of nucleic acid polymerases. *Adv Space Res* **12**: 207-216

Leipe, D.D., Aravind, L., Koonin, E.V. (1999) Did DNA replication evolve twice independently? *Nucleic Acids Res* **27**: 3389-3401

Li, F., Xiong, Y., Wang, J., Cho, H.D., Tomita, K., Weiner, A.M., Steitz, T.A. (2002) Crystal structures of the *Bacillus stearothermophilus* CCA-adding enzyme and its complexes with ATP or CTP. *Cell* **111**: 815-824

Martin, G., Keller, W., Doublie, S. (2000) Crystal structure of mammalian poly(A) polymerase in complex with an analog of ATP. **19**: 4193-4203

Mushegian, A.R., Koonin, E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* **93**: 10268-10273

Nakamura, T.M., Cech, T.R. (1998) Reversing time: origin of telomerase. *Cell* **92**: 587-590

Ollis, D.L., Brick, P., Hamlin, R., Xuong, N.G., Steitz, T.A. (1985) Structure of large fragment of Escherichia coli DNA polymerase I complexed with dTMP. *Nature* **313**: 762-766

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

Pelletier, H. (1994) Polymerase structures and mechanism. *Science* **266**: 2025-2026

Pelletier, H., Sawaya, M.R., Kumar, A., Wilson, S.H., Kraut, J. (1994) Structures of ternary complexes of rat DNA polymerase beta, a DNA template-primer, and ddCTP. *Science* **264**: 1891-1903

Ruff, M., Krishnaswamy, S., Boeglin, M., Poterszman, A., Mitschler, A., Podjarny, A., Rees, B., Thierry, J.C., Moras, D. (1991) Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(Asp). *Science* **252**: 1682-1689

Sakon, J., Liao, H.H., Kanikula, A.M., Benning, M.M., Rayment, I., Holden, H.M. (1993) Molecular structure of kanamycin nucleotidyltransferase determined to 3.0-A resolution. *Biochemistry* **32**: 11977-11984

Sarafianos, S.G., Das, K., Clark, A.D. Jr, Ding, J., Boyer, P.L., Hughes, S.H., Arnold, E. (1999) Lamivudine (3TC) resistance in HIV-1 reverse transcriptase involves steric hindrance with beta-branched amino acids. *Proc Natl Acad Sci U S A* **96**: 10027-10032

Sawaya, M.R., Pelletier, H., Kumar, A., Wilson, S.H., Kraut, J. (1994) Crystal structure of rat DNA polymerase beta: evidence for a common polymerase mechanism. *Science* **264**: 1930-1935

Steitz, T.A. (1999) DNA polymerases: structural diversity and common mechanisms. *J Biol Chem* **274**: 17395-17398

Steitz, T.A., Steitz, J.A. (1993) A general two-metal-ion mechanism for catalytic RNA. *Proc Natl Acad Sci U S A* **90**: 6498-6502

Steitz, T.A., Smerdon, S.J., Jager, J., Joyce, C.M. (1994) A unified polymerase mechanism for nonhomologous DNA and RNA polymerases. *Science* **266**: 2022-2025

Vassylyev, D.G., Sekine, S., Laptenko, O., Lee, J., Vassylyeva, M.N., Borukhov, S., Yokoyama, S. (2002) Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A resolution. *Nature* **417**: 712-719

Uemori T, Ishino Y, Toh H, Asada K, Kato I. (1993) Organization and nucleotide sequence of the DNA polymerase gene from the archaeon *Pyrococcus furiosus*. *Nucleic Acids Res* **21**: 259-265

Wang, J., Sattar, A.K., Wang, C.C., Karam, J.D., Konigsberg, W.H., Steitz, T.A. (1997) Crystal structure of a pol alpha family replication DNA polymerase from bacteriophage RB69. *Cell* **89**: 1087-1099

Woese, C (1998) The universal ancestor. *Proc Natl Acad Sci U S A* **95**: 6854-6859

Zhao, Y., Jeruzalmi, D., Moarefi, I., Leighton, L., Lasken, R., Kuriyan, J. (1999) Crystal structure of an archaebacterial DNA polymerase. *Structure Fold Des* **7**: 1189-1199

Zhou, B., Pata, J.D., Steitz, T.A. (2001) Crystal structure of a DinB lesion bypass DNA polymerase catalytic fragment reveals a classic polymerase catalytic domain. *Mol. cell.* **8**: 427-437

Table 1. Shows each one of the families of cellular polymerases. It is also shown their corresponding catalytic domains for those with known tertiary structure (see text), and a summary of their phylogenetic distribution at the level of the main cellular lineages. The kinds of polymerization associated to each family of polymerases are also shown. Keys for kind of polymerization: DdDp, DNA dependent DNA polymerization; DdRp, DNA dependent RNA polymerization; RdDp, RNA dependent DNA polymerization; ndDNAp, non dependent DNA polymerization; ndRNAp, non dependent RNA polymerization; RdRp, RNA dependent RNA polymerization. *Originally, although structurally and evolutionary unrelated, the catalytic domain of rat DNA polymerase β (a nucleotidyltransferase) was named as palm domain (Pelletier, et al, 1994), in analogy to the palm domain of Klenow fragment from *Escherichia coli* DNA polymerase I. In order to avoid confusions between the palm domain of DNA polymerase I and its homologous, and the so called palm domain of the DNA polymerase β from *Rattus norvegicus* (2bpl) of different evolutionary origin, here we propose to rename this latter domain as "head" domain, as an homologous domain has been described and named as "head" domain by Li et al, (2002) in another nucleotidyltransferase, the CAA adding enzyme. In gray we remark that the only polymerase with universal distribution is the cellular transcriptase ββ' and also remark the fact that several distinct families of polymerases share homologous domains.

Figure 1. Representative sequences from each of the families of nucleic acid polymerases sharing homologous domains in their catalytic subunits. Homologous domains are colored with the same pattern. The palm domain from DNA polymerases from families I, II, telomerase reversetranscriptase, RNA dependent RNA polymerase and family Y DNA polymerase, share enough structural similarity to suspect that they descend from a common ancestral palm domain. DNA polymerases from family I, II and III share an homologous 3'-5' exonuclease domain. Motifs described by Delarue (1990) are also shown. Motifs C from DNA polymerase II from *Desulfurococcus* sp., DinB and DNA polymerase III were not originally described by Delarue et al. (1990). The length in amino acid residues is shown for each sequence and for domain boundaries inside each sequence. For Telomerase reversetranscriptase and RNA dependent RNA polymerases, numbers in cursive are for the complete polypeptide. CATH numbers are shown for each domain when available.

Figure 2. Catalytic motifs in nucleic acid polymerases. Motifs (A) and (B) enclosed in squares are as proposed by Delarue (1990) with the modifications as suggested by Wang, et al, (1997). Important residues for metal binding are colored in gray. * For Motif C it is shown in which domain is found in the protein. Note that the order of motifs in DNA primase from *Pyrococcus furiosus* (1g71) is reversed in order to show the functional convergence.

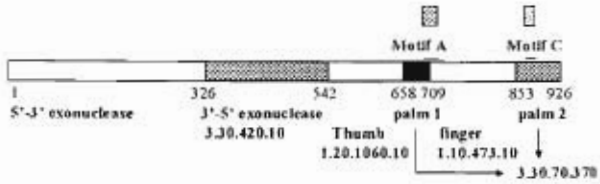
Figure 3. Domain structure for Nucleotidyltransferase family (NT) of polymerases. Motifs in DNA polymerase β from *Rattus norvegicus* and Terminal deoxynucleotidyltransferase from *Mus musculus* are as described by Delarue (1990). The corresponding motifs C in the other enzymes were found by visual inspection. Note that the only domain shared by all the members of the family is the catalytic "head" domain. The phylogenetic distribution is for the complete genomes analyzed here.

