UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

POSGRADO EN CIENCIAS BIOLOGICAS

FACULTAD DE CIENCIAS

"ANALISIS COMPARATIVO DEL TAMAÑO Y CONTENIDO DE LOS GENOMAS PROCARIONTES"

TESIS

QUE PARA OBTENER EL GRADO ACADEMICO DE

MAESTRA EN CIENCIAS

(BIOLOGIA)

PRESENTA

SARA ERNESTINA ISLAS GRACIANO

DIRECTOR: DR. ANTONIO EUSEBIO LAZCANO-ARAUJO REYES

MEXICO, D. F. AGOSTO 2003

Agradecimientos:

A la memoria de mi papá quién permanecerá en mí para siempre.

A mi mamá por su tranquilidad y ternura.

Para Eduardo y Diego Kazuo los amores de mi vida.

Agradezco de manera especial al Dr. Antonio Lazcano por su comprensión y ayuda en mi formación.
Gracias Toño por tu amistad.

A los miembros del jurado Dr. Antonio Lazcano, Dr. Víctor Valdes, Dra. Valeria Souza, Dr. Arturo Becerra y Dra. Alicia Negrón, por sus valiosos comentarios y suguerencias

A mis compañeros de Laboratorio de Microbiología Ana, Arturo, Luis, Erwin por su apoyo y gran sentido del humor.

A mis coautores y amigos que están por otros caminos Josetxu, Amanda y Héctor.

A las nuevas generaciones los tallarines y colateds

A Pilar y Olga por el apoyo bibliográfico.

A mis amigos.

Indice

Análisis comparativo del tamaño y contenido de los genomas procariontes

Resumen

Análisis comparativo del tamaño y contenido de los genomas procariontes

Los procariontes presentan una variación considerable en el tamaño de genoma, debida por una parte a su capacidad para modificar el contenido de DNA mediante transporte horizontal, slippage, duplicación de genes y genoma completo, así como a re-arreglos propios del genoma. Aunque no sabemos como eran los primeros organismos, es probable que su maquinaria genética fuera relativamente pequeña y sus capacidades codificantes fueran restringidas. Así, explicar los mecanismos por medio de los cuales se ha incrementado el tamaño del genoma no es una tarea fácil.
El cálculo con mayor precisión del contenido de DNA y la construcción de mapas genómicos ha sido refinado mediante la técnica de electroforesis de campo pulsado (PFGE), utilizada desde 1985 para estimar el tamaño de genoma. Con el propósito de estudiar los factores que intervienen en las variaciones del tamaño de los genomas procariontes, incluyendo la hipótesis que sugiere que estos son el resultado de rondas de duplicación del genoma completo (Ohono 1970; Wallace y Morowitz 1973; Zipkas y Riley 1975; Sparrow y Neuman 1976 Herdman 1985;), en este trabajo reportamos el resultado de un análisis estadístico de la distribución de 641 tamaños de genomas tanto de bacterias como de arqueas cuyas dimensiones han sido calculadas mediante la técnica de PFGE.
Se analizó una base de datos de 641 organismos procariontes construida de reportes publicados en NCBI, Scirus, Highwire, y fue completada con datos de posición filogenética, estilo de vida, temperatura, y metabolismo.
Con los datos disponibles hasta febrero de 2003, encontramos que el rango de tamaño de genomas procariontes es de 0.448 Mb (γproteobacteria) a 9.7Mb (αproteobacteria). Los organismos con genomas más pequeños son simbiontes y parásitos obligados pertenecientes a los grupos γproteobacteria, Anaeroplasma, Spiroplasma, Rickettsia y Spirochaetae. No todos los organismos con tamaño pequeño son anaerobios lo que puede ser explicado a través de una serie compleja de adaptaciones secundarias que han guiado a la reducción de su genoma. Se encontró sin embargo, que en general que los procariontes anaerobios y microaerofílicos están dotados con genomas más pequeños que los aerobios. No obstante, los genomas más pequeños no son por su propio tamaño una muestra que nos lleve a pensar que son formas ancestrales; igualmente, el rango relativamente pequeño del tamaño de genoma de los hipertrmófilos puede revelar una tendencia a que ambientes con altas temperaturas constriñen el contenido de DNA a un rango específico (0.5 Mb-5.10 Mb), probablemente por la reducción del tamaño promedio de sus genes. Los genomas más grandes son típicamente organismos aerobios de vida libre y con ciclos de vida complejos. Aunque esta base de datos claramente presenta un sesgo (organismos disponibles en WWW) y no representa toda la diversidad procarionte, en la distribución de los tamaños de genoma de la muestra no hay evidencias que nos permitan ratificar la hipótesis de Herdman (1985) es decir, los resultados sugieren que el contenido de DNA de los procariontes no proviene de duplicaciones totales del genoma.

Abstract

Comparative size analyses and DNA content of prokariotic genome.

There is a considerable variation in the prokariotic genome size this variation is a result of their ability to modify the DNA content by different means like horizontal transfer, slippage, gene duplication, whole genome duplication and arrangements of the genome itself. Although it is still unknown how does the first cells were, is probable that they were endowed with relatively small genetic machinery with reduced encoding capacities. Thus, to explain the mechanisms through which the genome size has been increased is not an easy task.

Pulsed field gel electrophoresis is the best technique to construct a genetic map and to estimate with accuracy the determination of DNA content. This technique has been used since 1985.

The purpose of this work was to analyze the factors involved in the variations of prokaryotic genome size including the hypothesis which suggests that these are the result of genome duplication (Ohno 1970; Wallace and Morowitz 1973; Zipkas and Riley 1975; Sparrow and Neuman 1976 Herdman 1985;), in this work we report a statistical analysis of a sample of 641 archaeal and bacterial genome sizes determined by pulsed-field gel electrophoresis (PFGE), reported in publications included in the NCBI/PubMed, Scirus, Highwire databases until February 2003. and was completed with phylogenetic data, life style , temperature, and metabolism.

In our sample the prokaryotic genome size rank was 0.448 (γ proteobacteria) to 9.7Mb (αproteobacteria). The organisms with the smallest genome size are obligated simbionts and parasites belong to γ proteobacteria, Anaeroplasma, Spiroplasma, Rickettsia and Spirochaetae groups.

Not all organisms with small genome are anaerobic this feature can be explained through a complex series of secondary adaptations that have leaded to reduce their genomes.

In general the anaerobic and microaerofilic procaryotes are endowed with smaller genomes than aerobic. Nevertheless the smallest genomes are not for its own size a sample that they are ancient forms; likewise, the relatively small rank of the hiperthermophilic genome size can reveal a tendency which shows that environments with high temperatures confine the few DNA content to a specific rank (0.5 Mb-5.10 Mb), probably by the reduction average of its genes. The largest genomes are typically aerobic organisms, free life and with complex life cycles.

The database analyzed here is biased (only available organisms on www) and does not reflect all the actual prokaryotic biodiversity, in our distribution there are no evidences which support the Herdman´s hipothesis, the results suggests that procaryotic DNA content does not outcome from the whole genome duplications

Análisis comparativo del tamaño y contenido de los genomas procariontes

Introducción.

Un genoma celular se puede definir como el contenido total de la información genética (DNA) utilizada por un organismo para mantenerse y reproducirse (Kolsto 1999). A lo largo del tiempo los seres vivos han sufrido modificaciones tanto en el contenido como en las dimensiones en su genoma. Así, aunque no sabemos como eran los primeros organismos, es probable que su maquinaria genética fuera relativamente pequeña y sus capacidades codificantes fueran restringidas.

El cálculo del contenido de DNA ha sido una tarea que se ha enfrentado desde hace varias décadas, utilizando diferentes técnicas como la colorimetría, la cinética de renaturalización del DNA, la electroforesis de campo pulsado (PFGE) (Cantor 1988) y, más recientemente, la secuenciación completa de genomas. Con el propósito de estudiar los factores que intervienen en las variaciones del tamaño de los genomas procariontes, incluyendo la hipótesis que sugiere que estos son el resultado de rondas de duplicación del genoma completo (Ohono 1970; Wallace y Morowitz 1973; Zipkas y Riley 1975; Sparrow y Neuman 1976 Herdman 1985; Trevors 1996), en este trabajo reportamos el resultado de un análisis estadístico de la distribución de tamaños de genomas tanto de bacterias como de arqueas cuyas dimensiones han sido calculadas mediante la técnica de PFGE.

Cuando intentamos reconstruir fases tempranas de la evolución de los seres vivos encontramos una gran cantidad de interrogantes que marcan las diferentes etapas de cambio, que arrancan desde los procesos de evolución prebiótica que llevaron a las primeras formas de vida, que muy probablemente estaban basadas en un polímero genético distinto al RNA mismo, hasta la aparición de formas celulares con genomas de DNA pasando por una forma intermedia en la que los genomas pueden haber estado formados por RNA. Los procariontes son los seres vivos más antiguos en la Tierra, con un registro fósil que data de hace 3.5 billones de

años (Schopf, 1993; Brasier *et al,*, 2002). Aunque el registro paleontológico no permite establecer con precisión ni como eran los primeros seres vivos ni el tipo de genoma que tuvieron las primeras células, se acepta que la atmósfera primitiva carecía de oxígeno libre y pudo, de hecho, haber sido reductora. Ello implica que los primeros seres vivos eran anaerobios y heterótrofos (Oparin 1938). Sus descendientes, en cambio, se fueron adaptando a un ambiente con una creciente cantidad de oxígeno liberado en la atmósfera, lo cual seleccionó nuevas capacidades metabólicas cuya presencia se refleja, al menos en parte, en las variaciones en el tamaño de los genomas procariontes.

Los genomas procariontes poseen, por un lado una capacidad mucho mayor que los eucariontes para adquirir genes y porciones de DNA mediante el transporte horizontal y, por otra, una estabilidad relativa que les confiere una identidad específica. Así, explicar los mecanismos mediante los cuales se ha incrementado el tamaño del genoma no es una tarea fácil. Se podría pensar que "organismos más complejos" (pluricelulares) requieren de más genes, es decir de una mayor cantidad de DNA (Petrov 2001). Sin embargo, existen algunas amibas que tienen 200 veces mas DNA que los humanos (pero no necesariamente tienen más genes). A diferencia de los eucariontes, el genoma en procariontes se traduce casi directamente a funciones bioquímicas, fisiológicas y complejidad organísmica, por que la mayoría de las secuencias procariontes corresponden a regiones codificantes, es decir, son proteínas o RNAs funcionales. Es decir, en los procariontes existe una correlación directa positiva entre el número de genes y el tamaño del genoma (Mira *et al,*, 2001). Así, se puede concluir que los genomas procariontes de mayor tamaño codifican para más proteínas, secuencias reguladoras, mecanismos de reparación, diversidad de rutas metabólicas y ciclos de vida complejos. En general, se puede decir también que los organismos con replicones* más grandes poseen, una tendencia metabólica "generalista", es decir,

* El término cromosoma fué acuñado para designar el aspecto adquirido por el material genético teñido en células eucariontes. Por analogía con los eucariontes, la molecula de DNA circular o linear de procariontes se denomina cromosoma, pudiendose designar igualmente como replicones pués este término hace referencia a la estructura de ácido nucleico con capacidad de

poseen capacidades metabólicas amplias y menos requerimientos por compuestos específicos en su medio de cultivo (Shimkets 1997). En cambio, los genomas más pequeños tienden a ser de organismos altamente especializados que ocupan nichos restringidos como aquellos procariontes parásitos que viven en hospederos bajo condiciones muy particulares (e.g. los micoplasmas, las rickettsias, etc). Sin embargo, esta correspondencia no es absoluta y determinante, dado que la distribución de tamaños se traslapa ampliamente entre estos dos niveles. Se ha propuesto que en el pasado remoto los genomas deben haber sido pequeños, codificando para enzimas poco específicas, proporcionando a dichas células, máxima flexibilidad bioquímica con un mínimo contenido de genes (Jensen 1976).

Existe una variación considerable entre los tamaños de genomas procariontes , que pueden ir desde los más pequeños con 580,000pb como *Mycoplasma genitalium* (Fraser *et al*, 1995) hasta los de *Stigmatiella erecta,* con 9,550,000 pb (Neumann *et al,* 1992). Algo similar ocurre con la geometría* de sus genomas. Hasta hace poco tiempo se pensaba que los procariontes poseían solo replicones circulares, pero algunas especies presentan cromosomas lineares como el genoma de *Borrelia burdogferi* (Casjens 1993), *Streptomyces lividans* (Lin *et al,* 1993), *Rhodococcus fasciens* (Bendich y Drlica, 2000), y *Azospirillum* (Martin-Didonet *et al,* 2000), pudiendo coexistir las dos formas en algunos organismos como *Agrobacterium*, *Azospirillum* y *Streptomyces*. La presencia de genomas lineares en grupos filogenéticamente muy separados sugiere que estos se han generado varias veces de manera independiente. Todos ellos pueden poseer elementos extracromosómicos o plásmidos, que a su vez pueden ser lineares o circulares, y que no son esenciales para la sobrevivencia del microorganismo. Los plásmidos suelen codificar para funciones "específicas ", que le permiten a la bacteria o arquea adaptarse a ambientes adversos. Tales funciones incluyen, por

---

autoduplicación. Por tanto son replicones los cromosomas de las células eucariontes, procariontes, los plásmidos y los ácidos nucleícos de los virus. Igualmente el término genoma y cromosoma procarionte se ha llegado a utilizar indistintamente.

* La geometría del genoma procarionte se refiere a la forma en que se presenta el DNA procarionte y puede ser circular o linear.

ejemplo, resistencia a los antibióticos, fertilidad (propician conjugación y transferencia de material genético), virulencia, degradación de sustancias, y fijación del nitrógeno. El tamaño de los plásmidos varía de 2 kb (2 genes aprox), a 600 kb y hasta 1600 kb como en *Rhizobium*, al que se le conoce un megaplásmido o cromosoma auxiliar. Las proteobacterias conforman una cohorte filogenética dividida en diferentes grupos, de los cuales las β y γ albergan una gran cantidad de plásmidos (Moreno 1998). El tamaño no es un rasgo único para diferenciar un plásmido de un cromosoma, ya que para distinguirlos se requiere que el plásmido posea genes esenciales, tamaño suficiente y control de replicación. Así, Ng *et al,* (2000) reportaron la secuencia completa de DNA del plásmido circular (pNRC100) de la arquea *Halobacterium*, y encontraron que contiene 191,346 pb (aprox. 186 genes) que son considerados genes esenciales, además de un gen para la replicación. Ello muestra lo difícil que es precisar la distinción entre plásmidos y cromosomas .

Factores que modifican la cantidad de DNA

El incremento en el contenido de DNA se produce principalmente por la transferencia horizontal, la endosimbiosis (en el caso de los eucariontes), slippage, y eventos de duplicación de genes y genomas (fig.1). A continuación se describen estos procesos.

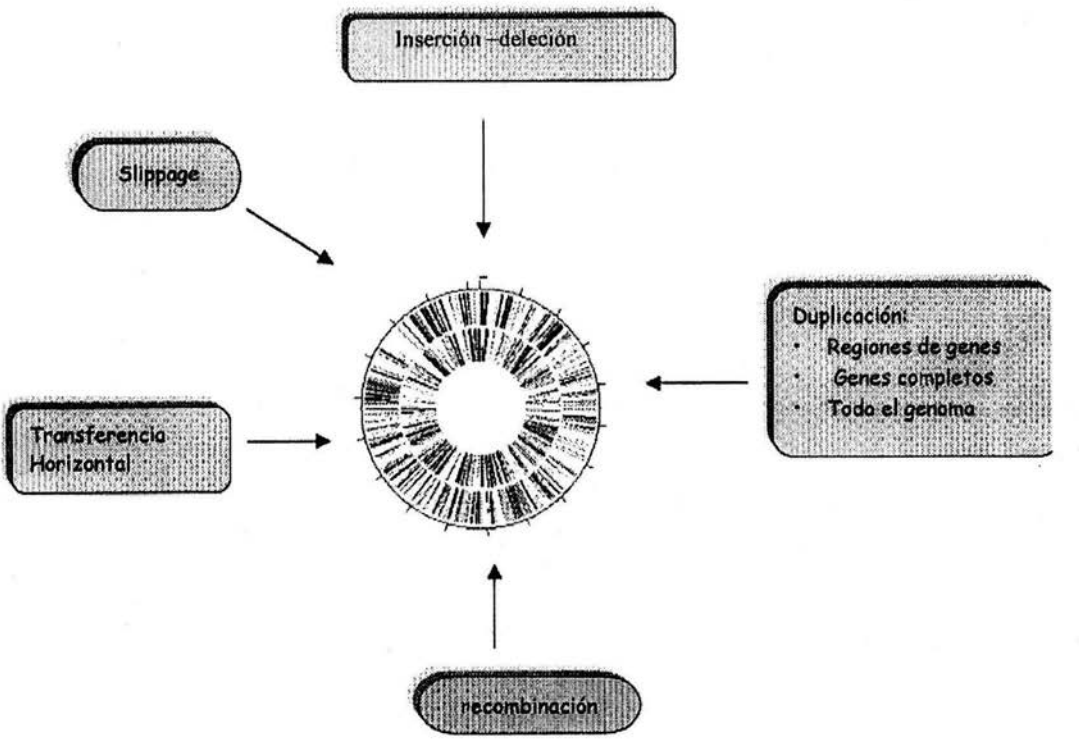


Fig 1. Factores que modifican la cantidad de DNA en procariontes.