Figure 4. The early evolution of polymerase function. The first protein polymerase that arose could have had a palm domain. This RNA polymerase RNA dependent eventually gave origin to Reversetranscriptase (RT) and to DNA dependent DNA polymerase. The replicative machinery of the LCA was inherited to the Archaeal-Eucaryal lineage, while in the Bacterial lineage a non-orthologous gene displacement gave origin to DNA polymerase III.

Family	DNA polymerase I	DNA polymerase II	Telomerase, reverse transcriptase and RNA dependent RNA polymerase	Family Y of DNA polymerases	Nulceotidyl-transferases (NT)	Bacterial primases	Archaeal-Eucaryal primases	Cellular transcriptase β and β' and Eucaryotic RNA polymerase RNA dependent	DNA polymerase III	Family D of DNA polymerases
Asociated polymerase activity	DdDp DdRp	DdDp	RdDp RdRp	DdDp	DdDp ndDNAp ndRNAp	DdRp	DdRp	DdRp RdRp	DdDp	DdDp
Catalytic domain	palm domain	palm domain	palm domain	palm domain	head* domain	toprim domain	prim domain	double psi-β barrel domain	no crystal structure	no crystal structu
Other domains shared between polymerases	3'-5' exonculease domain	3'-5' exonculease domain							3'-5' exonculease domain and OB fold domain	OB fold domain i regulatory subun (DP1)
Phylogenetic distribution	Eucarya Bacteria	Eucarya Archaea Bacteria	Eucarya Bacteria Archaea	Eucarya Archaea Bacteria	Eucarya Archaea Bacteria	Bacteria	Eucarya Archaea	Eucarya Archaea Bacteria	Bacteria	Archaea

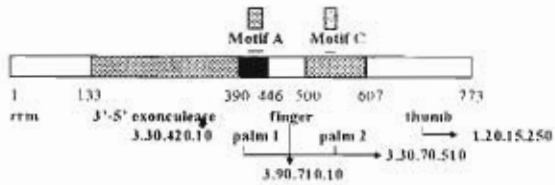
DNA polymerase I from *Escherichia coli* (1kfs)

DNA polymerase I family



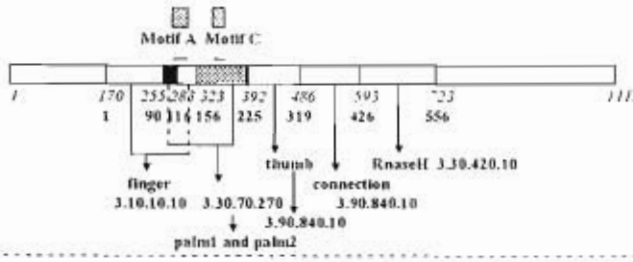
DNA polymerase II from *Desulfurococcus* sp. (1d5a)

DNA polymerase II Family



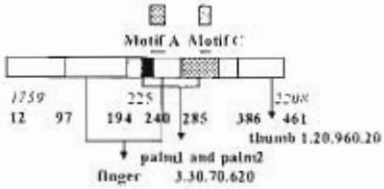
Reverse transcriptase from HIV-1 (1qe1)

Telomerase reverse transcriptases



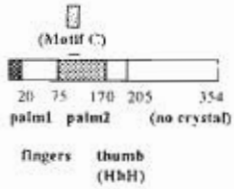
Poliovirus 3d polymerase RNA dependent RNA polymerase (1rdr)

RNA dependent RNA polymerase



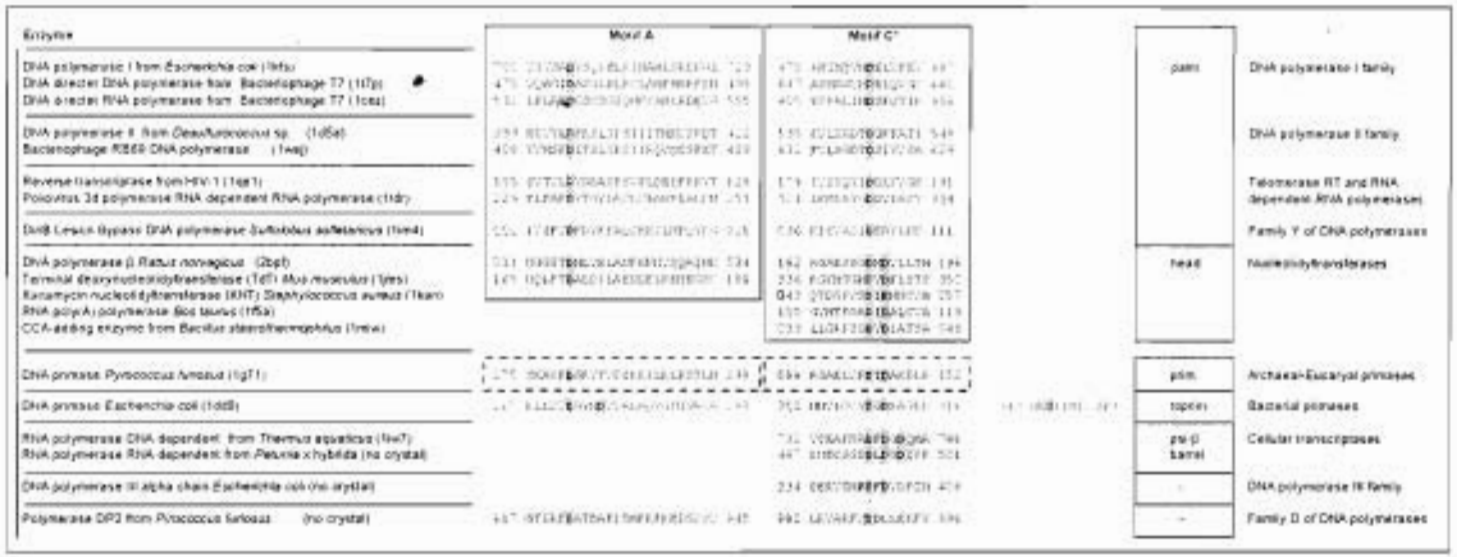
DinB Lesion Bypass DNA polymerase *Sulfolobus solfataricus* (1im4)

Family Y of DNA polymerases

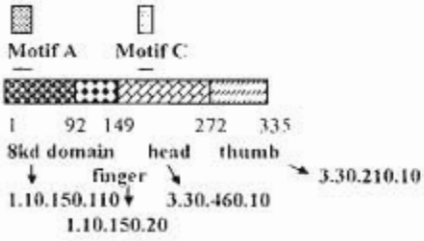


DNA polymerase III from *Ureaplasma parvum* (DPO3_UREPA)

DNA polymerase III family

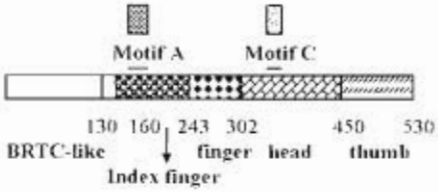

Enzyme
Motif A
Motif C*
palm
DNA polymerase I family
DNA polymerase II family
Telomerase RT and RNA dependent RNA polymerases
Family Y of DNA polymerases
head
Nucleotidyltransferases
prim
Archaeo-Eukaryal primases
topoi
Bacterial primases
pre-β barrel
Cellular transcriptases
-
DNA polymerase III family
-
Family D of DNA polymerases

DNA polymerase β *Rattus norvegicus* (class I) (2bpf)



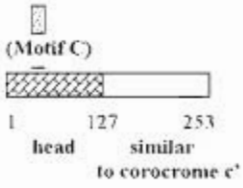
Animals, Plants, Fungi
few Bacteria
few Euryarchaea

Terminal deoxynucleotidyltransferase (TdT) *Mus musculus* (class I) (1jms)



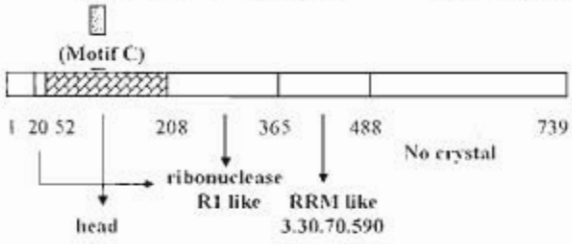
Animals, Plants, Fungi
few Bacteria
few Euryarchaea

Kanamycin nucleotidyltransferase (KNT) *Staphylococcus aureus* (class I) (1kan)



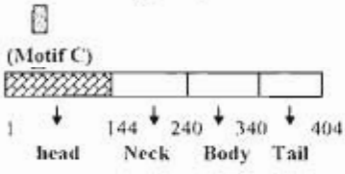
Staphylococcus sp.