Transferencia horizontal

Es el nombre que describe los diferentes procesos por medio de los cuales un organismo típicamente procarionte transfiere una parte de su material genético a otro organismo que puede o no ser de su misma especie (Eisen, 2000; Jain *et al,* 2002). En este proceso existe un componente temporal y espacial de los organismos para la adquisición y fijación del material transferido. Se ha establecido que no todos los genes tienen la misma probabilidad de ser transferidos. Por ejemplo, aquellos que codifican para el RNA ribosomal no son transferidos frecuentemente a otras especies mientras que los genes con mayor posibilidad de ser transferidos son los genes de mantenimiento del organismo. (Jain *et al,* 1999).

La comparación entre genomas completamente secuenciados con respecto a la composición de nucleotidos, análisis de uso de codones, y distribución filogenética basados en familias de genes son procedimientos que proporcionan evidencias de la transferencia horizontal de genes entre los dominios Arquea, Bacteria y Eucaria y pueden ser la base de adaptaciones bioquímicas y ambientales (Roy 1999) como en los casos de *Aquifex aeolicus* (bacteria hipertermofílica) y la arquea *Methanococcus jannaschiii* (Aravind *et al,* 1998). También se ha reportado transferencia de genes de Archaea a *Aquifex aeolicus* y *Thermotoga marítima.* (Nelson *et al,* 1999)

Endosimbiosis

Muchos procariontes muestran una tendencia para establecer asociaciones con células eucariontes, lo que conduce al establecimiento de diversos tipos de interacciones entre ellos, incluyendo la endosimbiosis que ha jugado un importante papel en la evolución (Margulis 1993). Las mitocondrias y los cloroplastos son vestigios de procariontes de vida libre y exhiben fuerte erosión genética durante su evolución como un resultado de la pérdida de genes innecesarios así como de transferencia de genes al núcleo. No obstante, no hay ejemplos hasta ahora conocidos de esta relación entre procariontes .

Patinaje de la polimerasa (Slippage)

Es una mutación que ocurre durante la replicación del DNA (en donde un mal apareamiento de las hebras genera un incremento o pérdida de material genético en un segundo evento de replicación, que tiene la peculiaridad de producir regiones de nucleótidos y por ende de proteínas con un sesgo en la composición del conteniendo de segmentos repetidos, mejor conocidos como secuencias de baja complejidad. Estas secuencias proporcionan una fuente de variabilidad fenotípica y genética en la evolución del tamaño de los genomas (Tautz *et al,* 1986; Hancock 1995; Becerra *et al,* en prep).

Procesos de duplicación

El significado evolutivo de la duplicación de genes fue reconocido desde hace mucho tiempo por Haldane (1932) quien sugirió que las copias del material genético extra (redundante) a través de sucesivas mutaciones pueden alcanzar funciones nuevas (Ohno 1970; Li 1980).

La duplicación de genes es el mecanismo más importante para la generación de nuevos genes durante la evolución del genoma y este mecanismo puede operar a diferentes niveles como se ve en la tabla 1

Tabla 1. Tipos de duplicación

Región de un gen	duplicación interna	P y E
Un gen completo	duplicación completa	P y E
Región de un cromosoma	polisomía	P y E
Cromosoma entero	aneuploidia	E
Genoma total	poliploidía	P y E

P procarionte E eucarionte

Duplicación de un gene

La amplificación de un gene es un fenómeno genético generalizado en organismos procariontes . Las duplicaciones pueden surgir por la recombinación desigual entre dos moléculas de DNA en una horquilla de replicación. La recombinación ocurre entre dos diferentes copias de una secuencia corta repetida representada por las líneas gruesas, formándose una amplificación del gene o bien una duplicación en tandem (Romero y Palacios 1977; Anderson y Roth 1977)

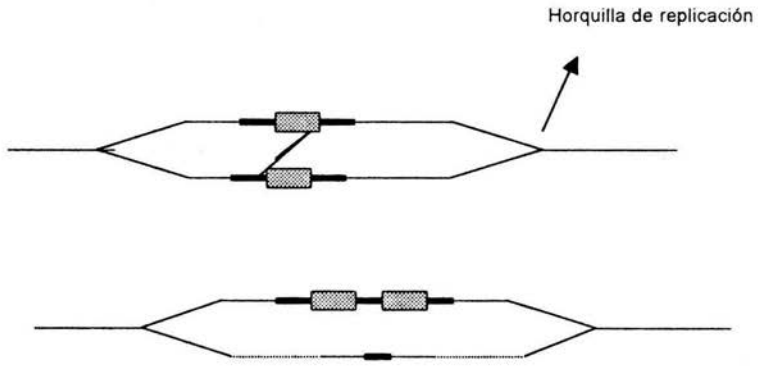


Fig 2 Modelo de duplicación de un gene durante la replicación del genoma bacteriano. La duplicación de un gene (rectángulos grises) puede ocurrir entre dos dobles hélices hijas. La recombinación desigual se realiza entre dos secuencias cortas repetidas (líneas gruesas) , resultando por un lado una duplicación en una hebra y una pérdida en la otra.

Una vez que un gen se duplica el nuevo gen puede mutar y eventualmente emergen nuevas funciones enzimáticas (*hisA* e *hisF)* (Alifano *et al,* ,1996*)* o bien solo sufrir elongación como el caso de *carB* (Lawson *et al,* 1996).

Duplicación de todo el genoma

La duplicación del genoma completo da como resultado una rápida expansión en el número de genes. En su libro *Evolution by gene duplication* Ohno (1970) propuso que nuevos genes eran producidos durante eventos de duplicación completa del genoma, y que estos eventos constituyeron un prerequisito para transiciones evolutivas mayores. Se ha planteado que la duplicación del genoma completo en procariontes puede ocurrir mediante: a) entrecruzamiento de genomas circulares idénticos b) unión cabeza cola de dos genomas lineares idénticos y c) un modo de replicación para el genoma bacterial primitivo similar al modo usado por algunos bacteriofagos actuales (Zipkas y Riley 1975).

El resultado de la duplicación completa del genoma es conocido como ploidía, y ha sido definida convencionalmente en células eucariontes y se refiere al número de cromosomas homólogos en los organismos ; la poliploidía es consecuencia de la no disyunción del material genético durante la meiosis (en eucariontes), por lo que hay un incremento en el contenido de DNA.

De acuerdo con el número de cromosomas homólogos presentes en las células, éstas pueden ser haploides (n) como los gametófitos y la mayoría de los procariontes, diploides (2n) como el ratón, o poliploides como algunas ranas arborícolas. La ploidía ha sido ampliamente observada en grupos biológicos muy separados entre sí tales como las levaduras, plantas (angiospermas pteridofitas) y en animales (ostracodos y algunos anfibios); es decir, es un proceso de origen polifilético (Soltis y Soltis 1999). La mayoría de los procariontes son haploides. Sin embargo, diferenciar este proceso durante los ciclos celulares en procariontes y eucariontes puede ser complejo. En algunos procariontes la ploidía ha sido percibida como consecuencia del desfasamiento entre una tasa más rápida de crecimiento con respecto a la replicación durante el ciclo celular (Trun 1999). Es decir, cuando *E.coli* crece a 60 minutos, cada nueva célula hereda un cromosoma; sí el valor de crecimiento de *E.coli* es más rápido que el tiempo de replicación del DNA, las células heredan cromosomas con horquillas de replicación, pudiendo suceder dos posibilidades:

a) La nueva célula puede heredar un cromosoma con más de una horquilla pero con un sólo sitio de término, en cuyo caso la célula es haploide; ó bien b) varios cromosomas en replicación, entonces la célula es técnicamente diploide (ver Trun 1999) Fig 3

Fig 3 .Estado del cromosoma a diferentes valores de crecimiento en *Escherichia coli* tomado de Turn (1999) y modificado

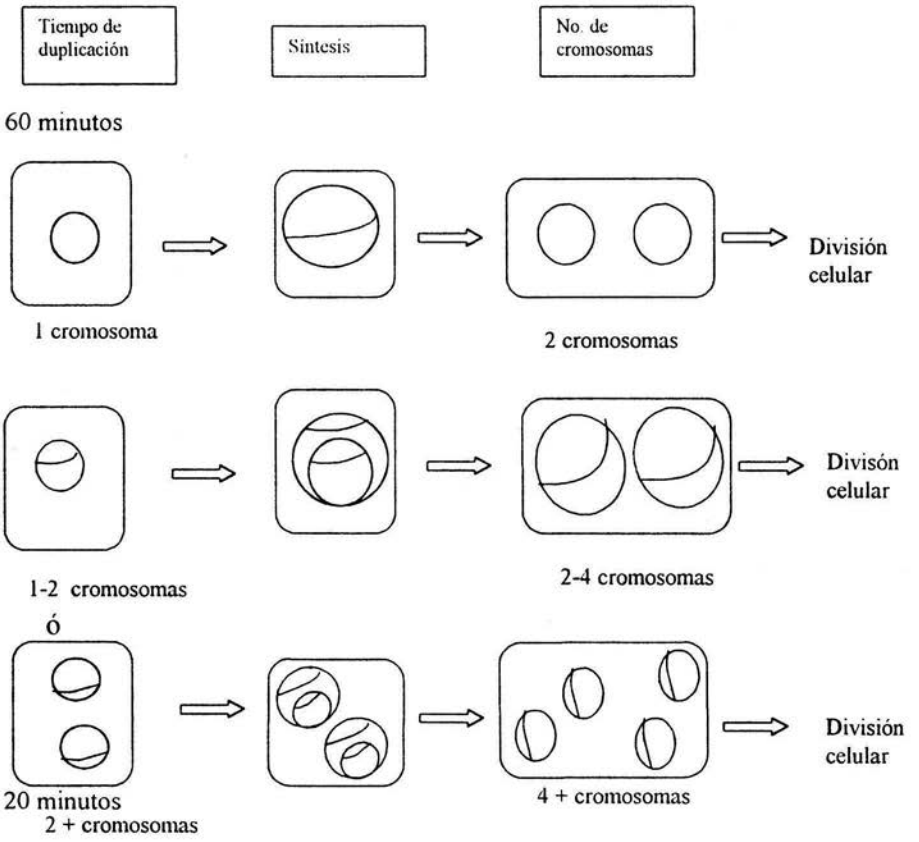


Tabla 2. No. de equivalentes por genoma en un población de procariontes en diferentes fases del ciclo celular . (Bendich y Drlica 2000)

Organismo	Fase estacionaria (No. Eq. x genoma al inicio)	Fase exponencial (No. Eq. x genoma al final)
Escherichia coli	2, 4, 8	11
Methanococcus jannaschii	3	7
Deinococcus radiodurans	4	10
Synechococcus PCC6301	--	8
Desulfovibrio gigas	--	9,17
Borrelia hermsii	--	16
Azotobacter vinelandii	--	80-100

El caso mas notable que se conoce es el que ocurre en la bacteria *Epulopiscium fishelsoni* en donde el contenido de DNA varía entre 4 y 5 ordenes de magnitud entre individuos en diferentes estados del ciclo de vida y puede exceder considerablemente la cantidad de DNA encontrada en el núcleo de mamíferos (Bresler 1998).

El incremento en la cantidad de material genético no asegura directamente una complejidad mayor en el genoma, porque, despues de todo, el resultado final implica que el organismo simplemente tiene una o más copias del genoma más que nuevos genes con capacidades metabólicas nuevas. Cuando un genoma se duplica hay muchas situaciones que pueden alterar el destino de la copia extra de las secuencias del genoma, entre las que se encuentran (a) la formación de pseudogenes; (b) la adquisición de nuevas funciones; (c) la adquisición de funciones parecidas.

La importancia de la duplicación total del genoma se ha convertido en un tema controvertido, porque las evidencias para afirmar que este tipo de eventos

sucedieron no es fácil de discernir, ni siquiera al analizar los genomas completamente secuenciados. Las evidencias más recientes (pero, al mismo tiempo, no aceptadas por todos) sobre la duplicación total del genoma provienen del análisis de dos genomas eucariontes completamente secuenciados: el de *Saccharomyces cerevisiae,* analizado por Wolfe y Shields (1997), quienes identificaron 55 grupos duplicados, comprendiendo 376 pares de genes conteniendo un mínimo de 3 genes en el mismo orden, y quienes concluyen que esta duplicación ocurrió hace 100 millones de años. Está también el caso de *Arabidopsis thaliana,* donde fueron encontradas 24 grandes regiones duplicadas en un genoma considerado relativamente compacto, (Simillon *et al*, 2002). Este descubrimiento llevó a pensar que la duplicación del genoma completo y su subsecuente contracción han sido un importante factor durante la evolución de genomas de plantas.

Después que la duplicación del genoma fue descrita como la mayor fuerza evolutiva en los vertebrados (Ohno 1970), este proceso se extrapoló para explicar la evolución de los procariontes. Tal postulado fue hipótesis la de trabajo central en varias investigaciones, que de forma general han seguido dos aproximaciones metodológicas distintas: (a) mediante el análisis del arreglo genómico de un solo organismo (Zipkas Riley (1975); y (b) el análisis de la distribución de un conjunto organismos con tamaño de genoma conocido (Wallace y Morowitz 1973; Sparrow y Neuman 1976, Herdman 1985;). En lo que se refiere al análisis del arreglo de secuencias en el genoma, Zipkas y Riley (1975) plantean que *E.coli* experimentó dos duplicaciones secuenciales de genoma en el pasado, lo cual se refleja en la posicion de algunos genes en un mapa circular del genoma de la bacteria. Al examinar el mapa genético de *E.coli*, encontraron que pares de genes bioquímicamente relacionados tienen una tendencia a orientarse cada 90 o 180 grados. Sin embargo sabemos que existen cambios y arreglos genómicos como movimiento de los genes y cambio en el orden de los mismos que suelen ocurrir (Fani et al., 1998). Si la duplicación del genoma tuvo lugar, estos sucesos deben haberse producido con poca frecuencia o muy simétricamente en todo el genoma

de tal forma que que se conservara dicho arreglo y que por tanto la distribución (orientación de los genes cada 90 o 180 grados) en el genoma no sea un arreglo aleatorio .

Por otra parte, Wallace y Morowitz (1973) analizaron la distribución de tamaños de genoma correspondientes a 98 especies bacterianas agrupadas en las familias Achromobacteraceae, Azotobacteraceae, Bacillaceae, Brevibacteriaceae, Brucellaceae, Corynebacteriaceae, Enterobacteriaceae, Lactobacillaceae, Micrococcaceae, Neissiaceae, Rhizobiaceae, Mycoplasmataceae, Nitrobacteraceae, Pseudomonadaceae, y Spirillaceae estimadas por cinética de renaturalización y microscopía electrónica . A través de su análisis ellos propusieron un esquema en donde los genomas más pequeños ($5x\ 10^9$ daltones) de los micoplasmas (en este caso, *Acholeplasma laidlawii*, un micoplasma saprofito aislado de aguas residuales) representan formas de vida primitiva y los denominaron genesistron, y concluyeron que la evolución subsecuente ocurrió por duplicación del DNA. Según Wallace y Morowitz Acholeplasma puede ser considerado como intermediario en la evolución de células protocariontes (es decir, una forma ancestral prebacteriana) a procariontes.

Sparrow y Neuman llevaron a cabo un analisis de las duplicaciones completas de muchos genomas distintos (1976) a partir de una distribución logarítmica de DNA por genoma de especies compiladas de diferentes reportes que expresan el contenido de DNA en una variedad de unidades (estas unidades fueron convertidas a picogramos para igualarlas). Su analisis comprendió 23 grupos filogenéticos ampliamente separados entre sí, incluyendo viroides, virus, bacterias, hongos, algas, protozoa, porifera, nematodos, insectos, cordados "inferiores", celenterados, angiospermas, vertebrados, moluscos equinodermos, anélidos, crustáceos y gimnospermas. La distribución del DNA tiende a formar varios picos a valores de múltiplos lo que parece representar duplicaciones de DNA intragrupo, en el caso de los eucariontes son independientes de la poliploidía, por lo que este fenómeno ha sido llamado criptopoliploidía y denota un incremento en el tamaño

de genoma por aumento en el tamaño del cromosoma. Cuando Sparrow y Neuman (1976) comparan duplicaciones teóricas contra valores mínimos de DNA en cada grupo, observaron una tendencia cíclica sobre 8 ordenes de magnitud a partir 300 nucleótidos (1.65×10^{-7} pg) calculados para un viroide de RNA que ellos interpretan como un genoma ancestral básico.

En cambio, Herdman (1985) estudió la distribución de 605 genomas de diversas cepas de bacterias, cuyos tamanos fueron calculados usando diversos métodos, incluyendo sobre todo cinética de renaturalización. La distribución fue discontinua mostrando picos modales de (1) .5, (2) 1.0-1.25, (3) 2.5-2.75, (4) una cola más larga que se extiende arriba de 4.75 y (5) muy pocos genomas entre 6 y 8.5×10^{-9} daltones. Además, Herdman postuló que los cambios en los tamaños de genoma son producto de dos principales procesos : a) fusión de genomas o b) duplicación de pequeños genomas ancestrales. Herdman también comentó que éste proceso ha ocurrido independientemente en diferentes grupos de bacterias y que el cambio de metabolismo anaerobio a aerobio ocurrió separadamente en cada una de las líneas bacterianas de descendencia y esto fue acompañado por una o más duplicaciones en el tamaño de genoma, aunque no sin dejar de hacer notar que existen mecanismos adicionales que intervienen en los cambios en el tamaño de genoma. Como se ve en los trabajos adjuntos, esta metodología y las conclusiones han sido modificadas substancialmente gracias a los datos disponibles.

Metodología

Se construyó una base de datos con los tamaños de genoma de 641 organismos procariontes determinados por por la técnica de Electroforesis de campo pulsado (PFGE), reportados en artículos incluidos en NCBI (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) Scirus (for a scientific information only http://www.scirus.com/) and Highwire (library of the sciences and medicine http://intl.highwire.org/) hasta febrero de 2003.

La base de datos fue completada con una descripción de la posición filogenética, el estilo de vida (simbiontes obligados, parásitos obligados y vida libre), el intervalo de temperatura de crecimiento (mesófilos 25-44 oC, termófilos 45-70oC e Hipertermófilos >70oC, y la respuesta al oxígeno de los organismos fue basada en reportes originales y en datos de Bergey's Manual of Bacterial Determination (Holt et al, 1994): anaerobios (organismos con respuesta al oxígeno negativa), microaerofílicos (organismos que requieren < 21% de oxígeno), facultativos anaerobios (organismos con doble respuesta al oxígeno) y aerobios (organismos que requieren 21 % de oxígeno). Se realizó un análisis estadístico (Ji cuadrada) usando el programa Microsoft Excel

EARLY METABOLIC EVOLUTION: INSIGHTS FROM COMPARATIVE CELLULAR GENOMICS

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

1. Introduction

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

2. Some problems in comparative genomic analysis

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

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

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

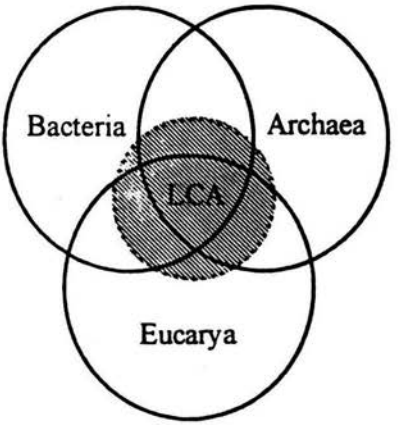


Figure 1. Intersection of the complete sequence spaces of the three domains defines the gene complement of the common ancestor (LCA). Identification of rapidly-evolving sequences would lead to a bigger set of ancestral genes (hatched areas).

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

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

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

TABLE I. Some processes in metabolic evolution.

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

3. Did metabolism evolve backwards?

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

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

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

4. The patchwork assemblage of biosynthetic routes

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

5. Gene duplication and arginine anabolism

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

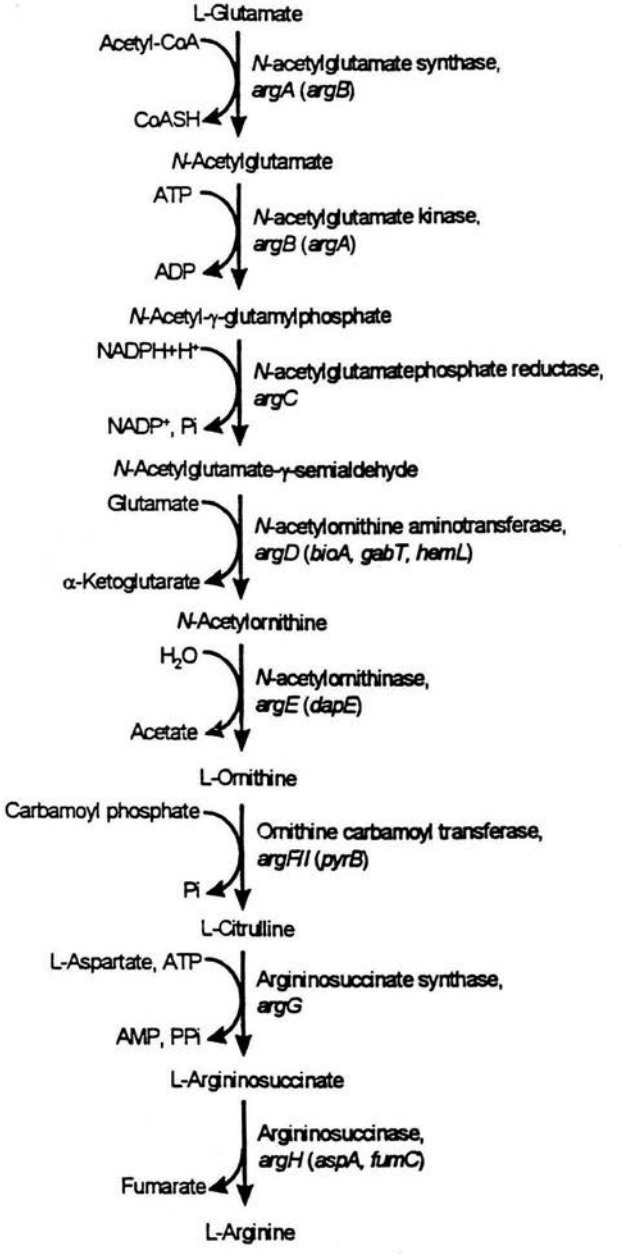


Figure 2. Arginine biosynthesis. The arginine biosynthetic genes paralogs are indicated within parenthesis.

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

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

6. Homologous enzymes can have different catalytic properties

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

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

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

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

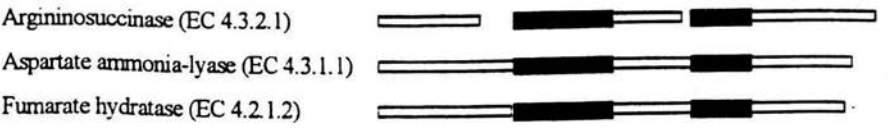


Figure 3. A three-member family of *E. coli* paralogous enzymes which different catalytic properties. The sequences were aligned using the Macaw program. The regions with statistically significant sequence similarity are shown in black.

7. Conclusions

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

Horowitz, N. H. (1965) The evolution of biochemical synthesis -retrospect and prospect, in V. Bryson and H. J. Vogel (eds.), *Evolving Genes and Proteins*, Academic Press, New York, pp. 15-23.

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

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

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

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

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

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

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

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

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

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

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

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

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

ON THE ROLE OF GENOME DUPLICATIONS IN THE EVOLUTION OF PROKARYOTIC CHROMOSOMES

S. ISLAS, A.CASTILLO, H.G. VÁZQUEZ, and A. LAZCANO
Facultad de Ciencias
Apartado Postal 70-407
Cd. Universitaria 04510
México D.F.,México.

1. Introduction

It is generally accepted that primitive cells were endowed with relatively small genetic systems with reduced encoding capacities. How the transition from such small genomes to the more complex ones observed in extant prokaryotes took place is still unknown, but it may have involved duplication of genes and of larger segments, horizontal transfer, cell fusion events, and perhaps even whole genome duplications (Casjens, 1998).

The role of whole genome duplications was first discussed for vertebrate evolution by Ohno (1970). Whether this took place or not is still a controversial issue (Hughes, 1999). On the other hand, compilation of data on mycoplasma DNA content led Wallace and Morowitz (1973) to suggest that the discontinuities in the frecuency distribution of genome sizes of their sample could be explained by succesive genome duplications that were assumed to have taken place also took place in other prokaryotes. Some time ago this idea appeared to be supported by the position between functional-related genes in the circular map of *Streptomyces coelicolor* (Hoopwood 1967) and *Escherichia coli* (Zipkas and Riley 1975). Furthemore, the results of statistical analysis of a sample of 603 prokaryotic genomes was interpreted as evidence of several rounds of entire duplications that had started with a modal value of aproximately 0.8 Mb, and led to genome of 1.6 Mb, 4 Mb, and other minor peaks with higher values (Herdman 1985).

With the development of pulsed-field gel electrophoresis (PFGE), a technique that allows the separation and analysis of large DNA fragments and the direct study of the physical structure of prokaryotic genomes, the accuracy in the determination of genome size has been significantly improved (Shimkets, 1998). Here we report the results of an analysis of a sample of 246 prokaryotic genome sizes obtained by PFGE, that includes both Bacteria and Archaea, available as of June 1999. Although this database is likely to be biased and does not represents the full range of prokaryotic diversity, the different peaks we have observed in the discontinuous distribution of DNA content (Figure1) do not support the idea that several duplications beginning an

J. Chela-Flores et al. (eds.), Astrobiology, 289–292.

hypothetical ancestral minigenome have taken place during evolutionary time.The posibility of a correlation between oxygen response and genome size is also discussed.

2. Material and methods

A genome size database was constructed with the 246 prokaryotic genome sizes determined by PFGE reported in the publications included in the NCBI database PubMed (http://www.ncbi.nlm.mih.gov/PubMed/). This database is available upon request. The information was completed with a description of the organisms' phylogenetic positions and lifestyle. The organisms' response to oxygen was based on the original reports and on data from the *Bergey's Manual of Bacterial Determination* (1994): anaerobe (organisms with negative oxygen response), microaerophilic (organisms that need ≤ 21% of oxygen), facultative anaerobe (organisms with double response to free oxygen), and aerobes (organisms that require ≥ 21% of oxygen). Statistical analyses were performed using the X^2 in Microsoft Excel ^{tm} program.

3. Results

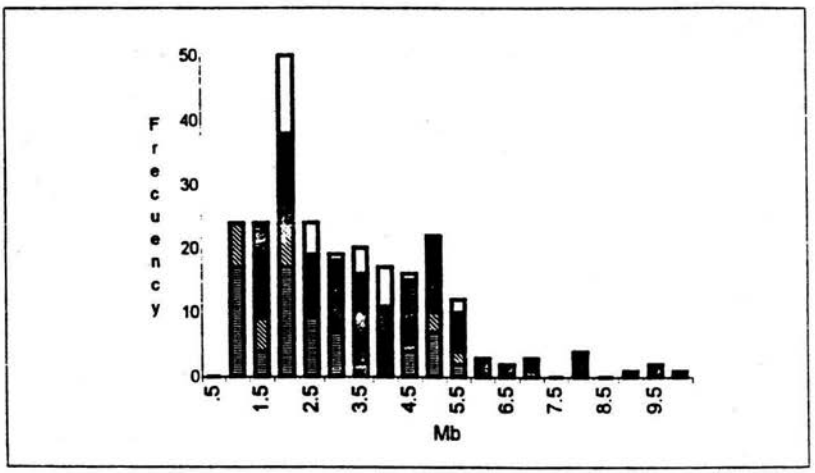


Figure 1. Distribution of prokaryotic genome sizes with respective oxygen response: the white bars correspond to anaerobes, diagonal lines, to microaerophilics; horizontal lines, to facultatives and grey points, to aerobes.

The distribution of prokaryotic genome size and their oxygen response in our sample is shown in Figure 1. The first two bars corresponding to organisms with parasitic lifestyle; in the first bar only low and double response to oxygen are represented. From the third bar onwards, the four types of metabolic response to oxygen are

represented, including both parasitic and free-living organisms. The facultative anaerobes have genome sizes ranging from the smallest ones to middle sized (0.57-570 Mb). The aerobic organisms have the largest genome reported, while the strict anaerobes include the genomes with sizes between (1.12- 9.50 Mb), and the negative response include the genomes sizes with DNA content is (1.60-5.40 Mb) are never extreme values include as 9.50 Mb.

4. Discussion and conclusions

Herdman (1985) has argued that the peaks observed in the discontinuos size distribution of his sample of bacterial genomes provided support for the hypothesis that the evolution of prokaryotic DNA content could be explained by whole genome duplications. However, although our sample is larger and based in a more accurate technique for detecting DNA content, we have found no evidence corroborating his conclusions. The range of prokaryotic genome sizes available as of June 1999, ranges from 0.573 Mb (*Mycoplasma genitalium)* to 9.5 Mb (*Stigmatiella erecta).* Since the smallest free-living prokaryotes included in our sample have genome sizes in the range of 1.6 Mb to 1.9 Mb, according to the whole genome duplication hypothesis we would expect to find peaks with modal values of 3.2 Mb, 6.4 Mb, and 12.8 Mb. This is clearly not the case (Figure 1).

The smaller genome sizes (0.57 to 1.5 Mb) in our sample correspond to parasitic organisms of the Mollicutes group. The later include other mycoplasma that have, whose slightly larger genomes such as *Anaeroplasma, Asteroleplasma,* and *Spiroplasma*, oscillate between 1.5 to 1.78 Mb. Other groups with reduced genome sizes are the rickettsia and the spirochaete. Not all these organisms with reduced DNA content are anaerobes.. which can be explained by recognizing that these organisms (genome size $\leq$ 1.78 Mb), are the outcome of a complex series of secondary adaptations that have led to the polyphyletic reduction of their genome dimensions. Thus, neither the mycoplasma or the rickettsia are accurate models of ancestral Archean organisms.

It has been suggested that the first microorganisms were anaerobic heterotrophes, and the later availability of oxygen promoted the apparence of new metabolic capacities (Oparin, 1938). Our results show a correlation between genome size and oxygen response. As seen in Figure 1, the larger genomes are found solely in bacteria, all which are strict aerobes with complex life cycles. These results strongly suggest that such species evolved during late Proterozoic times, once the levels of free-O_2 in the terrestial atmosphere had reached values comparable to the extant ones.

There are no reports available in the literature of Archaea with genome size comparable to those of *Stigmatiella* (Casjens, 1998) (Figure 1). Whether this reflects the

evolutionary strategies of the Archaea domain, it is not clear, but our interpretation may be limited by the current descriptions of prokaryotic diversity.

The database analyzed here is biased by the medical and economical significance of the organisms, and does not reflect in an accurate way the actual biodiversity of prokaryotes. Pathogens and parasites are clearly overepresented because of their medical and economic significance in human, animal, and crop plant life. Nonetheless, the large number of microaerophilic and facultative organisms from different phylogenetic groups in our sample probably reflects the succesful adaptation to increasingly higher levels of oxygen in the terrestrial atmosphere.

Finally, we would like underline the fact that in spite of the limitation of our database there are no indications of free-living prokaryotes with genome sizes smaller than 1.53 Mb (*Fervidobacterium islandicum*). This observation casts doubts on the existence of nanobacteria (Kajander and Cifícioglu 1998; Aboll, 1999), whose genomes are assumed to be at least one order of magnitude smaller than those of mycoplasma.