RNA poly(A) polymerase *Bos taurus* (class I) (1f5a)

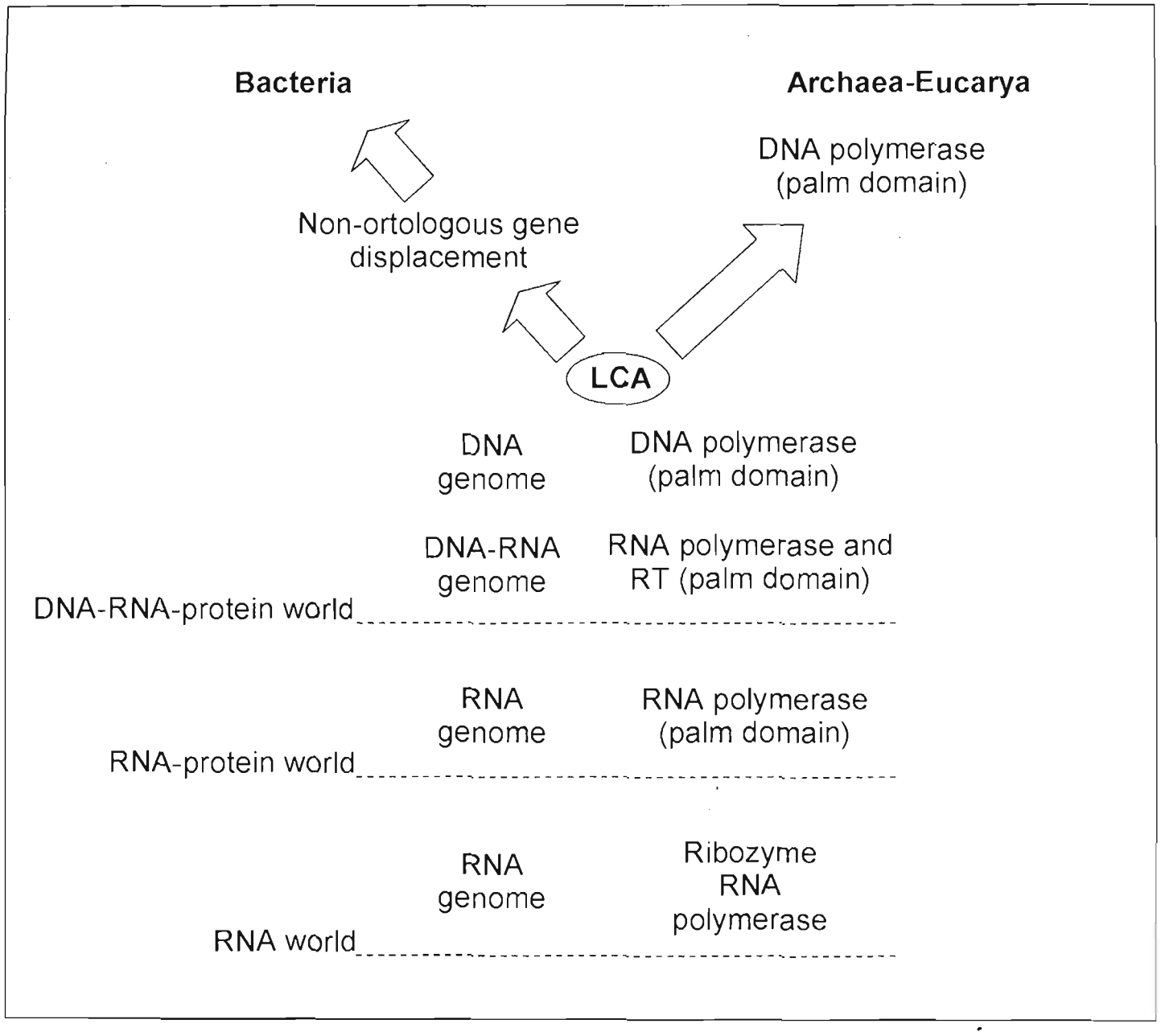


Animals, Plants, Protists, Fungi

CCA-adding enzyme from *Bacillus stearothermophilus* (class II) (1miw)



Animals, Plants, Protists, Fungi
Bacteria


Bacteria
Archaea-Eucarya
DNA polymerase
(palm domain)
Non-ortologous gene
displacement
LCA
DNA
genome
DNA polymerase
(palm domain)
DNA-RNA
genome
RNA polymerase and
RT (palm domain)
DNA-RNA-protein world
RNA
genome
RNA polymerase
(palm domain)
RNA-protein world
RNA
genome
Ribozyme
RNA
polymerase
RNA world

V. DISCUSIÓN

¿Se ha perdido el pasado biológico?

De acuerdo a Sober (1988) la posibilidad de reconstruir el pasado a partir de los datos que el presente nos ofrece, depende de si el proceso que conecta al pasado con el presente es *destructor de información* o *conservador de información*. En un proceso *destructor de información*, sin importar el estado inicial del sistema, el estado final del sistema será siempre el mismo. Un caso sencillo pero ilustrativo lo representa el sistema formado por una canica que dejamos caer en una concavidad. Inicialmente la canica oscilará de un lado a otro, pero cuando finalmente la canica descanse al fondo de la concavidad no tendremos forma de saber en que orilla del pozo inició su recorrido. Un proceso *destructor de información* es un proceso convergente. Por el contrario, un proceso *conservador de información* es aquél en el cual el estado final del sistema es altamente dependiente de las condiciones iniciales del mismo. Esto es, los estados finales del sistema tienden a ser divergentes dependiendo de variaciones en el estado inicial. Por lo tanto, en un proceso *conservador de información* es posible inferir el estado ancestral a partir de los datos actuales.

Aspectos de la evolución que conservan las huellas del pasado

La enorme diversidad de la biota sugiere que en principio, el proceso evolutivo es en esencia un proceso divergente y por lo tanto puede conservar las huellas de su pasado. Es decir, que los caminos que puede tomar la evolución son tan numerosos que las convergencias absolutas en la evolución biológica, son prácticamente inexistentes. El ojo de los pulpos y el de los humanos han convergido por la acción de la selección natural, sin embargo, son distintos en sus detalles estructurales. Esto también implica que la similitud entre dos estructuras (aunque se encuentren desempeñando funciones distintas en especies separadas, como por ejemplo las extremidades de los caballos y de las ballenas), es debida a ancestría común (Darwin, 1859).