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

References

Aboll, A. (1999) Battle lines drawn between nanobacteria researchers. Nature **401**: 105
Casjens, S. (1998) The diverse and dynamic structure of bacterial genomes *Annu Rev genet* **32**:339-377
Herdman M (1985) The evolution of bacterial genomes . In: Cavalier Smith T (ed) *The Evolution of genome size*. John Wiley, London
Holt J G *et al* (1994) *Bergey's Manual of Determinative Bacteriology*. Williams and Wilkins. Baltimore, Maryland, USA, p.787
Hopwood D A (1967) Genetic analysis and Genome structure in *Streptomyces coelicolor*. *Bacteriol Rev* **31**: 373-403
Hughes, A. L. (1999) Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history *J Mol Evol* **48**:565-576
Kajander,E.O. and Cifícioglu, N. (1998) Nanobacteria: an alternative mechanism for pathogenic intra-and extracelluklar calcification and stone formation Proc. Natl. Acad Sci USA **95**: 8274-8279
Ohno S (1970) *Evolution by Gene Duplication*. Springer Verlag, New York
Oparin A. I. (1938) *The Origin of Life*, MacMillan, New York.
Shimkets, L .J. (1998) Structure and sizes of the genomes of the Archaea and Bacterial In Bruijn F *et al* (eds).*Bacterial Genomes: Physical Structure and Analysis*. (Chapman & Hall, New York)
Wallace D C and Morowitz H J (1973) genome size and evolution *Chromosome* **40**: 121-126
Zipkas, D. and Riley, M. (1975) Proposal concerning mechanism of evolution of the genome of *Escherichia coli Proc. Nat. Acad. Sci. USA* **72**(4):1354-1358

NCBI

PubMed

Search PubMed for Go Clear

Limits Preview/Index History Clipboard Details

Display Abstract Show: 0 ort Send to Text

1: Int Microbiol. 2003 Jun 28 [Epub ahead of print]. Related Articles, Links

Hyperthermophily and the origin and earliest evolution of life.

Islas S, Velasco AM, Becerra A, Delaye L, Lazcano A.

Facultad de Ciencias, UNAM, Ciudad Universitaria, Apdo. Postal 70-407, 04510, Mexico D.F., Mexico.

The possibility of a high-temperature origin of life has gained support based on indirect evidence of a hot, early Earth and on the basal position of hyperthermophilic organisms in rRNA-based phylogenies. However, although the availability of more than 80 completely sequenced cellular genomes has led to the identification of hyperthermophilic-specific traits, such as a trend towards smaller genomes, reduced protein-encoding gene sizes, and glutamic-acid-rich simple sequences, none of these characteristics are in themselves an indication of primitiveness. There is no geological evidence for the physical setting in which life arose, but current models suggest that the Earth's surface cooled down rapidly. Moreover, at 100 degrees C the half-lives of several organic compounds, including ribose, nucleobases, and amino acids, which are generally thought to have been essential for the emergence of the first living systems, are too short to allow for their accumulation in the prebiotic environment. Accordingly, if hyperthermophily is not truly primordial, then heat-loving lifestyles may be relics of a secondary adaptation that evolved after the origin of life, and before or soon after separation of the major lineages.

PMID: 12838394 [PubMed - as supplied by publisher]

Display Abstract Show: 0 ort Send to Text

International Microbiology

DOI 10.1007/s10123-003-0113-4

Review Article

Hyperthermophily and the origin and earliest evolution of life

Sara Islas · Ana M. Velasco · Arturo Becerra · Luis Delaye · Antonio Lazcano (✉)

S. Islas · A. M. Velasco · A. Becerra · L. Delaye · A. Lazcano
Facultad de Ciencias, UNAM, Ciudad Universitaria, Apdo. Postal 70-407, 04510 México D.F., México

✉ A. Lazcano
Phone: +52-5-6224823
Fax: +52-5-6224828
E-mail: alar@correo.unam.mx

Received: 30 January 2003 / **Accepted:** 25 February 2003

Abstract The possibility of a high-temperature origin of life has gained support based on indirect evidence of a hot, early Earth and on the basal position of hyperthermophilic organisms in rRNA-based phylogenies. However, although the availability of more than 80 completely sequenced cellular genomes has led to the identification of hyperthermophilic-specific traits, such as a trend towards smaller genomes, reduced protein-encoding gene sizes, and glutamic-acid-rich simple sequences, none of these characteristics are in themselves an indication of primitiveness. There is no geological evidence for the physical setting in which life arose, but current models suggest that the Earth's surface cooled down rapidly. Moreover, at 100 °C the half-lives of several organic compounds, including ribose, nucleobases, and amino acids, which are generally thought to have been essential for the emergence of the first living systems, are too short to allow for their accumulation in the prebiotic environment. Accordingly, if hyperthermophily is not truly primordial, then heat-loving lifestyles may be relics of a secondary adaptation that evolved after the origin of life, and before or soon after separation of the major lineages.

Keywords Hyperthermophily · Comparative genomics · Organic-compound stability · Last common ancestor · Origin of life

Introduction

A thermophilic origin of life is not a new idea. "Heat has been justly regarded the mother of all generations," wrote Jean-Baptiste Lamarck in his 1804 *Philosophie Zoologique*, adding that "it cannot be doubted that suitable portions of inorganic matter, occurring amidst favorable surroundings, may by the influence of Nature's agents, of which heat and moisture are the chief, receive an arrangement of their parts that foreshadows cellular organization, and thereafter pass to the simplest organic state and manifest the earliest movements of life" [22].

Lamarck's ideas are echoed in a number of contemporary proposals on a hot origin of life. It is not surprising that the correlation between hyperthermophily and antiquity has led to suggestions of a high-temperature emergence of life. This interpretation has been reinforced by a number of facts, including large-scale analysis suggesting that, soon after its formation, the surface of primitive Earth was extremely hot. The planet is generally thought to have remained molten for some time after its formation 4.6×10^9 years ago, although evidence of a 4.4×10^9-year-old hydrosphere implies that its surface cooled down rapidly [52]. However, the Earth underwent late accretion impacts that may have boiled-off the oceans as late as 3.8×10^9 years ago [41]. Moreover, both paleontological and molecular fossil records appear to support the possibility of a hyperthermophilic origin of life: (a) the 3.49- to 3.43×10^9-year-old Australian Warrawoona stromatolitic chert horizons [37] are endowed with the diagnostic features of a microbial community associated with a seafloor hydrothermal system [49]; and (b) rooted universal single-gene phylogenies have shown that hyperthermophiles are not randomly distributed in universal trees, but occupy the deepest, shorter branches towards the lowest portion of molecular rRNA-based cladograms [1, 34, 44].

However, attempts to infer the conditions of life based on the traits of heat-loving prokaryotes have led to opposing suggestions: while some advocate a hyperthermophilic heterotrophic emergence of life [7, 19], others hypothesize that mineral surfaces in hot volcanic settings fueled the appearance of primordial chemoautolithotrophic biological systems lacking genetic material [50]. Regardless of these differences, all hot-origin-of-life

scenarios face the same problem, i.e., the chemical decomposition of presumed essential biochemical compounds, such as amino acids, ribose, nucleobases, RNA, and other thermolabile molecules, whose half-lives for decomposition at temperatures between 250 °C and 350 °C are at the most a few minutes [28, 51].

Is it possible, then, that the evidence supporting a hot origin of life is being misinterpreted, i.e., that the extrapolation of molecular phylogenies into prebiotic times is misleading? The purpose of this paper is to review the evidence against an extremophilic origin of life and a heat-loving RNA world, thus supporting the possibility that (hyper) thermophilic microbial lifestyles are in fact the outcome of secondary adaptations during early stages of cell evolution.

The genomes of heat-loving prokaryotes

Comparison of archaeal and bacterial genomes has led to the identification of a number of thermophilic/hyperthermophilic-specific signatures, including the low abundance of the dinucleotide CG in their DNA [21], amino acid compositional biases [48], reduced protein-encoding gene length [55, 48], and the presence of reverse gyrase, an ATP-dependent topoisomerase described as a hallmark of a heat-loving lifestyle [13]. As discussed below, hyperthermophilic genomes have additional characteristic traits.

Although considerable variations in DNA content exist within closely related bacterial species and strains, the available data suggest that genome sizes of each of the three domains appear to lie within defined ranges [40]. As part of an attempt to study the size and organization of prokaryotic chromosomes, a database was constructed with 641 archaeal and bacterial genome sizes determined by pulsed-field gel electrophoresis (PFGE), published as of December 2002. This database is available upon request, and will be published elsewhere.

Figure 1 shows the genome size distribution of our sample. Mesophilic proteobacteria have the largest genomes, ranging from 0.448 Mb for *Buchnera* sp. to 9.7 Mb for *Azospirillim lipoferum* 59B. The smallest bacterial genome sizes correspond to obligate symbionts such as *Buchnera* sp. and pathogens such as *Mycoplasma*, *Chlamydia* and *Rickettsia*, whose small DNA content is a derived character that reflects secondary gene loss due to their parasitic lifestyles.

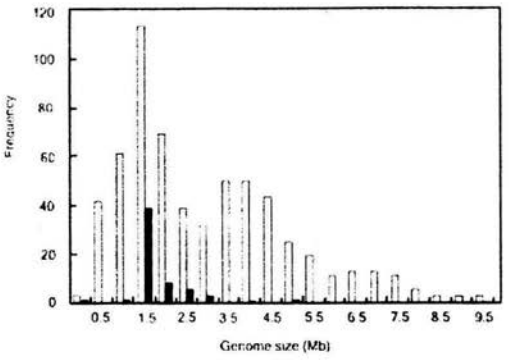


Fig. 1. Chromosomal DNA content distribution for mesophiles (*white bars*) and hyperthermo- and thermophilic (growth temperature 80 °C or more) prokaryotes (*black bars*). Prokaryotic genome sizes shown here were reported in the literature as determined by pulsed-field gel electrophoresis (n=641)

Of course, the data summarized in Fig. 1 are biased by an inadequate sampling that does not fully represent the true levels of microbial biodiversity, but is clearly skewed, on the one hand, towards pathogenic bacteria and, on the other, to extremophilic archaea. Nevertheless, it provides useful insights into the size and organization of prokaryotic genomes. Although thermophilic and hyperthermophilic bacterial and archaeal genomes follow a trend similar to that of their mesophilic counterparts, they depart from a normal distribution and fall within a well-defined size range (from 0.5 Mb for the thermophilic ectosymbiont *Nanoarchaeon equitans,* to the 5.10 Mb of the facultative thermophilic *Methanosarcina acetivorans*), with a maximum at approximately 2 Mb. However, this size range does not necessarily reflect a correlation between DNA content, heat-loving microbial lifestyles and antiquity, since many different mesophilic bacterial species, including *Leptospira*, green-sulfur bacteria, cyanobacteria, spirochaetes, fusobacteria and actinobacteria, are endowed with similarly small-sized chromosomes.

Genomic analysis has shown that thermophilic and hyperthermophilic genomes are endowed with smaller protein-encoding genes than their mesophilic counterparts [48, 55]. Detailed statistical analysis of 56 complete genomes, including seven eukaryotes, 14 archaeal and 35 bacterial species, has shown that the mean protein length of heat-loving prokaryotes (283±5.8) is significantly smaller than in mesophiles (340±9.4) [48]. It is possible that these reduced gene sizes are correlated with an extremophilic lifestyle, such as protein thermostability. As shown in Fig. 2, however, reduced gene size is a polyphyletic trait: small protein-encoding genes that fall within the same size range of hyperthermophilic genes are

also found in a wide diversity of non-extremophiles, including proteobacteria, green-sulfur bacteria, low GC gram-positives, fusobacteria, and mesophilic euryarchaea.

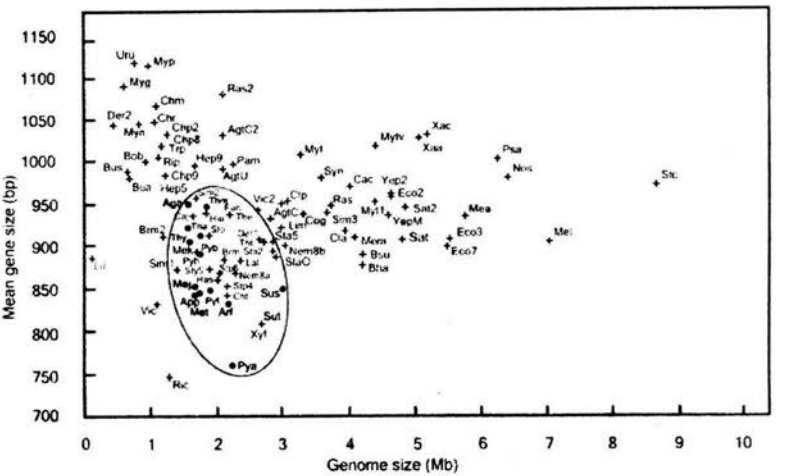


Fig. 2. Protein-encoding gene size distribution as a function of genome size. *Black dots* correspond to extremophilic proteomes of species that live at 80 °C or more. The area within which extremophilic ORFs are located is *circled* for clarity. *Symbols*: Archaea hyperthermophilic crenarchaea: *Aeropyrum pernix* (App), *Sulfolobus solfataricus* (Sus), *Sulfolobus tokodaii* (Sut), *Pyrobaculum aerophilum* (Pya); hyperthermophilic euryarchaea: *Pyrococcus abyssi* (Pyb), *Pyrococcus furiosus* (Pyf), *Pyrococcus horikoshii* (Pyh), *Methanococcus jannaschii* (Mcj), *Thermoplasma acidophilum* (Tha), *Methanobacterium thermoautotrophicum* (Met), *Archaeoglobus fulgidus* (Arf), *Thermoplasma volcanium* (Thv); non-hyperthermophilic euryarchaea: *Halobacterium* sp. (Has), *Methanopyrus kandleri* (Mek), *Methanosarcina mazei* (Mem), *Methanosarcina acetivorans* (Mea): Bacteria hyperthermophilic bacteria: *Aquifex aeolicus* (Aqa), *Thermotoga maritima* (Thm); α-Proteobacteria: *Rickettsia conorii* (Ric), *Rickettsia prowazekii* (Rip), *Sinorhizobium meliloti* (Sim1, Sim2, Sim3), *Brucella melitensis* (Brm, Brm2), *Agrobacterium tumefaciens* C58 Uwash (AgtU), *Agrobacterium tumefaciens* C58 Cereon (AgtC, AgtC2), *Mesorhizobium loti* (Mel), *Caulobacter crescentus* (Cac); Beta proteobacteria: *Neisseria meningitidis* MC58 (Nem8a, Nem8b), *Ralstonia solanacearum* (Ras, Ras2); γ-Proteobacteria: *Xylella fastidiosa* (Xyf), *Vibrio cholerae* (Vic, Vic2), *Haemophilus influenzae* (Hai), *Salmonella typhi* (Sat), *Salmonella typhimurium* LT2 (Sat2), *Escherichia coli* K12 (Eco2), *Escherichia coli* O157H7 (Eco7), *Escherichia coli* O157H7 EDL933 (Eco3), *Yersinia pestis* KIM (YepM), *Yersinia pestis* CO92 (Yep2), *Pseudomonas aeruginosa* (Psa), *Xanthomonas citri* (Xac), *Xanthomonas campestris* (Xaa), *Pasteurella multocida* (Pam), *Buchnera aphidicola* Sg (Bua), *Buchnera* sp. (Bus);δ/ε-Proteobacteria: *Campylobacter jejuni* (Caj), *Helicobacter pylori* 26695 (Hep5), *Helicobacter pylori* J99 (Hep9); green sulfur bacteria: *Chlorobium tepidum* TLS (Cht); gram-positive,

low-GC: *Streptococcus pneumoniae* TIGR4 (Stp4), *Streptococcus pneumoniae* R6 (Stp6), *Streptococcus pyogenes* MGAS315 (Sty5), *Listeria innocua* (Lii), *Listeria monocytogenes* (Lim), *Thermoanaerobacter tengcongensis* (Tht), *Staphylococcus aureus* MW2 (Sta2), *Staphylococcus aureus* Mu50 (Sta0), *Staphylococcus aureus* N315 (Sta5), *Lactococcus lactis* (Lal), *Streptococcus pyogenes* (Sto), *Clostridium perfringens* (Clp), *Clostridium acetobutylicum* (Cla), *Bacillus subtilis* (Bsu), *Bacillus halodurans* (Bah), *Mycoplasma pneumoniae* (Myn), *Mycoplasma genitalium* (Myg), *Mycoplasma pulmonis* (Myp), *Ureaplasma urealyticum* (Uru); gram-positive, high-GC: *Streptomyces coelicolor* (Stc), *Mycobacterium tuberculosis* CDC1551 (Myt1), *Mycobacterium tuberculosis* H37Rv (Mytv), *Mycobacterium leprae* (Myl); radioresistant bacteria: *Deinococcus radiodurans* (Der1, Der2); Fusobacteria: *Fusobacterium nucleatum* (Fun); cyanobacteria: *Synechocystis* CC6803 (Syn), *Thermosynechococcus elongatus* (The), *Nostoc* sp. (Nos); actinobacteria: *Corynebacterium glutamicum* (Cog); chlamydia: *Chlamydophila pneumoniae* AR39 (Chp9), *Chlamydophila pneumoniae* J138 (Chp8), *Chlamydophila pneumoniae* CWL029 (Chp2), *Chlamydia trachomatis* (Chr), *Chlamydia muridarum* (Chm); spirochete: *Borrelia burgdorferi* (Bob), *Treponema pallidum* (Trp)

Like their mesophilic counterparts, hyperthermophilic genes are endowed with simple sequences, i.e. homopolymeric tracts and tandem arrays of multiple short repeat motifs. These low-complexity regions have their origin in mutational processes, such as slipped-strand mispairing and unequal crossing-over, that take place during DNA replication and are known to represent a major source of genetic variation in pathogenic prokaryotes [31]. Analysis of all the completely sequenced hyperthermophilic and thermophilic genomes available as of December 2002 shows that the natural amino acid composition of each proteome is enhanced with respect to its corresponding simple sequences, which have a compositional bias as shown by the abundance of small, α-helix forming amino acids, i.e., alanine, leucine, lysine, serine and glutamic acid (Becerra, Cocho, Delaye and Lazcano, unpublished results). As shown in Fig. 3, however, simple sequences in hyperthermophiles are clearly enriched in glutamic acid. The stability of the α-helix structure of glutamic acid homopolymers under acid pH values [35] probably explains why, with the exception of *Thermoplasma acidophilum*, simple sequences of acid-resistant heat-loving prokaryotes tend to be rich in this amino acid. Enrichment of glutamic acid in extremophilic simple sequences explains the relative abundance in hyperthermophilic genomes, as noted by Tekaia et al. [48].