Si estudiamos detenidamente el proceso básico de evolución al nivel molecular, observaremos que las moléculas pueden guardar una gran cantidad de información histórica (como había sido sugerido por Zuckerkandl y Pauling (1965)). La evolución molecular es básicamente un proceso de dos etapas: a) generación de variación heredable, y b) transmisión diferencial de ésta variación. Como el DNA es la molécula de la herencia, solo la variación que en ella se genere, será en última instancia trascendente para la evolución. Por otro lado, el destino de la variación heredable (su eliminación o su perpetuación en todos los miembros de la especie) estará determinado en última instancia por un juego entre las fuerzas de selección natural y deriva génica (Figura 7). Con respecto a la primer etapa del proceso evolutivo, la cantidad de variación que se puede generar es enorme, por ejemplo, haciendo un cálculo grueso, el número de posibles secuencias distintas de DNA para un gen de 900 pares de bases es astronómico 4^{900}. Claramente la evolución solo ha explorado un espacio muy pequeño del "*espacio de secuencia*" posible. ¿Cómo exploran las secuencias este espacio? La primer etapa del proceso evolutivo es la generación de variación azarosa. Ello

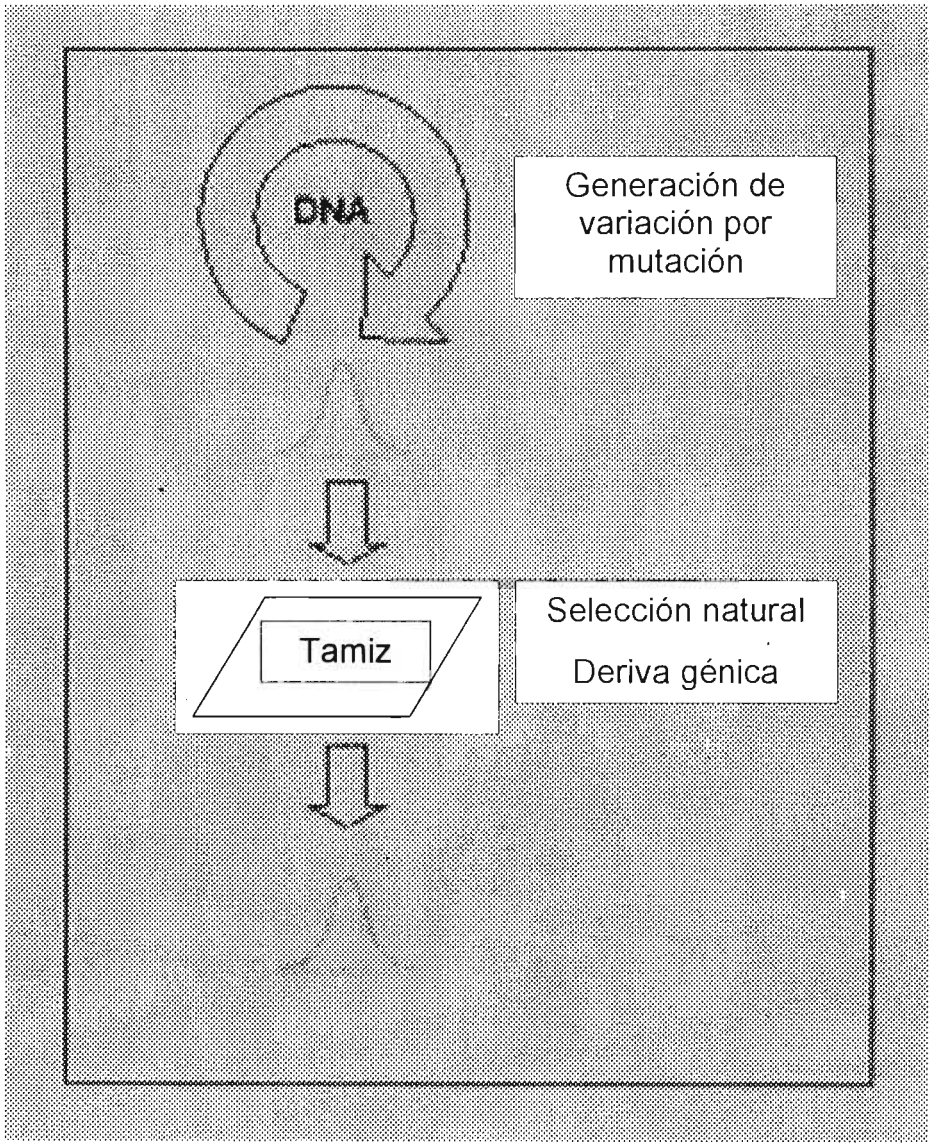
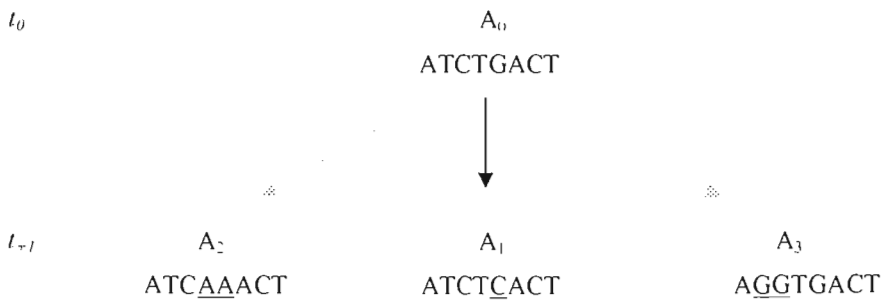


Figura 7. Evolución molecular. De forma reduccionista, la evolución se puede representar como un proceso de dos etapas: a) generación de variación y b) destino de la variación por selección natural y/o deriva génica.

significa dos cosas, por un lado significa que no podemos determinar ni en que región de la molécula de DNA, ni en que tiempo ocurrirá la siguiente mutación. Por otro lado significa que no existe la mutación dirigida (es decir, el ambiente no puede influenciar la aparición preferencial de aquellas mutaciones que redundarán en una mayor adecuación para el organismo en un ambiente determinado). En el DNA la información se codifica por un arreglo lineal de 4 símbolos distintos (A, T, C, G), el estado de la variación en el tiempo t_{+1} que se puede alcanzar por mutación (por ejemplo una mutación puntual), dependerá estrictamente del estado de la variación en el tiempo t. Como se muestra a continuación los estados A_2 y A_3 no pueden ser alcanzados con una mutación puntual a partir del estado A_0.



Esta combinación de factores (un enorme *espacio de secuencia*, generación de variación azarosa y dependencia de la variación alcanzable por mutación en el tiempo t_{+1} del tiempo t) hacen que sea muy poco probable que una vez que una determinada secuencia recorrió un camino en el espacio de secuencia, regrese al estado anterior "sobre sus propios pasos". En la segunda etapa del proceso evolutivo (el proceso de sustitución alélica), tanto la deriva génica (que fija alelos o los elimina de la población de forma azarosa) como la selección natural (que elimina alelos o los fija en la población, dependiendo de la relación entre la función y su desempeño en un ambiente determinado, el cual es variable también) actúan solamente a partir de la variación disponible en un determinado momento de la historia de la población. La combinación de los factores que vimos previamente hacen de la evolución prácticamente un proceso irrepetible.

El proceso evolutivo no borra las huellas de su propio pasado. Si encontramos dos secuencias de DNA o de aminoácidos que comparten un grado de similitud superior a lo que podríamos encontrar por azar, entonces podemos sospechar que son homólogos. En este sentido, si las convergencias estructurales de proteínas son raras o inexistentes, la similitud que existe entre los dominios "*palm*" de las distintas polimerasas, así como las topologías similares de los distintos dominios de unión a RNA sugieren ancestría común.

Por ejemplo, con respecto a la evolución temprana de la vida, la existencia del código genético universal sugiere firmemente que los seres vivos actuales somos monofiléticos. El hecho de que en la actualidad existen algunos códigos genéticos ligeramente distintos del código genético "universal" en algunas ramas terminales

del árbol universal, indica que otros códigos genéticos son posibles (es decir, que la estructura del código genético, la asignación de determinados aminoácidos con determinados codones, no es determinista, es decir que no es necesariamente el resultado de la interacción entre aminoácidos y codones debido a sus propiedades fisicoquímicas) y abre la posibilidad de que durante la evolución temprana del código pudieron haber existido distintas variantes y que los seres vivos actuales heredamos solamente una de ellas.

Aspectos de la evolución que borran las huellas del pasado

Sin embargo, como se mencionó anteriormente existen dos procesos que pueden borrar las huellas del pasado biológico. El primero de ellos es la pérdida polifilética o total de algún carácter. Por ejemplo, si existió una molécula informacional y catalítica previa al RNA, las evidencias de su pasada existencia se han perdido por completo en los seres vivos actuales, y su existencia puede ser inferida solamente a partir de conjeturas. Es decir, la existencia de dicha molécula es una necesidad hipotética. Por un lado, la biología indica que el RNA evolucionó previo al DNA y las proteínas, sin embargo los experimentos sobre química prebiótica no han encontrado un camino directo de la sopa prebiótica al mundo del RNA. Por lo tanto se hace necesaria la existencia de un mundo intermedio entre la sopa prebiótica y el mundo del RNA. Por otro lado, si un gen determinado o una ruta metabólica completa, la cual se encontrase en el último ancestro universal y se perdió en uno o más linajes, habremos perdido la evidencia directa que indique su presencia en el *cenancestro*.

El otro proceso que puede nublar la reconstrucción del pasado es la transferencia horizontal de genes entre linajes seguido de sustituciones no ortólogas. Sin embargo, hay razones para pensar que la transferencia horizontal de genes no ha sido tan extensa como para borrar toda huella del pasado. Por ejemplo, los fenogramas construidos a partir de los proteomas completos rescatan los tres dominios celulares. Por otro lado existen moléculas centrales tales como la DNA polimerasa III que claramente se encuentran en un solo linaje, el linaje bacteriano en este caso y que indican que por alguna razón la transferencia horizontal entre dominios no a abarcado a todas las clases de genes, cuando menos entre los principales dominios celulares. Este sigue siendo un tema abierto en los estudios sobre evolución molecular.

VI. CONCLUSIONES

Sobre el significado de las reconstrucciones y la naturaleza del cenancestro

Existe una serie de complicaciones no triviales relacionadas a la inferencia de la evolución temprana de la vida mediante la comparación de genes y genomas. Los procesos de pérdidas polifiléticas secundarias, las sustituciones no ortólogas y la transferencia horizontal de genes, aunado al sesgo en el muestreo, nublan nuestra capacidad de reconstruir la naturaleza del último ancestro común (Becerra, et al 1997). Además, cada grupo de genes ortólogos identificados como altamente conservados mediante la comparación de genomas celulares, se pudo haber originado en distintas etapas de la evolución celular (Doolittle, 2000; Zhaxybayeva y Gogarten, 2004), y no necesariamente todos ellos tuvieron que estar presentes en el genoma del *cenancestro*. Algunos genes pudieron haber sido transferidos horizontalmente a los linajes actuales, desde linajes que posteriormente se extinguieron. Sin embargo, si la transferencia horizontal de genes es un fenómeno más común entre especies cercanas que entre especies filogenéticamente lejanas, entonces, es probable que la transferencia horizontal no haya borrado del todo el patrón de herencia vertical. Además, distintos linajes pueden estar sujetos a distintas tasas de transferencia horizontal (Zhaxybayeva, et al 2004) y es probable que distintos genes estén sujetos a distintas tasas de transferencia horizontal, en especial los genes relacionados a procesos *informacionales* de la célula parecen estar menos sujetos a transferencia horizontal que los genes *operacionales* (Rivera y Lake, 1999).

En la literatura existen numerosos trabajos que intentan describir la naturaleza del último ancestro común utilizando metodologías distintas, pero todas basadas finalmente en la identificación de caracteres homólogos. Recientemente, Harris et al (2003) buscaron en la base de datos de genes ortólogos COG (Tatusov, et al 2001) genes ortólogos que estuviesen universalmente conservados y que además mostraran una filogenia universal en donde los tres linajes celulares estuviesen claramente diferenciados. De los casi 3100 grupos de COG's existentes en la base de datos, solamente 80 se encontraron universalmente conservados, y solamente 50 de ellos mostraron tener una filogenia compatible con la filogenia del rRNA (genes denominados *tres dominios*). De los 50 genes *tres dominios* 37 de ellos están fisicamente asociados con el ribosoma en las células actuales (proteínas ribosomales y factores de traducción y transcripción). El resto de los genes tres dominios está formado por proteínas asociadas a funciones ribosomales como: la metionina aminopeptidasa, metiltransferasas, proteína *ffh;* proteínas asociadas a la transcripción y replicación del DNA tales como: tres subunidades de la RNA polimerasa DNA dependiente, y la proteína *NusG*, la denominada *sliding clamp* (*DnaN* en *E. coli*), la exonculeasa 5'-3' de la DNA polimerasa I y la proteína de recombinación *RecA*; y dos proteínas sin función asociada (una ATPas y una GTPasa). El resto de las 30 proteínas universalmente conservadas que no muestran una filogenia tres dominios está conformado por tRNA aminoacil transferasas,

timidin cinasa, topoisomerasa 1A, fosfomanomutasa, proteasas, una subunidad de la DNA polimerasa III, entre otros.

En un análisis distinto, basado en un simple esquema de clasificación que utiliza la presencia o la ausencia de superfamilias de proteínas (definida según la base de datos SCOP), Yang, et al (2005) construyen una matriz de presencias y ausencias de familias de proteínas en 174 genomas completos, con la cual reconstruyen una filogenia universal en donde Arqueas, Bacterias y Eucariontes se muestran como grupos monofiléticos. Mediante su estudio, identifican a 49 superfamilias universalmente conservadas. Las superfamilias identificadas comprenden una variedad de funciones que van más allá de proteínas asociadas al ribosoma (Yang, et al 2005).

Claramente, las diferencia entre los resultados realizados por Harris et al (2003), Yang, et al 2005 y Delaye et al, (en prensa), se debe a las distintas metodologías utilizadas en cada caso. En el caso de Harris et al (2003) un primer filtro se impone al momento de elegir solo aquellos genes universalmente conservados identificados mediante la metodología de la base de datos COG (Tatusov, et al 2001). El siguiente filtro es filogenético, al elegir solo aquellas secuencias que reflejan la filogenia de los tres dominios. Es probable que algunos genes no sean asignados al genoma del último ancestro común debido a errores en la reconstrucción filogenética tales como atracción de ramas largas. En el caso de Yang, et al (2005) se eligen solamente familias de proteínas con estructura terciaria conocida. Es de esperarse que el número de genes asignados al LCA se incremente conforme se conozcan nuevas estructuras terciarias. Sin embargo, un estimado del número de proteínas que serán finalmente asignadas al LCA mediante esta metodología se puede obtener graficando el número de estructuras universalmente conservadas contra el tamaño de la base de datos (SCOP en este caso). El estudio de Yang et al (2005) está restringido a la fracción de familias de proteínas para las cuales se conoce su estructura terciaria. Sin embrago, es el estudio que más proteomas abarca (174) y el número final de estructuras no varia demasiado si se utiliza solo una fracción de los proteomas analizados (19 de cada linaje) o si se utilizan las topologías (en lugar de las superfamilias) para hacer el análisis (Yang et al 2005). Además, utilizar perfiles basados en estructuras terciarias de proteínas tienen la ventaja de identificar homólogos lejanos ya que los métodos de búsqueda basados en estructura primaria no son capaces de identificar homología entre secuencias extremadamente divergentes. Finalmente, la precisión del análisis realizado por Delaye, et al (en prensa), depende de la habilidad de BLAST para detectar todos los homólogos (minimizar el número de falsos negativos) y excluir los falsos positivos. El número de familias de proteínas (en este caso dominios Pfam) asignados al genoma del LCA es mayor para el análisis de Delaye, et al (en prensa) (115) que para el caso de Harris, et al (2003) (80 universalmente conservados de los cuales solo 50 presentan una filogenia de tres dominios) y el caso de Yang, et al (2005) (49). Las diferencias se puede deber a la exclusión de genomas de parásitos por parte de Delaye, et al (en prensa). Un análisis futuro podría

consistir en realizar filogenias a los genes universalmente conservados detectados por Delaye et al (en prensa) para determinar tratar de ampliar el conjunto de genes encontrados por Harris et al (2003).