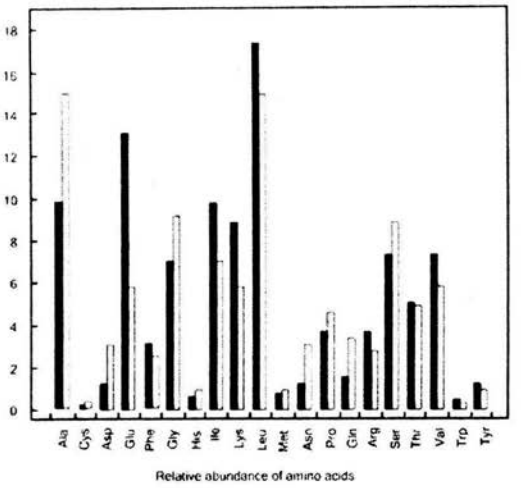


Fig. 3. Relative abundances of amino acids in simple sequences in all available proteomes as of December, 2002. Simple sequences were identified using the SEG program [53], which identifies low-complexity regions in which an enhanced concentration of short repeats not due to chance events can be detected. *White bars* show the average simple-sequence amino acid composition of mesophiles, and *dark bars* show those of hyperthermophilic prokaryotes (80 °C or more). The hyperthermophilic species represented here are *Pyrococcus furiosus, P. horikopshi, P. abyssi, Aeropyrum pernix, Methanococcus jannaschii, Archaeoglobus fulgidus, Sulfolobus solfataricus* and *S. tokodaii*

The faulty records of archaean life

As shown by recent debates, the identification of the oldest paleontological traces of life can be a highly contentious issue. The early archaean geological record is scarce, and most of the preserved rocks have been metamorphosed to a considerable extent. However, the evidence suggests that life emerged on Earth as soon as it was possible to do so. Although the biological origin of the microstructures interpreted as cyanobacterial remnants in the 3.5×10^9 year old Apex sediments of the Australian Warrawoona formation [38] have been questioned [3], there is additional paleontological evidence that highly diverse microbial communities were thriving during the early and middle Archaean [32].

Unfortunately, it is unlikely that data on how life originated will be provided by the geological record. There is no direct evidence of the environmental conditions on the Earth at the time of the origin of life, nor is there 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 that may or may not have been important for the emergence of living systems.

The attributes of the first living organisms are also 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 sugar-phosphate backbones. It is possible that the only property they shared with extant organisms was the structural complementarity between monomeric subunits of replicative informational polymers, e.g. the joining together of residues in a growing chain whose sequence is directed by preformed polymers. Such ancestral polymers may have not even involved nucleotides. Accordingly, the most basic questions pertaining to the origin of life relate to much simpler replicating entities predating by a long series of evolutionary events the oldest recognizable heat-loving prokaryotes represented in molecular phylogenies.

The rooting of universal cladistic trees determines the directionality of evolutionary change and allows ancestral characters to be distinguished from those that were derived. Determination of the rooting point of a tree normally imparts polarity to most or all characters. 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 do they appear to be more rudimentary in their metabolic abilities than their aerobic counterparts. Primitive living systems would initially refer to pre-RNA worlds, in which life may have been based on polymers using backbones other than ribose-phosphate and possibly bases different from adenine, uracil, guanine and cytosine, followed by a stage in which life was based on RNA as both genetic material and catalysts [23].

Molecular cladistics may provide clues to some very early stages of biological evolution, but it is difficult to see how the applicability of this approach can be extended beyond a threshold that corresponds to a period of cellular evolution in which protein biosynthesis was already in operation, i.e., an RNA/protein world. Older stages are not yet amenable to molecular phylogenetic analysis. A cladistic approach to the origin of life itself is not feasible, since all possible intermediates that may have once existed have long since vanished.

Was the last common ancestor a hyperthermophile?

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. No paleontological remains will bear testimony of its existence, as the search for a fossil of the cenancestor is bound to prove fruitless. 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 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 now available, then quite unexpectedly the resulting repertoire is formed by relatively few features and by incompletely represented biochemical processes [8]. 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 poorly preserved [11].

Analysis of an increasingly large number of completely sequenced cellular genomes has revealed major discrepancies in the topology of rRNA trees. Very often these differences have been interpreted as evidence of horizontal gene-transfer events between different species, questioning the feasibility of the reconstruction and proper understanding of early biological history [10]. There is clear evidence that genomes have a mosaic-like nature whose components come from a wide variety of sources. 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.

The resulting 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 canonical 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 [53] has argued that the LCA was not a single organismic entity, 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 viewpoint, the development of the essential features of translation and of metabolic pathways took place before the earliest branching event, but what led to the three domains was not a single ancestral lineage, rather 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 decreasing sequence exchange and increasing genetic isolation would eventually lead to the observed tripartite division of the biosphere.

Did the hypothetical communal progenote ancestor proposed by Woese [53] diverge sharply into the three domains soon after the appearance of the code and the establishment of translation? Not necessarily, since inventories of LCA genes clearly include sequences that originated in different pre-cenancestral epochs. The origin of the mutant sequences ancestral to those found in all extant species, and the divergence of the Bacteria, Archaea, and Eukarya 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 LCA [8].

Universal gene-based phylogenies ultimately reach a single universal entity, but the bacterial-like LCA [8] that we favor was not alone. Company must have been provided by its siblings, a population of entities similar to it that existed throughout the same period. They may not have survived, but some of their genes did if they became integrated via lateral transfer into the LCA genome. The cenancestor should thus be considered as the evolutionary outcome of a series of ancestral events, including lateral gene transfer, gene losses, and paralogous duplications, that took place before the separation of Bacteria, Archaea, and Eukarya.

Comparisons of combined ortholog protein data sets that exclude sequences which may have undergone lateral transfer are consistent with rRNA trees [5]. Genomic trees also exhibit an excellent broad-level agreement with rRNA-based phylogenies [47]. Genomic trees are not cladograms but phenograms, i.e., they are hierarchical representations of similarities and differences in gene content, in which 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 rooted in the same area as rRNA phylogenies, which suggests that massive lateral transfer events between distant groups has not obliterated the early history of life. Thus, although hyperthermophiles may be displaced from their basal position if molecular markers other than elongation factors or ATPase subunits are employed, or if alternative

phylogeny-building methodologies are used [4], it can still be argued that rRNA-based phylogenies provide one of the best-preserved historical records of cell evolution [53].

The recognition that the deepest branches in rooted universal phylogenies are ocuppied by hyperthermophiles does not by itself provide 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. [16] to conclude that the universal ancestor was a mesophile. This possibility has been contested by Di Giulio [9], who has argued for a thermophilic or hyperthermophilic LCA. However, since the time factor is absent from the methodology developed by Galtier et al. [16], the inferred low G+C content of the cenancestral rRNA does not necessarily belong to the cenancestor itself, but may correspond to a mesophilic predecessor that may have been located along the trunk of the universal tree.

Chemical evolution and extreme environments

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. The first successful synthesis of organic compounds under plausible primordial conditions was accomplished 50 years ago by the action of electric discharges acting for a week over a mixture of CH_4, NH_3, H_2, and H_2O, and led to complex mixture of monomers that included racemic mixtures of several proteinic amino acids, in addition to hydroxy acids, urea and other molecules [27]. Prebiotic synthesis of amino acids largely proceeds by a Strecker synthesis that involves the aqueous-phase reactions of highly reactive intermediates (Structure

$$RCHO + HCN + 2\,NH_3 \rightleftharpoons RCH(NH_2)CN + H_2O$$

1, Structure

$$RCH(NH_2)CN + 2\,H_2O \longrightarrow RCH(NH_2)COOH + NH_3$$

2).

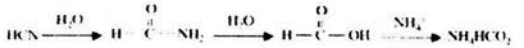
Detailed studies of the equilibrium and rate constants of these reactions demonstrated that both amino acids and hydroxy acids can be synthesized at high dilutions of HCN and aldehydes in a simulated primitive ocean. The reaction rates depend on temperature, pH, HCN, NH_3, and aldehyde concentrations, and are rapid on a geological time scale; the

half-lives for the hydrolysis of the intermediate products in the reactions, amino nitriles and hydroxy nitriles, are less than a 1,000 years at 0 °C, and there are no known slow steps [30].

The remarkable ease by which adenine can be synthesized by the aqueous polymerization of ammonium cyanide demonstrated the significance of HCN and its derivatives in prebiotic chemistry [33]. As summarized elsewhere [30], the prebiotic importance of HCN has been further substantiated by the discovery that the hydrolytic products of its polymers include amino acids, purines, and pyrimidines. The reaction of cyanoacetylene or cyanoacetaldehyde (a hydrolytic derivative of HCN) with urea leads to high yields of cytosine and uracil, especially under simulated evaporating pond conditions which increase the urea concentration [36].

The ease of formation under reducing conditions (CH_4+N_2, NH_3+H_2O, or $CO_2+H_2+N_2$) of amino acids, purines, and pyrimidines in one-pot reactions strongly suggests that these molecules were present in the prebiotic broth. In addition, 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 prebiotic 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-year-old Murchison carbonaceous meteorite, which also yielded evidence of liquid water in its parent body [12].

A major advantage of high temperatures is that chemical reactions go faster, and the primitive enzymes, once they appeared, could have thus been less efficient but nonetheless effective. However, the price paid is manifold: 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 (Structure



3). (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 (see Structure 1); and (d) loss of organic compounds by thermal decomposition and diminished stability of genetic polymers.

Extremophilic genomes are protected against thermal decomposition by a number of enzyme-dependent mechanisms [18], but these would have not been available during prebiotic times or at the time of the origin of life. In fact, the existence of an RNA world with ribose appears to be incompatible with a (hyper)thermophilic environment [29]. Survival of nucleic acids is limited by the hydrolysis of phosphodiester bonds [24], and the stability of Watson-Crick helices (or their pre-RNA equivalents) is strongly diminished by high-temperatures. For an RNA-based biosphere, reduced thermal stability on the geologic time scale of ribose and other sugars is the worst problem, but the situation is equally bad for pyrimidines, purines and some amino acids. As summarized elsewhere [23, 39], measurements by different groups have shown that 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. The half-life for hydrolytic deamination of cytosine at 100 °C is 19–21 days, although at 100 °C the half-life of uracil is approximately 12 years. At 100 °C, the thermal stability of purines is also reduced: 204–365 days for adenine, with guanine having a low half-life.

A hyperthermophilic pyrite-dependent origin of life?

An alternative to the problem of low half-lives of biochemical monomers at temperatures of 100 °C or more is to assume an autotrophic origin of life. Such proposals are periodically resurrected, but they are generally made without supportive evidence. The most elaborate chemoautotrophic-origin-of-life scheme has been proposed by Wächtershäuser [50]. 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 resembled 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 sulfur 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 [25].

The reaction $FeS+H_2S \rightarrow FeS_2+H_2$ is a very favorable 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. Pyrite-mediated CO_2 reduction to amino acids, purines and pyrimidines is yet to be achieved. However, as reviewed elsewhere [6, 20, 25], the FeS/H_2S combination has been shown to: (a) reduce nitrate and acetylene; (b) induce peptide-bond formation that results from the activation of amino acids with carbon monoxide and (Ni, Fe)S; and (c) to induce the synthesis of acetic acid and pyruvic acid from CO under simulated hydrothermal conditions in the presence of sulfide minerals [6, 20, 25]. However, support for Wächtershäuser's central tenets is meager. Life does not consist solely of metabolic cycles, and none of these experiments prove that enzymes and nucleic acids are the evolutionary outcome of multistep autocatalytic metabolic cycles surface-bounded to FeS/FeS_2 or some other mineral. In fact, 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 [2].

Summary and conclusions

As the initially molten young Earth cooled down, global temperatures of 100 °C must have been reached but could not have persisted for more than 20 million years [42]. Deep-sea hydrothermal vents and other local high-temperature milieus have existed throughout the history of the planet and have played a major role in shaping the early environments. However, the rates of thermal decomposition of amino acids, nucleobases, and genetic polymers are very short on the geological time scale and argue against a hot origin of life in such extreme environments.

Since high salt concentrations protect DNA and RNA against heat-induced damage [26, 46], this and other non-biological mechanisms, such as adsorption to minerals surfaces and formation of clay–nucleic acid complexes [15] might have played a significant role in the preservation of organic compounds and genetic polymers in the primitive environments. However, such mechanisms would be inefficient at temperatures above 100 °C. Because adsorption involves the formation of weak noncovalent bonds, mineral-based concentration

and protection would have been most effective at low temperatures [43]; at high temperatures any adsorbed monomers would drift away into the surrounding aqueous environment and become hydrolyzed. However, some minerals could also have the opposite effect: as shown by the Cu^{+2}-montmorillonite catalyzed decomposition of adenine to hypoxanthine [45], the association of organic compounds with some minerals may in fact reduce their half-lives.

If hyperthermophily is not truly primordial, then heat-loving lifestyles may be relics of a secondary adaptation that evolved after the origin of life and before or soon after the separation of the major lineages. As argued here, the so-called root of universal trees does not correspond to the first living system, but is the tip of a trunk of still undetermined length in which the history of a long (but not necessarily slow) series of archaic evolutionary events such as an explosion of gene families and multiple events of lateral gene transfer are still preserved. Is it possible that traces of the emergence of hyperthermophily persist in the molecular records of earliest biological evolution somewhere along the trunk of rRNA-based phylogenic trees? If hyperthermophiles were not the first organisms, then their basal position in molecular trees could be explained as: (a) a relic from early archean high-temperature regimes that may have resulted from a severe impact regime [17, 41]; (b) adaptation of Bacteria to extreme environments by lateral transfer of reverse gyrase [14] and other thermoadaptative traits from heat-loving Archaea; and (c) outcompetition of older mesophiles by hyperthermophiles originally adapted to stress-inducing conditions other than high temperatures [29].

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. Like vegetation in a mangrove, the roots of universal phylogenetic trees are submerged in the muddy waters of the prebiotic broth, but how the transition from the non-living to the living took place is still unknown. Given the huge gap existing in current descriptions of the evolutionary transition between the prebiotic synthesis of biochemical compounds and the LCA of all extant living beings, it is probably naive to attempt to describe the origin of life and the nature of the first living systems from molecular phylogenies. A high-temperature origin of life may be possible, but if this was the case then it could have not involved the usual purines and pyrimidines, or other biochemical monomers.

Acknowledgements AL is an affiliate of the NSCORT-University of California, San Diego. This paper was completed during a sabbatical leave of absence in which one of us (AL)

enjoyed the hospitality of Stanley L. Miller and his associates at the University of California, San Diego. Support from the National Aeronautics and Space Administration Specialized Center of Research and Training in Exobiology (NSCORT) is gratefully acknowledged.