Cada uno de los métodos descritos anteriormente posee ciertas ventajas y desventajas. Sin embargo, es claro que independientemente de las diferencias metodológicas, los resultados son similares entre si (entre 50 y 100 genes universalmente conservados, la mayoría relacionados a la maquinaria ribosomal, genes de transcripción, componentes accesorios de la replicación del DNA y algunos genes relacionados al metabolismo, en especial de nucleótidos y amino ácidos). Claramente, este conjunto de genes es insuficiente para mantener un sistema celular. En el mejor de los casos, este conjunto de genes representa el genoma del *cenancestro* de la misma forma que un fósil representa evidencia de la existencia de vida pasada. Al igual que en un fósil en donde solo se conservan algunas de la estructuras del organismo ancestral, no todas los genes que pudieron estar presentes en el genoma del *cenancestro* se encuentran conservados, pero es posible inferir la existencia de algunas funciones si algunas de las moléculas involucradas en la función se conservan. En especial si se trata de genes ortólogos. De los genes detectados por Harris et al (2003) como genes *tres dominios* tres de ellos están relacionados directamente con la replicación y reparación del DNA, estos son la *sliding clamp* (la cual es necesaria para la alta procesividad de las DNA polimerasas durante la replicación), exonculeasa 5'-3' de la DNA polimerasa I y la enzima de recombinación RecA. Relacionado al DNA pero de forma distinta, también pertenece a éste grupo las subunidades centrales de la transcriptasa celular. Si bien, la maquinaria de replicación central no está universalmente conservada, el hecho de que estas proteínas si lo estén, sugiere que el último ancestro común utilizaba DNA para su continuidad genética y que probablemente las DNA polimerasas no estén conservadas debido a otros factores, tales como la sustitución no homóloga, como se sugiere en esta tesis.

Carl Woese (1998) ha sugerido que el ancestro universal estaba formado por una comunidad de células diversas que sobrevivía y evolucionaba como una unidad, la cual poseía una historia física más no genealógica. Esta comunidad estaba conformada por entidades celulares (*progenotes*) con una maquinaria de procesamiento de la información muy simple y poco precisa, en donde la tasa de mutación y la transferencia horizontal de genes era muy elevada, a tal grado que la dinámica evolutiva estaba regida por la transferencia horizontal. De acuerdo a Woese (1998), conforme evolucionaron estructuras biológicas más complejas y precisas, las *temperaturas* evolutivas disminuyeron (es decir, la tasa de mutación y la transferencia horizontal), distintos sistemas celulares *cristalizaron* (probablemente el primero en hacerlo fueron los aparatos de traducción y transcripción y debido a ello se encuentran altamente conservados) y la dinámica evolutiva comenzó a parecerse a la de las células actuales. Si bien, es muy probable que los primeros sistemas replicativos fueron muy sencillos, el patrón de conservación de los genes en los genomas actuales se puede comprender también si suponemos un *cenancestro* similar en complejidad a un procarionte actual y apelamos a diferencias en tasas de evolución entre genes homólogos, pérdidas secundarias, transferencia horizontal y

sustituciones de genes, todos ellos procesos que sabemos que ocurren. Por lo tanto, a pesar de lo atractiva que pueda parecer la hipótesis de Woese (1998) no hay en principio razón para preferir su hipótesis con respecto a otras, salvo la existencia de una tasa de transferencia horizontal mucho mayor a la de los genomas actuales durante la existencia del *cenancestro*. El último ancestro común no tiene que estar necesariamente cercano cualitativamente al origen de la vida (Becerra, et al 2000; Zhaxybayeva, et al 2004).

1. Los genes universalmente conservados en 20 genomas de organismos no parásitos, sugieren la existencia previa de un mundo de RNA/proteína.
2. Debido a la universalidad del DNA como molécula de la herencia y a la conservación en los tres linajes celulares de un grupo de enzimas relacionadas a su replicación, la parsimonia sugiere que el último ancestro común poseía un genoma de DNA. Tal como parece haber ocurrido en el caso de las Euryarchaeas y las mitocondrias, las sustituciones no ortólogas han jugado un papel importante en la evolución de algunas DNA polimerasas. Sugerimos que es más parsimonioso suponer que la enzima central replicativa no está universalmente conservada por haber sufrido una sustitución no ortóloga posterior a la existencia del *cenancestro*, a sugerir que el último ancestro universal poseía un genoma de RNA y la replicación del DNA evolucionó independientemente dos veces, una en el linaje bacteriano y otra en el linaje arqueo/eucarionte.
3. Las DNA polimerasas pertenecientes a las familias I, II, RNA polimerasas virales y reverso transcriptasas comparten un dominio *palm* homólogo, el cual proviene de una etapa anterior al *cenancestro* y probablemente era responsable de replicar el material genético en el mundo del RNA/proteína.
4. Entre las proteínas más antiguas que se pueden identificar (los dominios de unión a RNA universalmente conservados), se encuentran evidencias de evolución por *patchwork*.
5. Como lo sugieren los dominios de unión a RNA universalmente conservados, la duplicación de genes seguida de divergencia parece haber jugado un papel central en la evolución de nuevas funciones desde etapas muy tempranas de la evolución de la vida en la Tierra.

VII. PERSPECTIVAS

El origen de nuevas proteínas

Dos de las preguntas centrales en evolución molecular son: cuántas familias de genes existen y ¿cuál es el mecanismo(s) mediante el(los) cual(es) se originan los nuevos genes?. Con respecto a la primer pregunta, lo que tratamos de saber es cuántos grupos monofiléticos de enzimas existen. Como veremos, la respuesta a esta pregunta depende en cierto grado del mecanismo de origen de nuevos genes.

Después de la publicación de *El Origen de las Especies* (Darwin, 1859), la similitud entre estructuras de origen biológico pasó a interpretarse como evidencia de ancestría común, es decir, de homología. De esta forma, la similitud entre las extremidades de los vertebrados, se interpretó como evidencia de que todos los vertebrados compartimos un ancestro común con extremidades. A nivel molecular, dos secuencias de aminoácidos que se parecen más de lo que esperaríamos por azar suponemos que tienen un ancestro común (por ejemplo de forma simplificada, dos proteínas que se parecen en más de 30 aminoácidos por cada 100 sospechamos que son homólogas). De igual forma, dos proteínas que comparten la misma estructura terciaria (o la misma topología, definida ésta como el mismo arreglo espacial de estructuras secundarias con la misma conectividad), en principio suponemos que comparten un ancestro común. El argumento anterior puede estar basado por ejemplo en la siguiente observación: debido a que el número de posibles secuencias de proteínas es tan grande 20^n (donde n representa el número de aminoácidos de una secuencia), la cantidad de formas distintas en las cuales se pueda plegar una secuencia de aminoácidos probablemente sea muy grande también, y por lo tanto es poco parsimonioso suponer que pueda existir convergencia a este nivel (aunque no todas las secuencias posibles de aminoácidos se puedan plegar en una proteína globular). De esta forma, casos en los cuales dos proteínas comparten la misma topología a pesar de compartir una similitud en aminoácidos menor a la que esperaríamos por azar, suponemos que son homólogas. Además la estructura primaria parece divergir más rápidamente que la estructura terciaria. Las hipótesis planteadas en este trabajo, (la homología del domino *palm* entre las distintas familias de polimerasas y el modelo de evolución en *patchwork* para los dominios de unión a RNA) están basados en la suposición de que dos proteínas que comparten la misma estructura terciaria comparten un ancestro común.

Por otro lado, con la acumulación de diversos cristales de proteínas, comenzó a ser evidente que distintas regiones de una proteína se pueden plegar de formas muy diferentes en el espacio tridimensional (una proteína puede estar formada por la fusión de dos o más dominios distintos cada uno con una topología diferente) y es común encontrar que dos proteínas distintas son homólogas solamente en uno de los dominios. Con ello llegó la hipótesis de que el dominio es la unidad estructural y evolutiva de las proteínas. De esta forma, parte de la evolución de proteínas se explica mediante el modelo de "Lego" (Henikoff, 1996). Es decir, una nueva proteína, se puede formar mediante la fusión, o duplicación de dominios preexistentes (las

unidades básicas). Si los dominios son en realidad las unidades estructurales y evolutivas de las proteínas, entonces para trazar el origen de un grupo de proteínas homólogas debemos trazar el origen de sus dominios.

Entonces, la pregunta de ¿cuántas familias de proteínas existen?, se puede replantear como, ¿cuántas familias de dominios existen?. En un artículo ya clásico, Chothia (1992) estimó que deberían existir alrededor de 1000 topologías distintas. Un número extremadamente pequeño de formas, si pensamos que por ejemplo una bacteria como *Escherichia coli* tiene alrededor de 4000 genes y que la evolución biológica ha estado experimentando con la evolución de las proteínas por al menos 3,500 millones de años. En la actualidad, la base de datos SCOP (Structural Classification of Proteins versión 1.65) tiene clasificados alrededor de 800 topologías diferentes clasificados en unas 1,294 superfamilias.

Existen una serie de preguntas fundamentales con respecto a estas unidades estructurales y evolutivas de las proteínas que son los dominios: ¿Cómo y cuando se originaron estos dominios en la naturaleza? ¿Es posible la evolución de una topología determinada a otra distinta mediante mutaciones puntuales? ¿O la evolución de estos dominios estuvo precedida por la recombinación de elementos estructurales más pequeños, tal y como en la actualidad los dominios actuales se combinan para formar las proteínas multidominio? ¿Realmente todas las proteínas que comparten una misma topología son homólogas o es probable la convergencia a este nivel? Con respecto a esta última pregunta las opiniones son encontradas. Recientemente Denton y Marshall (2001) han sugerido que el número posible de topologías de proteínas es relativamente pequeño debido a las "reglas de construcción" de las proteínas. De acuerdo a ellos, el cálculo realizado por Chothia (1992) apoya esta visión.

Si bien, es posible realizar una serie de argumentos que contradicen la interpretación de Denton y Marshall (2001) la forma ideal de abordar la pregunta de "que tan común es la convergencia de topologías" o dicho de otra forma "que tan grande es el espacio de plegamiento" es realizando un experimento, el cual describiremos a continuación.

La evolución de proteínas ocurre principalmente debido a la duplicación genética seguida de divergencia y/o recombinación. De esta forma se generan las familias de proteínas. En principio, aproximadamente cada topología que existe en la naturaleza y que define una familia de secuencias homólogas, se debió haber inventado por lo menos una vez en la naturaleza (esto solo si la hipótesis que sugiere que todos los genes actuales provienen de un solo gen es falsa). Lo que necesitamos hacer es encontrar casos en donde un determinado gen se haya originado *de novo*. Estos genes formados *de novo* representan nuestro experimento evolutivo. Podemos comparar las topologías de las proteínas originadas *de novo*, con el resto de las topologías de proteínas. Si encontramos que reiteradamente los nuevos genes codifican para proteínas con topologías que ya han sido "descubiertas" por otros genes (con un origen independiente), entonces el numero posible de

topologías será más bien pequeño y la convergencia a nivel estructural será común. Por el contrario. si cada gen nuevo que identifiquemos, codifica para una topología nueva, esto será una indicación de que el número de topologías que pueden existir será con seguridad más grande que lo que se a sugerido.

Parte del éxito de este experimento tendrá que ver con la correcta identificación de los genes nuevos. Afortunadamente, la naturaleza parece habernos brindado una oportunidad para poder realizar este experimento. Se ha encontrado que un segmento de DNA puede codificar para más de un gen mediante el uso de distintos marcos de lectura. Este fenómeno parece ser común en virus, y probablemente ocurra en bacterias. El caso canónico es el del bacteriofago de cadena sencilla de DNA ϕX174. Varios genes sobrelapados se pueden observar en este caso, por ejemplo el gen B está completamente contenido en el gen A (Figura, 8). ¿Cuál es el origen de estos genes sobrelapados? Si bien es probable que ambos genes hayan existido previamente y que una mutación que eliminó la señal de termino de uno de los genes (el gen A por ejemplo) lo cual hizo recorrer el marco de lectura hasta la siguiente señal de término (que pudo estar después de que terminara el gen B). También es probable que el gen B se haya originado *de novo* a partir de una mutación que iniciara un marco de lectura junto con un sitio de iniciación de la transcripción al interior del gen A y en un marco de lectura distinto (Li. 1997). De cualquier forma, el nuevo polipéptido que se ha generado, representa un caso *de novo*, y si codifica para una proteína globular, se puede estudiar su topología. El reto, será identificar todos los genes codificados en los genomas secuenciados que contengan proteínas sobrlapadas y buscar si dichos genes tienen un homólogo cristalizado en la base de datos de estructuras terciarias PDB (Protein Data Bank Brokhaven). Para después hacer la comparación con las bases de datos de genes originados independientemente. Si este experimento tiene resultado, ayudará a comprender el origen de las formas en las proteínas y la importancia relativa de la selección natural en conjunto con las leyes de plegamiento que determinan las formas biológicas cuando menos en el ámbito de lo muy pequeño.