References

1. Achenbach-Richter L, Gupta R, Stetter KO, Woese CR (1987) Were the original eubacteria thermophiles? System Appl Microbiol 9:34–39
2. Bada JL, Lazcano A (2002) Some like it hot, but not biomolecules. Science 296:1982–1983
3. Brasier MD, Green OR, Jephcoat AP, Kleppe AK, van Kranendonk MJ, Lindsay JF, Steele A, Grassineau NV (2002) Questioning the evidence for Earth's oldest fossils. Nature 416:76–81
4. Brochier C, Philippe H (2002) A non-hyperthermophilic ancestor for Bacteria. Nature 417:244
5. 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
6. Cody GD, Boctor NZ, Filley TR, Hazen RM, Scott JH, Sharma A, Yoder HS Jr. (2000) Primordial carbonylated iron-sulfur compounds and the synthesis of pyruvate. Science 289:1337–1340
7. Corliss JB, Baross JA, Hoffman SE (1981) An hypothesis concerning the relationship between submarine hot springs and the origin of life on Earth. Oceanologica Acta (Suppl.) 4:59–69
8. Delaye L, Becerra A, Lazcano A (2002) The nature of the last common ancestor. In: Ribas de Pouplana L (ed) The genetic code and the origin of life. Landes Bioscience, Georgetown (in press)
9. Di Giulio M (2000) The universal ancestor lived in a thermophilic or hyperthermophilic environment. J Theor Biol 203:203–213
10. Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124–2129
11. Edgell RD Doolittle WF (1997) Archaea and the origin(s) of DNA replication proteins. Cell 89:995–998
12. 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 (2002) Astrophysical and astrochemical insights into the origin of life. Reports Prog Phys 65:1427–1487
13. Forterre P (2002) A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. Trends Genet 18:236–237
14. Forterre P, Bouthier de la Tour C, Philippe H, Duguet M (2000) Reverse gyrase from hyperthermophiles: probable transfer of a thermoadaptation trait from Archaea to Bacteria. Trends Genet 16:152–154
15. Franchi M, Bramanti E, Morassi Bonzi LM, Orioli PL, Vettori C, Gallori E (1999) Clay-nucleic acid complexes: characteristics and implications for the preservation of genetic material in primeval habitats. Origins Life Evol Biosph 29:297–315
16. Galtier N, Tourasse N, Gouy M (1999) A nonhyperthermophilic common ancestor to extant life forms. Science 283:220–221
17. Gogarten-Bockels M, Hilario E, Gogarten JP (1994) The effects of heavy meteorite bombardment on the early evolution of life—a new look at the molecular record. Origins Life Evol Biosph 25:78–83
18. Grogan DW (1998) Hyperthermophiles and the problem of DNA instability. Mol Microbiol 28:1043–1049
19. Holm NG (ed) (1992) Marine hydrothermal systems and the origin of life. Kluwer Academic, Dordrecht
20. Huber C, Wächtershäuser G (1998) Peptides by activation of amino acids with CO on (Ni, Fe)S surfaces: implications for the origin of life. Science 281:670–672

21. Karlin S, Mrázek J (1998) Prokaryotic genome-wide comparisons and evolutionary implications. In: de Bruijn FJ, Lupski JR, Weinstock GM (eds) Bacterial Genomes: physical structure and analysis. Kluwer Academic , Boston, pp 196–212

22. Lamarck JB (1804) Zoological Philosophy: an exposition with regard to the Natural History of Animals. The University of Chicago Press, Chicago (Translated 1984), 458 pp

23. Levy M, Miller SL (1998) The stability of the RNA bases: implications for the origin of life. Proc Natl Acad Sci USA 95:7933–7938

24. Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362:709–715

25. Maden BEH (1995) No soup for starters? Autotrophy and the origins of metabolism. Trends Biochem Sci 20:337–341

26. Marguet E, Forterre P (1994) DNA stability at temperatures typical for hyperthermophiles. Nucleic Acid Res 22:1681–1686

27. Miller SL (1953) A production of amino acids under possible primitive Earth conditions. Science 117:528

28. Miller SL, Bada JL (1988) Submarine hot springs and the origin of life. Nature 334:609–611

29. Miller SL, Lazcano A (1995) The origin of life—did it occur at high temperatures? J Mol Evol 41:689–692

30. Miller SL, Lazcano A (2002) Formation of the building blocks of life. In: Schopf JW (ed) Life's origin: the beginnings of biological evolution. California University Press, Berkeley, pp 78–112

31. Moxon ER (1999) Whole-genome analysis of pathogens. In: Stearns SC (ed) Evolution in health and disease. Oxford University Press, New York pp 191–204

32. Nisbet EG, Sleep NH (2001) The habitat and nature of early life. Nature 409:1083–1091

33. Oró J (1960) Synthesis of adenine from ammonium cyanide. Biochem Biophys Res Commun 2:407–412

34. Pace NR (1991) Origin of life—facing up to the physical setting. Cell 65:531–533

35. Palmer T (1995) Understanding enzymes. Prentice Hall, Hertfordshire

36. Robertson MP, Miller SL (1995) An efficient prebiotic synthesis of cytosine and uracil. Nature 375:772–774

37. Schopf JW (ed) (1983) Earth's earliest biosphere: its origin and evolution. Princeton University Press, Princeton

38. Schopf JW (1993) Microfossils of the early Archaean Apex chert: new evidence of the antiquity of life. Science 260:640–646

39. Shapiro R (1995) The prebiotic role of adenine: a critical analysis. Origins Life Evol Biosph 25:83–98

40. Shimkets LJ (1998) Structure and sizes of the genomes of the Archaea and Bacteria. In: de Bruijn FJ, Lupski JR, Weinstock GM (eds) Bacterial genomes: physical structure and analysis. Kluwer Academic, Boston, pp 5–11

41. Sleep NH, Zahnle KJ, Kastings JF, Morowitz HJ (1989) Annihilation of ecosystems by large asteroid impacts on the early Earth. Nature 342:139–142

42. Sleep NH, Zahnle K, Neuhoff PS (2001) Initiation of clement surface conditions on the earliest Earth. Proc Natl Acad Sci USA 98:3666–3672

43. Sowerby SJ, Mörth C-M, Holm NG (2001) Effect of temperature on the adsorption of adenine. Astrobiology 1:481–487

44. Stetter KO (1994) The lesson of archaebacteria. In: Bengtson S (ed) Early life on earth: Nobel symposium no. 84. Columbia University Press, New York, pp 114–122

45. Strasak M, Sersen F (1991) An unusual reaction of adenine and adenosine on montmorillonite: a new way of prebiotic synthesis of some purine nucleotides? Naturwissenschaften 78:121–122

46. Tehei M, Franzetti B, Maurel M-C, Vergne J, Hountondji C, Zaccai G (2002) The search for traces of life: the protective effect of salt on biological macromolecules. Extremophiles 6:427–430

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

48. Tekaia F, Yeramian E, Dujon B (2002) Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene 297:51–60

49. Van Kranendonk MJ (2002) The flourishing of early life on Earth at hydrothermal vents: geological evidence from the 3.49–3.43 Ga Warrawoona Group, Pilbara Craton, Western Australia. Abstracts of the IAU Symposium 213 Bioastronomy 2002: Life among the stars. Australian Centre for Astrobiology, Hamilton Island, Great Barrier Reef, Australia, July 8–12, 2002, p 33

50. Wächtershäuser G (1988) Before enzymes and templates: theory of surface metabolism. Microbiol Rev 52:452–484

51. White RH (1984) Hydrolytic stability of biomolecules at high temperatures and its implication for life at 250 °C. Nature 310:430–432

52. Wilde SA, Valley JW, Peck WH, Graham CM (2001) Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. Nature 409:175–178

53. Woese CR (2002) On the evolution of cells. Proc Natl Acad Sci USA 99:8742–8747

54. Wootton J, Federhen S (1993) Statistics of local complexity in amino acid sequence and sequence database. Compt Chem 17:149–163

55. Zhang J (2000) Protein-length distributions for the three domains of life. Trends Genet 16:107–109

FACULTAD DE CIENCIAS
Departamento de Biología Evolutiva
E-mail: alar@correo.unam.mx
FAX 525/55/622.4828

February 7, 2003

Re: "Comparative genomics and the gene complement of a minimal cell", by S. Islas, A. Becerra, P.L. Luisi and A. Lazcano

Prof. Dr. Peter Walde
Institute of Polymers
Materials Science
ETH-Zentrum, CNB D 90.2
Universitaetstrasse 6
CH-8092 Zurich
Switzerland
Tel: +41-1-63 204 73 or +41-1-63 230 53

Dear Peter,

Enclosed please find the printed copies of the corrected version of the manuscript "Comparative genomics and the gene complement of a minimal cell", by Sara Islas, Arturo Becerra, Pier Luigi Luisi and Antonio Lazcano, together with a diskette in which the corresponding file may be found in Word.

In preparing this version we have (a) rewritten the references following the format required by *Origins of Life and Evolution of the Biosphere*; (b) we have added a column in Table 2 in which the number of ORFs found in every one of the genomes listed there has been added, in order to facilitate the comprehension of the column in which % of redundancies is shown; and (c) have added a short explanation in the Material and Methods section on the way the % of redundancies was estimated.

I trust you will find everything in order, but please feel free to contact me in any further information is required.

With warmest personal regards and many thanks indeed,

Antonio Lazcano
Professor

enclosures

Comparative genomics and the gene complement of a minimal cell

Sara Islas[1], Arturo Becerra[1], P. Luigi Luisi[2] and Antonio Lazcano[1]*

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

[2]ETH-Zentrum
Institut fur Polymere
Universitatstrasse 6
CH-8092 Zurich, Switzerland

*corresponding author

Abstract

The concept of a minimal cell is discussed from the viewpoint of comparative genomics. Analysis of published DNA content values determined for 641 different archaeal and bacterial species by pulsed field gel electrophoresis has lead to a more precise definition of the genome size ranges of free-living and host-associated organisms. DNA content is not an indicator of phylogenetic position. However, the smallest genomes in our sample do not have a random distribution in rRNA-based evolutionary trees, and are found mostly in (a) the basal branches of the tree where thermophiles are located; and (b) in late clades, such as those of Gram positive bacteria. While the smallest-known genome size for an endosymbiont is only 450 kb, no free-living prokaryote has been described to have genomes < 1450 kb. Estimates of the size of minimal gene complement can provide important insights on the primary biological functions required for a sustainable, reproducing cell nowadays and throughout evolutionary times, but definitions of the minimum cell is dependent on specific environments.

Key words: minimum gene set, minimal cellular genomes, genetic redundancy, DNA content

I. Introduction

Definition of the properties of a minimal cell is a notoriously complex question which is related not only to the understanding of the essential properties of a living system, but is also germane to the issue of the origin of life and early stages of cellular evolution. Several different, complementary approaches to this problem are already feasible or may be available in the near future, including the development of experimental systems based on populations of replicating polymers such as RNA molecules (Joyce, 2002), the *in vitro* synthesis of artificial cells which can metabolize, multiply and adapt (Szostak et al., 2001; Pohorille and Deamer, 2002), the empirical characterization of intracellular endosymbionts and obligate parasites with highly streamlined genomes (Morowitz,1967; Morowitz and Wallace, 1973; Mira et al., 2001; Gil et al., 2002), the trimming of extant prokaryotic genomes by knock-out experiments and transposon mutagenesis (Itaya, 1995; Hutchinson et al., 1999), and the recently advertised attempt to design a novel form of life with a completely artificial genome (Marshall, 2002).

The characteristics of a minimal cell may be inferred from the existence of the basic components required for reproduction and self-maintenance under given environmental conditions (Luisi et al., 2002). From the viewpoint of comparative genomics, the characterization of a minimal cell is equivalent to the identification of the minimum number of genes required by an unicellular organism. Such estimates can provide important insights on the primary biological functions required for a sustainable, reproducing cell nowadays and throughout evolutionary times. However, the definition of minimal genome is determined to a considerable extent by the specific

environment in which the presumed minimal cell is found (Space Science Board/National Research Council, 1999; Riley and Serres, 2000). Free-living, unicellular organisms may exist with genomes smaller than the 1.45 Mb lower-limit exhibited by extant prokaryotes (see below), but all the available evidence suggests that nowadays reduced, highly-streamlined genomes like those of *Buchnera* and the mycoplasma are viable only under the permissive, nutrient-rich, stable intracellular environment of their hosts (Mira et al., 2001; Gil et al., 2002). However, the situation must have been different during the earliest stages of biological evolution, when it is assumed that simpler, free-living cells with genomes even smaller than those of *Buchnera* and *Mycoplasma genitalium* must have proliferated.

A minimal gene set can be estimated by the presence or absence of homologous genes based on whole-genome computational sequence comparisons (Mushegian and Koonin, 1996) and, similarly, by the determination of the set of sequences shared among fully sequenced proteomes, i.e., the universal families protein families (Kyrpides et al., 1999; Hutchinson et al., 1999). Significant variations may exist between the lenghts of prokaryotic genes (Tekaia et al., 2002). However, on a first approximation bacterial genes may be considered of similar size and tightly packed, i.e., the number of prokaryotic genes is proportional to genome size (Casjens, 1998). Hence, additional insights on the minimal amount of DNA required by extant cells may also be achieved by an statistical analysis of prokaryotic genome sizes (Herdman, 1985; Casjens, 1998; Shimkest, 1998). Previous attempts to analyze the distribution of bacterial DNA content were based on a sample of 603 prokaryotic genome sizes derived by different methodologies, such as renaturation kinetics and colorimetric techniques

(Herdman, 1985), which have very different degrees of accuracy. With the development of pulse-field gel electrophoresis (PFGE), a technique that allows the separation and analysis of large DNA fragments and the direct study of the physical structure of genomes, however, the accuracy in the determination of genome sizes has been significantly improved. Here we report the results of an analysis of a database of 641 prokaryote genome sizes determined by PFGE that we have compiled from the published literature, and discuss its significance in providing insights on a minimal cellular genome. The approach developed here is very similar to that reported by Shimkets (1998), and may be considered complementary. We also discuss here how the high levels of genetic redundancy detected in all sequenced genomes can be used to obtain insights in simpler living systems without the large sets of enzymes and the sophisticated regulatory abilities of contemporary organisms, that are hypothesized to have existed prior to the divergence of the three major domains, which lacked.

2. Material and methods

A genome size database has been constructed with the 641 prokaryotic DNA content values determined by PFGE reported in publications included in the NCBI/PubMed database (http://www.ncbi.nlm.nih.gov/PubMed/) as of November 2002. The organisms in this database have been divided into four major groups: (i) free-living Archaea and Bacteria (including pathogens and symbionts that remain separate and have free-living stages); (ii) thermophilic prokaryotes (optimal growth temperature > 45 °C); (iii) obligate parasites; and (iv) endosymbionts, excluding mitochondria and chloroplasts. The

information was completed with the phylogenetic position (not shown) and lifestyle of each organism, based both on the original reports and on data from the *Bergey's Manual of Bacterial Determination* (Holt et al., 1994). The database is periodically updated and is available upon request. We have estimated the levels of genetic redundancy in the smallest genomes of endosymbionts and obligate parasites using the database of levels of paralogy (Total Proteins Hits) available from the Institute for Genomic Research (TIGR, http://www.tigr.org). To be considered redundant, all the ORFs in a given genome, whether annotated or not, were compared using BLAST and had to exhibit at least 60% sequence similarity ($P<0.0001$). The result of this comparison is shown in Table 2, where the sizes of some of the smallest known cellular genomes are indicated in kb, together with the number of ORFs, the number of redundants found in each genome, and the corresponding percentage per genome

3. Results

The genome size distribution in our database is shown in Figure 1. The values of DNA content of free-living prokaryotes can vary over a tenfold range, from *Halomonas halmophila*, a moderately halophilic gamma proteobacteria endowed with a small 1450 kb genome (Mellado et al., 1998), to the 9700 kb genome of *Azospirillium lipoferum* Sp59b (Martin-Didonet et al., 2000). The widest range of genome sizes is exhibited by the proteobacteria, from the 450 kb *Buchnera* genome, to the largest ones in the sample, which correspond to aerobic organisms with complex life cycles which can include formation of spores and mycelia. There are not reports of

archaeal genomes as large as those of *Azospirillum* and *Stigmatella*, perhaps due to incomplete sampling. All the archaeal genomes in our sample are small and fall within the 500 to 5100 kb range. These size range corresponds in fact to those of thermophilic bacterial and archaeal genomes, were the lower and upper limits appear to correspond to extreme cases, i.e., the 500 kb chromosome of the thermophilic ectosymbiont *Nanoarchaeon equitans* (Hubert et al., 2002), and the 5100 kb of the facultative thermophilic *Methanosarcina acetivorans* (Sowers et al., 1988).

Classification of endosymbionts as a group by themselves shows that although their genome size distribution overlaps with that of obligate parasites (Figure 1), their DNA content can reach values significantly smaller that those of the smallest parasites, i.e., the mycoplasma. The smallest-known cellular genome is only 450 kb and corresponds to the obligate endosymbiont proteobacterium *Buchnera* spp. (Gil et al., 2002), significantly smaller than the lower limit of 580 kb of the Mollicutes, which corresponds to the obligate parasite *Mycoplasma genitalium* (Fraser et al., 1995). Other groups with reduced genome sizes are the rickettsia and several spirochaete. The DNA content values of other obligate parasites and organisms with stringent growth conditions, which we have grouped with the mycoplasma, however, can reach values as large as the 5016 kb of *Mycobacterium intracellulare* (Kim et al., 1996).

4. Discussion

The data summarized in Figure 1 is clearly biased and does not reflect in an accurate way the actual levels of prokaryotic diversity. Because of their significance in medical and economical significance in human, animal, and crop plant life, pathogens and parasites are clearly overepresented in our sample. Moreover, the overlap in the 2000 to 3000 kb region in Figure 1 of several of the categories used here to group the species in our sample shows that prokaryotes with similar genome sizes but different lifestyles can have very different complement of genes.

In spite of these limitations, the data summarized in Figure 1 provides useful insights into the evolution of prokaryotic DNA content and the size of a minimal cellular gene set. Considerable variations in DNA content may exist even within closely related bacterial species and strains (Bergthorsson and Ochman, 1995; Casjens, 1998), but as shown by the genomes of genera like *Helicobacter* and *Streptomyces*, this is not always the case (Shimkets, 1998). The size range of bacterial genome sizes are clearly less constrained than that of the archeal chromosomes. Our results also demonstrate the unsurpassed genome plasticity of the proteobacterial clade. While some members of the group like the myxobacteria have undergone major expansion of their encoding abilities adapting to oxygen-rich environments and developing complex life cycles, others like *Buchnera* have followed an opposite direction and lost considerable amounts of DNA as they adapted to an intracellular environment (Gil et al., 2002).

The thermophilic bacterial and archaeal genomes tend to be relatively small, with the lowest limit represented by the 500 kb chromosome of the thermophilic ectosymbiont *Nanoarchaeon equitans* (Huber et al., 2002). The

5100 kb genome of the facultative thermophilic *Methanosarcina acetivorans* is probably atypical. However, the size range of thermophilic genomes does not necessarily reflects a correlation between DNA content, heat-loving microbial lifestyles and antiquity, since a wide variety of mesophilic bacterial groups, including leptospira, green-sulfur bacteria, cyanobacteria, spirochaetes, fusobacteria, and actinobacteria, can also exhibit small-sized genomes.

The smallest, highly-streamlined genomes in our sample do not have a random phylogenetic distribution. The phylogenetic mapping of genome sizes on the 16/18S rRNA tree (not shown) demonstrates that the reduction of prokaryotic genome size has occurred independently multiple times in separate lineages, and persists as an end-state character with the organisms deriving essential nutrients from a host. Although endosymbionts and intracellular parasites have many features in common, including massive gene losses as they adapted into the nutrient-rich environment provided by their hosts, grouping them into two different categories allows some insights into the differences that exist between these two lifestyles (Figure 1). For instance, it is likely that the larger size of intracellular parasite genomes, as compared to those of endosymbionts, is due to the presence of genetically encoded specifically related to parasitic lifestyles, such as sequences involved in host-parasite recognition and infection mechanisms.

Figure 1 provides no support for the hypothesis that the size distribution of extant prokaryotic chromosomes is the outcome of a series of whole genome duplications that begun with an ancestral 800 kb minigenome as suggested by Wallace and Morowitz (1973) and Herdman (1985). Since there are no

known free-living prokaryotes with genomes smaller than the 1450 kb, 1500 kb, and 1530 kb of *Halomonas halmophila* (Mellado et al., 1998), *Aquifex pyrophilus* (Shao et al., 1994) and *Fervidobacterium islandicum*, respectively, the extrapolation of a normal distribution curve beyond this cut-off value does not seems justified. However, as argued by Shimkets (1998) on the basis of a smaller sample of 141 chromosomes of prokaryotes grouped as generalists and specialists, the minimum genome size for a living organism is approximately 600 kb, a figure that fits nicely with the small genomes of *Mycoplasma genitalium* and the different *Buchnera* species (Fraser et al., 1995; Gil et al., 2002). The independent, massive gene losses that these two types of bacteria have undergone suggest that their limited encoding capacities are feasible only because of their adaptation to the highly permissive intracellular environments provided by their hosts.

5. How small can viable cells be?

One of the earliest attempts to describe both in functional and evolutionary terms the minimal set of characteristics that a cell must fulfill to be considered alive was undertaken by Morowitz (1967). Based on the enzymatic components of primary metabolism whose presence he assumed was required for DNA-based cell reproduction, Morowitz estimated the size of a minimal cell that turned out to be about one-tenth smaller than mycoplasma.

As reviewed elsewhere (Luisi et al., 2002), the defining characteristics of a minimal cell now and throughout the past has been discussed by Varela et al.

(1974), Woese (1983), Oro and Lazcano (1984), Dyson (1985), Jay and Gilbert (1987), Morowitz (1992), Walde et al. (1994), Oberholzer et al (1995), Ganti (1997), and Szostak et al (2001). Perhaps not surprisingly, the rapid pace at which more and more completely sequenced cellular genomes become available has shifted the emphasis towards deducing the minimum number of protein-encoding genes required for cellular life outside a host cell and under laboratory conditions.

Following the publication of the complete genomes of *Haemophilus influenza* and *M. genitalium*, Mushegian and Koonin (1996) published the results of a detailed comparison of these two species in conjuction with the fragmentary data from other organisms then available. Once parasite-specific sequences were discarded, the final outcome was an inventory of 256 genes that according to Mushegian and Koonin resembles not only the genetic complement of the ancestor of the Gram-negative and Gram-positive lineages to which *H. influenza* and *M. genitalium*, respectively, belong, but also the amount of DNA required to sustain a modern type minimal cell under permissible conditions. Since most of the 256 sequences shared by these two organisms have eukaryotic and/or archaeal homologs, Mushegian and Koonin also discussed how this figure could be reduced to describe the genome of the last common ancestor of the Bacteria, Archaea and Eukarya, and suggested that their results could provide insights into the earliest stages of biological evolution.

As underlined by Koonin (2000), the estimated 256 minimal gene set complement derived from the comparison of the *H. influenzae* and *M. genitalium* genomes is quite similar to the values of viable minimal genome

sizes inferred by site-directed gene disruptions in *B. subtilis* (Itaya, 1995) and transposon-mediated mutagenesis knock-outs in *M. genitalium* and *M. pneumoniae* (Hutchinson et al., 1999). These figures are also consistent with the estimate that the universal family of proteins shared among fully sequenced cellular genomes comprises 324 sequences (Kyrpides et al., 1999) and, as summarized in Table 1, with the sizes of the *Buchnera* genomes (Gil et al., 2002), and the 551 kb vestigial nucleus or nucleomorph found in cryptomonads, and which is the outcome of a secondary endosymbiotic event in which a protist engulfed an already existing unicellular eukaryotic alga which was then reduced to a secondary plastid (Douglas et al., 2001).

However, considerable caution is required to avoid an overinterpretation of these different estimates. Although the backtrack methodology proposed by Mushegian and Koonin (1996) is quite straightforward, their estimates do not consider proteins that perform the same function but have different sequences (Riley and Serres, 2000), either because they have diverged beyond recognition or because they are in fact analogous. Equally important, they failed to consider polyphyletic gene losses which have been involved in the size reduction of the *M. genitalium* and *H. influenzae* genomes, and which led to the loss of purine- and pyrimidine nucleotide biosynthetic pathways, among others (Becerra et al., 1997).

As the number of fully sequenced genomes has increased, their comparison has led to smaller sets of minimum gene complements, which are now reduced to approximately 80 orthologous sequences common to all life forms (Koonin, 2000). 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 1996; Böhlke et al., 2002). If the term "universal distribution" is restricted to its most obvious sense, i.e., that of traits found in all completely sequenced genomes now available, 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). As argued elsewhere (Islas et al., submitted), such inventories include sequences that originated in different epochs, incldуing some which may have arisen in the RNA/protein world (Tekaia et al., 1999; Delaye and Lazcano, 2000; Lazcano, 2001; Anantharaman et al., 2002). Hence, the figures reported by Mushegian and Koonin (1996) and Koonin (2000) represent, at he best, lower limits of the actual size of minimal gene-encoded functions required by a cell living under highly permissive environmental conditions. Thus, such estimates do not provide accurate models for the properties of ancestral Archean genomes.

6. The search for a minimal cell: beyond genetic and functional redundancy

Recognition that the biochemical complexity of extant organisms is the outcome of process of biological evolution that started perhaps 4×10^9 years ago can lead to some inferences on smaller ancestral cells endowed with less-complex genome replication apparatus and simpler gene expression mechanisms. In spite of the structural and functional similarities between the

template-directed enzymatic synthesis of RNA and DNA, double-stranded DNA cellular genomes replicate via a large, complex array of molecular components in which proof-reading DNA polymerases play a central role. However, a number of experimental results and sequence comparisons suggest that replication of a DNA genome can be achieved with a simplified set of catalysts (Delaye et al., 2002). For instance, the RNA-primer formation is catalyzed in mitochondria not by a primase but by the organellar DNA-dependent monomeric RNA polymerase (Frick and Richardson, 2001). This suggests that a smaller set of less-specific polymerases could be functional and, in fact, may have existed during the early stages of cell evolution. Thus, a working model of a simpler DNA-cell may be envisioned in which a single ancestral polymerase, whose evolutionary vestiges appear to be present in the catalytic palm domain of the DNA pol I and its homologs such as the T7 phage RNA polymerase (Delaye et al., 2001), could play multiple roles as a DNA polymerase, a transcriptase and a primase.

Similar arguments can be advocated for a simplified version of protein synthesis requiring less components. For instance, the fact that RNA molecules are capable of perfoming by themselves all the reactions involved in peptide-bond formation suggests that protein biosynthesis evolved in an RNA world (Zhang and Cech, 1998), 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 (Nissen et al., 2000).

Additional clues to the genetic organization of primitive forms of translation involving less components are provided by paralogous genes, which are sequences that diverge not through speciation but after a duplication event. Such genetic redundancies are a common feature of all known cellular genomes, including those of the smallest described lifeforms (Table 2). Accordingly, 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 (Gavrilova et al., 1976; Spirin, 1986) strongly supports the possibility of an older ancestral protein synthesis apparatus prior to the emergence of elongation factors.

7. Concluding remarks

The properties of a minimal cell can be approached in two different but complementary directions. One possibility involves the laboratory synthesis of encapsulated cell-like systems which may eventually metabolize, multiply and adapt (Szostak et al., 2001). An alternative approach involves the study of extant minimal genomes in order to describe cells with decreasing degrees of complexity. As discussed here, the small values of DNA content found in widely separated microbial species do not represent a primitive trait, but are in fact the outcome of polyphyletic sequence losses that have occurred in recent clades. Thus, they are excellent laboratory models to study the

properties of the genetic and metabolic repertoire of minimal cells, but the information they provide on the their evolutionary predecessors, specially those that may have existed during Archaean times, is rather limited. Primitive cells were probably endowed not only with less genes, but also with less complex sequences and simpler mechanisms of gene expression.

As discussed here, an examination of the distribution of DNA content of Archaea and Bacteria complements other genomic approaches, even if our conclusions are hindered by the nature of the available information. All known organisms share a core of highly conserved, genetically-encoded features, a significant portion of which corresponds to the translation machinery and is maintained even in highly streamlined genomes such as those of Table 1. However, our methodology is hindered by the fact that prokaryotes with similar genome sizes can have very different complements of genes. Regardless of one's definition of life, the size and content of the minimal gene set required for life will be strongly determined by the environment of the minimum cell itself. The search for minimal living systems under highly permissive conditions should thus be complemented with the search for free-living prokaryotes with genomes smaller than those of *H. halmophila*, in order to understand the minimum gene content for sustaining viability. The existence of extremely reduced 55S mitochondrial ribosomes in *Caenorhadbditis elegans* (Mears et al., 2002), as compared to its 70S prokaryotic counterpart, suggest that other organisms may exist with novel or reduced version of the essential moleuclar machinery. Whether such prokaryotes exist or not is not yet known, but the current cut-off values of genome size distribution curves (Figure 1) suggest that considerable

attention should be given to the search for similar free-living prokaryotes and the sequencing of their genomes.

The experimental efforts to define the essential genes required for life under highly permissive conditions have shown mutant *M. genitalium* populations with 265 to 350 genes can growth and divide under laboratory conditions (Hutchinson et al., 1999). Extrapolation of these results to the early evolution of life may help us to understand some of the essential characteristics, but additional efforts are required for a proper understanding of the evolutionary transition between putative RNA-cells and full-flegged DNA/protein cells. Insights into such intermediate stages are provided by analysis of genetic redundancy (Table 2) and by the experimental evidence reviewed here that has demonstrated that under *in vitro* conditions protein synthesis can take place even in the absence of some of its molecular components. Indeed, the selection and maintenance of laboratory strains in which paralogous copies of highly conserved genes such as those encoding the two elongation factors involved in protein synthesis would be substituted by one single, less-specific catalyst appear to be feasible with the available experimental techniques.

Acknowledgements

The suggestions of Dr. Cesar Hernandez and the assistance of Mlle. Ana Maria Velasco are gratefully acknowledged. A.L. is an Affiliate of the NSCORT (NASA Specialized Center for Research and Training) in Exobiology at the University of California, San Diego.

Figure caption

Figure 1. Prokaryotic genome size distribution (N=641). Open boxes, free-living prokaryotes; grey boxes, obligate parasites; black boxes, thermophiles; boxes with horizontal lines, endosymbionts.