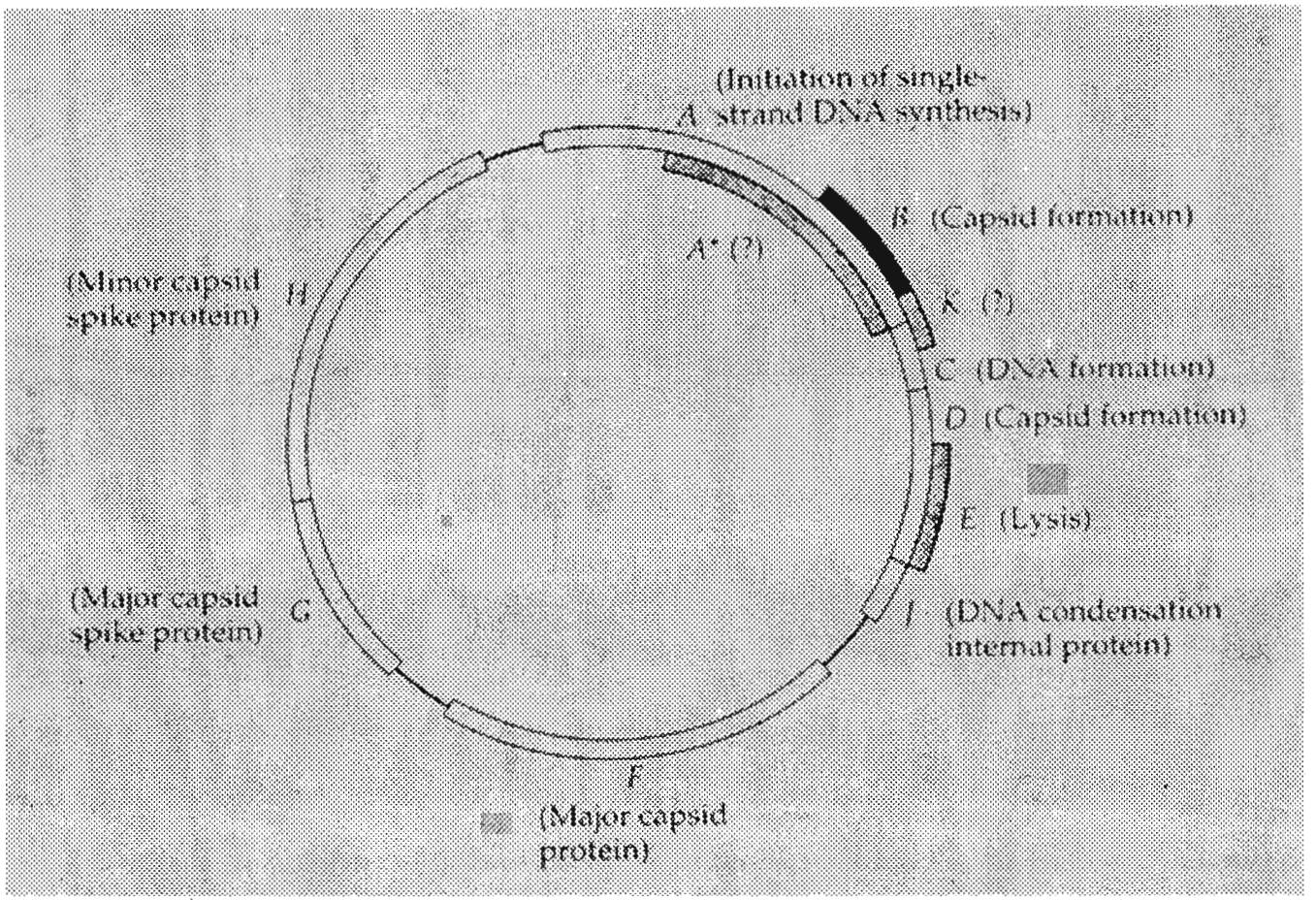


Figura 8. Genes sobrelapados. El bacteriofago de cadena sencilla de DNA ϕX174 contiene varios genes sobrelapados, obsérvese el gen B, el cual está totalmente incluido dentro del gen A (Tomado de Li, 1997).

VIII. REFERENCIAS

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

Becerra, A., Silva, E., Lloret, L., Islas, S., Velasco, AM. And Lazcano, A. (2000) Molecular Biology and the reconstruction of microbial phylogenies: des liaisons dangereuses? *In* J.Chela-Flores, G. Lemerchand, and J. Oró (eds.) *Astrobiology: Origins from the Big-Bang to Civilisation Proceedings of the First Ibero-American School of Astrobiology* (Kluwer Academic Publishers). pp. 135-151

Brown, JR., Douady, CJ., Italia, MJ., Marshall, WE., Stanhope, MJ. (2001) Universal trees based on large combined protein sequence data sets. *Nat Genet.* **28**: 281-285

Darwin, C., The Origin of the Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (Murria, London, 1859)

Daubin, V., Moran, NA., Ochman, H. (2003) Phylogenetics and the Cohesión of Bacterial Genomes. *Science* **301**: 829-832

Delaye, L., Becerra, A., and Lazcano, A. (2005) The Last Common Ancestor: what's in a name? *Origin of Life and Evolution of the Biosphere* (en prensa)

Denton, M. and Marshall, C. (2001) Laws of form revisited. *Nature* **410**: 417

Chatton, E. (1938) *Titres et Travoux Scientifiques (1906-1937) de Edouard Chatton* (E. Sótano, Sète, France).

Chothia, C. (1992) One thousand families for the molecular biologist. *Nature* **357**, 543-544

Doolittle, WF. (1999) Phylogenetic classification and the universal tree. *Science* **284**: 2124-2129

Doolittle, WF. (2000) The nature of the universal ancestor and the evolution of the proteome. *Curr Opin Struct Biol.* **10**: 355-358

Fitch, WM., Upper, K. (1987) The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harb Symp Quant Biol.* **52**:759-767

Fitz-Gibbon, ST., House, CH. (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* **27**: 4218-4222

Freeland, S.J., Knight, R.D. and Landweber, L.F. (1999) Do proteins predate DNA? *Sciencie.* **286**: 690-692

Forterre, P., Philippe, H. (1999) The last universal common ancestor (LUCA), simple or complex? *Biol Bull.* **196**:373-5

Gilbert, W. (1986) The RNA world. *Nature.* **319**, 618

Gogarten, JP., Kibak, H., Dittrich, P., Taiz, L., Bowman, EJ., Bowman, BJ., Manolson, MF., Poole, RJ., Date, T., Oshima, T., et al. (1989) Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc Natl Acad Sci U S A.* **86**:6661-6665

Haldane, J.B.S. (1929) The origin of life. The Rationalist Annual. Reeditado el 1967 en: Bernal, J.D., The Origin of Life. The Weidenfeld and Nicolson Natural History. R. Carrington. ed. London: Readers Union. p.p. 242-249

Harris, JK., Kelley, ST., Spiegelman, GB., and Pace, NR. (2003) The genetic core of the universal ancestor. *Genome Res.* **13**. 401-412

Henikoff, S., Greene, EA., Pietrokovski, S., Bork, P., Attwood, TK., Hood, L. (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science*. **278**:609-614

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., Miyata, T. (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci U S A*. **86**:9355-9359

Joyce, GF. (1989) RNA evolution and the origins of life. *Nature*. **338**: 217-224

Joyce, GF. (2002) The antiquity of RNA-based evolution. *Nature*. **418**: 214-221

Kronberg, A. and Baker, TA. (1992) *DNA replication*, second edition. W.H. Freeman and Company. USA

Lazcano, A., Fox, GE., Oró, J. (1992) Life before DNA: the origin and evolution of early Archaean cells. *In* Mortlock, R. P. (ed.), The Evolution of Metabolic Function. pp. 237-295. CRC Press. Boca Raton. FL.

Lazcano, A., Guerrero, R., Margulis, L., Oro, J. (1988) The evolutionary transition from RNA to DNA in early cells. *J Mol Evol*. **27**:283-290

Lazcano, A. & Miller SL. (1999) On the origin of metabolic pathways. *J Mol Evol*. **49**: 424-431

Lecompte, O., Ripp, R., Thierry, JC., Moras, D., Poch, O. (2002) Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res*. **30**:5382-5390

Leipe, D.D., Aravind, L., Koonin, E.V. (1999) Did DNA replication evolve twice independently? *Nucleic Acids Res 27*:3389-3401

Li, Wen-Hsiung. *Molecular Evolution*. (Sinauer Associates. USA. 1997)

Miller, SL. (1953) A production of amino acids under possible primitive earth conditions. *Science* **117**: 528-529

Oparin, A.I. (1936) The Origin of Life.

Rivera, JR., and Lake, JA. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A*. **96**: 3801-3806

Rokas, A., Williams, BL., King, N., and Carroll, SB. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*. **425**: 798-803

Snel, B., Bork, P., Huynen, MA. (1999) Genome phylogeny based on gene content. *Nat Genet*. **21**: 108-110

Sober, Elliott (1988) *Reconstructing the Past, Parsimony, Evolution, and Inference*. MIT Press. London England

Tatusov, RL., Natale, DA., Garkavstev, IV., Tatusova, TA., Shankavaram, UT., Rao, BS., Kiryutin, B., Galperin, MY., Fedorova, ND., Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**: 22-28

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

Woese, CR., & Fox, GE. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA*. **74**: 5088-5090

Woese, CR., Kandler, O., Wheelis, ML. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.

Woese, CR. (1998) The universal ancestor. *Proc Natl Acad Sci USA*. **95**: 6854 - 6859

Yang, S., Doolittle, RF. And Bourne, P. (2005) Phylogeny determined by protein domain content. *Proc Natl Acad Sci USA*. **102**: 373-378

Zhaxybayeva, O., Gogarten, JP. (2004) Cladogenesis, coalescence and the evolution of the three domains of life. *Trends Genet*. **20**: 182-187

Zhaxybayeva, O., Lapierre, P., Gogarten, JP. (2004) Genome mosaicism and organismal lineages. *Trends Genet*. **20**: 254-260

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