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

Bergthorsson, U. and Ochman, H.: 1995, Heterogeneity of genome sizes among natural isolates of *Escherichia coli*, *J. Bact.* **177**, 5784-5789

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

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

Casjens, S.: 1998, The diverse and dynamic structure of bacterial genomes, *Annu. Rev. Genet.* **32**, 339-377

Delaye, L., Vázquez, H., and Lazcano, A.: 2001, The cenancestor and its contemporary biological relics: the case of nucleic acid polymerases, *In* J. Chela-Flores, T. Owen, and F. Raulin (eds) *First steps in the origin of life in the Universe* (Kluwer Academic Publisher, Dordrecht, 223-230

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), 285-288

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

Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L.-T., Wu, X., Reith, M., Cavalier-Smith, T., and Maier, U.-G.: 2001, The highly reduced genome of an enslaved algal nucleus, *Nature.* **410**, 1091-1096

Dyson, F. J.: 1985, *Origins of life,* (Cambridge University Press, Cambridge)

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

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M. et al.: 1995, The minimal gene complement of *Mycoplasma genitalium*, *Science.* **270,** 397-403

Frick, D. N. and Richardson, C. C. :2001, DNA primases, *Annu. Rev. Biochem.* **70**, 39-80

Ganti, T.: 1997, Biogenesis itself, *J. Theoret. Biol.* **187**, 583-593

Gavrilova, L. P., Kostiashkina, O. E., Koteliansky, V. E., Rutkevitch, N. M., and Spirin, A. S.: 1976, Factor-free (non-enzymic) and factor-dependent systems of translation of polyuridylic acid by *Escherichia coli* ribosomes, *J. Mol. Biol.* **101**, 537-552

Gil, R., Sabater-Munoz, B., Latorre, A., Silva, F. J., and Moya, A.: 2002, Extreme genome reduction in *Buchnera* spp: toward the minimal genome needed for symbiotic life, *Proc. Natl. Acad. Sci. USA* **99**, 4454-4458

Herdman, M.: 1985, The evolution of bacterial genomes. *In* Cavalier-Smith, T. (ed), *The evolution of genome size.* (John Wiley and Sons, New York), 37-68

Holt, G., Krieg, R., Sneath, A., Staley, T., and Williams, T.: 1994, *Bergey's manual of determinative bacteriology*. 9th Ed (Williams and Wilkins)

Huber, H., Hohn, M. J., Rachel R., Fuchs, T., Wimmer, C. V., and Stetter, K.: 2002, A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont, *Nature.* **417**, 63-67

Hutchinson, C. A., Peterson, S.N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., Smith, H. O., and Venter, J. C.: 1999 Global transposon mutagenesis and a minimal mycoplasma genome, *Science* **286,** 2165-2169

Islas, S., Velasco, A. M., Becerra, A., Delaye, L., and Lazcano, A.: 2003, Hyperthermophily and the origin and earliest evolution of life, Inter. Microbiol. (submitted)

Itaya, M.: 1995, An estimation of the minimal genome size required for life, *FEBS Letters.* **362**, 257-260

Jay, D. and Gilbert, W. : 1987, Basic protein enhances the encapsulation of DNA into lipid vesicles: model for the *formation* of primordial cells, *Proc. Natl. Acad. Sci. USA* **84**, 1978-1980

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

Kim, J. R., Kang, B. S., Ko, J. H., Park, J. S., Kim, S. J., Bai, G. H., Chung, T. H., Nam, K. S., Choi, Y. K., Choi, I. S., Chung, T. W., Lee, Y. C., and Kim, C. H.: 1996, Genomic heterogeneity in clinical strains of *Mycobacterium tuberculosis, M. terraecomplex, M. gordonae, M. avium-intracellulare complex*, and *M. fortuitum* by pulsed-field gel electophoresis, *J. Biochem. Mol. Biol.* **29**, 569-573

Koonin, E. V.: 2000, How many genes can make a cell: the minimal-gene-set concept, *Annu. Rev. Genomics Human Genet.* **1**, 99-116

Kyrpides, N., Overbeek, R., and Ouzonis, C.: 1999, Universal protein families and the functional content of the last universal common ancestor, *J. Mol. Evol.* **49**, 413-423

Lazcano, A.: 2001, El último ancestro común, *In* E. Martínez Romero y J. Martínez Romero (eds) *Microbios en Línea.* (UNAM, México), 421-429

Luisi, P. L., Oberholzer, T., and Lazcano, A. :2002, The notion of a DNA minimal cell: a general discourse and some guidelines for an experimental approach, *Helv. Chim. Acta.* **85**, 1759-1777

Marshall, E.: 2002, Genetics: Venter gets down to life basics, *Science.* **298**, 1701

Martin-Didonet, M., Chubatsu, S. L., Souza, M. E., Klein, A. M., Rego, G. M., Rigo, U. L., Yates, G. M., and Pedrosa, O.: 2000 Genome structure of the genus *Azospirillum, J. Bacteriol.* **182**, 4113-4116

Mears, J. A., Cannone, J. J., Stagg, S. M., Gutell, R. R., Agrawal, R. K., and Harvey, S. C.: 2002, Modeling a minimal ribosome based on comparative sequence analysis, *J. Mol. Biol.* **321**, 215-234

Mellado, E., Garcia, M. T., Roldan, E., Nieto, J. J., and Ventosa, A.: 1998, Analysis of the genome of the Gram-negative moderate halophiles *Halomonas* and *Chromohalobacter* by using pulsed-field gel electrophoresis, *Extremophiles.* **2**, 435-438

Mira, A., Ochmanm H., and Moran, N. A.: 2001, Deletional bias and the evolution of bacterial genomes, *Trends Genet.* **17**, 589-596

Morowitz, H. J.: 1967, Biological self-replicating systems, *Prog. Theor. Biol.* **1**, 35-58

Morowitz, H. J.: 1992, *The beginnings of cellular life,* (Yale University Press, New Haven)

Morowitz, H. J. and Wallace, D. C.: 1973, Genome size and the life cycle of the mycoplasma, *Ann. N.Y. Acad. Sci.* **225**, 62-73

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

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

Oberholzer, T., Wick, R., Luisi, P. L., and Biebricher, C. K.: 1995, Enzymatic RNA replication in self-reproducing vesicles: an approach to a minimal cell, *Biochem. Biophys. Res. Comm.* **207**, 250-257

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

Oro, J. and Lazcano, A.: 1984, A minimal living system and the origin of a protocell, *Adv. Space Res.* **4**, 167-176

Pohorille, A. and Deamer, D.: 2002, Artificial cells: prospects for biotechnology, *Trends Biotech.* **20,** 123-128

Riley, M. and Serres, M. H.: 2000, Interim report on genomics of *Escherichia coli, Annu. Rev. Microbiol.* **54**, 341-411

Shao, Z., Mages, W., and Schmitt, R.: 1994, A physical map of the hyperthermophilic bacterium *Aquifex pyrophilus* chromosome. *J. Bacteriol.* **176**, 6776-6780

Shimkets, L. J.: 1998, Structure and sizes of genomes of the Archaea and Bacteria. *In* de Bruijn, F. J., Lupskin, J. R., and Weinstock, G. M. (eds), *Bacterial genomes: physical structure and analysis,* (Kluwer Academic Publishers, Boston), 5-11

Sowers, K. and Gunsaluz, R. R.: 1988, Plasmid DNA from acetotrophic methanogen *Methanosarcina acetivorans, J. Bacteriol.* **170**, 4979-4982

Space Study Board/National Research Council.: 1999, *Size limits of very small organisms,* National Research Council. (National Academy of Sciences, Washington, D.C.)

Spirin, A. S.: 1986, *Ribosome structure and protein synthesis,* (Benjamin/Cummings, Menlo Park), 414 pp.

Szostak, J. W., Bartel, D.P., and Luisi, P. L. : 2001, Synthesizing life, *Nature.* **409**, 387-390

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

Tekaia, F., Dujon, B., and Lazcano, A.: 1999, 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 c4.6, p. 53

Tekaia, F., Yeramian, E., and Dujon, B.: 2002, Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis, *Gene.* **297**, 51-60

Varela, F. J., Maturana, H. R. and Uribe, R.: 1974, Autopoiesis: the organization of living systems, its characterization, and a model, *Curr.Mod. Biol.* **5**, 187-196

Walde, P., Goto, A., Monnard, P. A., Wessicken, M., and Luisi, P. L.: 1994, Oparin's reaction revisited: enzymatic synthesis of poly(adenylic acid) in micelles and self-reproducing vesicles, *J. Am.Chem. Soc.* **116**, 7541-7547

Wallace, D. C. and Morowitz, H. J. : 1973, Genome size and evolution, *Chromosome.* **40**, 121-126

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

Zhang, B. and Cech, T. R.: 1998, Peptidyl-transferase ribozymes: trans reactions, structural characterization and ribosomal RNA-like features, *Chem. Biol.* **5**, 539-553

Figure 1. Prokaryotic genome size distribution (N=641)

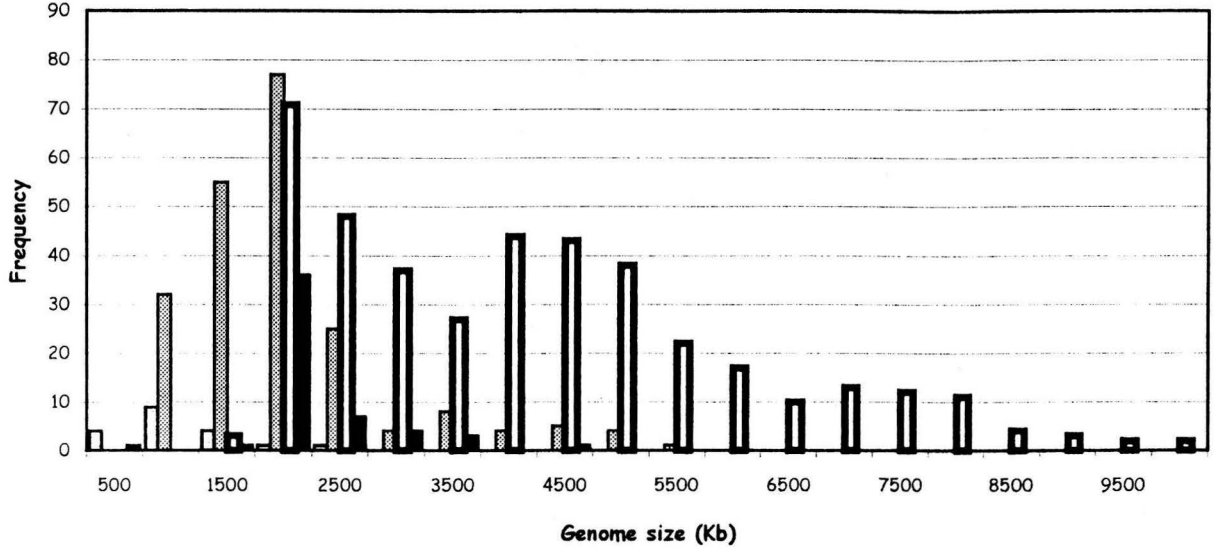


Table 1. Some miniature cellular genomes

Species	Genome size (kb)	Lifestyle	Reference
Mycoplasma genitalium	580	obligate parasite	Fraser et al., 1995
Buchnera spp.	450	endosymbiont	Gil et al., 2002
crytomonad nucleomorph	551	secondary endosymbiont	Douglas et al., 2001

Table 2. Genetic redundancies in small genomes of endosymbionts and obligate parasites*

Proteome	Genome sizes (kb)	number of ORFs	number of redundant sequences	% of redundancy
Mycoplasma genitalium	580	480	52	10.83
Mycoplasma pneumoniae	816	688	134	19.47
Buchnera sp. APS	640	574	67	11.67
Ureaplasma urealyticum	751	611	105	17.18
Chlamydia trachomatis	1000	895	60	6.71
Chlamydia muridarum	1000	920	60	6.52
Chlamydophila pneumoniae J138	1200	1070	148	13.83
Rickettsia prowazekii	1100	834	49	5.87
Rickettsia conorii	1200	1366	189	13.83
Treponema pallidum	1100	1031	78	7.56

*Genome sizes, complete proteomes, and the number of ORFs were all retrieved from NCBI http://www.ncbi.nlm.nih.gov

Conclusiones:

Diferentes aproximaciones teóricas y experimentales provenientes de diversas disciplinas han sido probadas con el propósito de reconstruir la historia evolutiva de los procariontes desde los procesos químicos que pudieron haber ocurrido antes del surgimiento de los primeros sistemas vivos, hasta los modelos encaminados a establecer los mecanismos por los cuales se transforman.

En este sentido la discusión aquí presentada se enfocó en el papel que ha jugado la duplicación total del genoma en la evolución de los procariontes y la interpretación que Wallace y Morowitz (1973) hicieron a partir de la relación que existe entre ésta y el cálculo para determinar un posible genoma ancestral; posteriormente la misma hipótesis fue retomada y corroborada por Herdman (1985).

Comparando la distribución de los tamaños de genoma de nuestra muestra con las distribuciones de los autores antes mencionados descubrimos que en ambos trabajos se encontró una distribución discontinua, mientras que los primeros aprecian valores modales en los intervalos 0.5, 1 -1.5, 2, y por arriba de 2.5 x 10^{-9} daltones, lo que les permitió sugerir que los genomas más grandes provenían de formas primitivas más pequeñas como los micoplasmas. Ahora sabemos que los micoplasmas son organismos que han reducido su genoma como resultado de una adaptación secundaria.

Para Herdman (1985) los máximos observados en la distribución de los tamaños de genoma bacterianos de su muestra son picos modales a 0.5, 1-1.5, 2.5-2.75 x 10^{-9} daltones más una cola extendida por arriba de los 4.75 y muy pocos genomas grandes; quién además infiere que los organismos actuales con metabolismo fermentativo son representantes "modernos" de aquellos que existieron cuando la atmósfera era anaeróbica y presumiblemente el contenido de los genomas era relativamente pequeño y un cambio de metabolismo anaerobio a aerobio ocurrió separadamente en cada una de las mayores líneas de descendencia bacterianas, acompañado por una o más duplicaciones totales de genoma.

Los datos mostrados aquí representan, por un lado, la diversidad procarionte (en cuanto al contenido de DNA calculado por PFGE que es una técnica más precisa para su cuantificación) estudiada hasta la fecha y rebasa las listas publicadas por (Wallace y Morowitz 1973; Sparrow y Neuman 1976; y Herdman 1985).

La distribución de la muestra en este trabajo (Apéndice 2a) no es una distribución normal (los datos tienden a concentrarse fuertemente en los primeros cuatro intervalos y con un brusco descenso que más o menos se mantiene hasta el intervalo de 5.5 y una cola extendida hasta el intervalo de 10 Mb. Resalta un valor máximo en el intervalo de 2-2.5 Mb . En nuestra muestra los organismos con tamaño de genoma más pequeño (0.448 - 0.5 Mb) corresponden a simbiontes y parásitos obligados, de los grupos Proteobacteria, Nanoarchaea, y los organismos con tamaño de genoma >0.5-1.5 Mb pertenecen a Micoplasmatales, Proteobacteria, Spirochaetales, Chlamydiales y Acholeplasmatales.

No obstante, encontramos valores del contenido de DNA de otros parásitos obligados mayores como el de *Mycobacterium intracellulare* con 5.016 Mb (Kim et al., 1996). Excluyendo todos los organismos >0.5-1.5 Mb por su dependencia obligada a un hospedero, el mínimo contenido de DNA de un organismo de vida libre se encuentra a partir de 1.45 Mb y sí nuestros datos corroboraran la hipótesis de la duplicación total del genoma en la distribución esperaríamos encontrar picos con valores modales de 3.2, 6.4, y 12.8 Mb aproximadamente y claramente este no es el caso. Además si este fuera el caso en cada barra de la distribución o por lo menos en intervalos cercanos encontraríamos una "cierta" homogeneidad filogenética que marcaría no sólo relaciones ancestrales sino una "complejidad" genómica (metabólica) creciente y constante difícil de detectar por la propia dinámica del genoma.

Por otra parte, el que en algunos procariontes haya sido observada la "posibilidad" de poder duplicar completamente el genoma temporalmente y bajo condiciones de laboratorio (Turn 1999) (manteniendo la duplicación o no) como es el caso de *Azotobacter vinelandii, Deinococcus radiodurans* y *Methanococcus jannaschii* (Bendich y Drlica 2000) no implica necesariamente que este hecho se pueda interpretar como el mecanismo único o más importante que operó en las primeras células, para la adquisición a gran escala de material genético.

En este trabajo el intervalo de tamaños de genoma procarionte disponible hasta marzo de 2003 va de 0.448 Mb (*Buchnera sp CCE)* a 9.7 Mb (*Azospirillum lipoferumSp59b*), (Islas et al 2003a ; Islas et al; 2003b). Como ya se mencionó anteriormente los procariontes más pequeños de vida libre en la muestra incluyen organismos a partir de 1.45 Mb *Halomonas halmophila* (Islas et al; b). Sin embargo, esto no quiere decir que éste dato corresponda a la cantidad de DNA mínimo que pudo haber tenido el ancestro común de todos los seres vivos, y mucho menos el "primer organismo vivo", sino que se intenta mostrar que existen características de genomas adicionales que merecen tomarse en cuenta para abordar el problema de una célula mínima. Puesto que procariontes dentro del mismo intervalo de tamaño de genoma pueden pertenecer al mismo grupo taxonómico como los Micoplasmatales, ó en un mismo intervalo de tamaño de genoma se encuentran diferentes grupos taxonómicos (Apéndice 2b) cepas bacterianas pueden variar en un amplio intervalo como *Burkholderia cepaica* 4.6-8.6 Mb se concluye que el tamaño de genoma por sí mismo no determina una correlación entre este y capacidades metabólicas específicas.

Debido a que el contenido de DNA es un atributo que representa una amplia diversidad metabólica que proviene de rutas biosintéticas comunes en algunos organismos, otros difieren en ciertos pasos enzimáticos debido a pérdidas secundarias o adiciones enzimáticas, redundancias funcionales etc. (Islas et al 1998)

Entonces el interés por evaluar la cantidad mínima de DNA de una célula corresponde a la posibilidad de ubicarla retrospectivamente en las primeras fases de la evolución de la vida y sería relevante no solo por calcular el número de genes sino para tratar de caracterizar las capacidades metabólicas que podría haber ejercido dicho "organismo" en un determinado ambiente y evaluar sus posibilidades de cambio; tal como se ha pretendido definir al último ancestro común.

El intento por describir la naturaleza del último ancestro común es un asunto que mantiene la atención de diversos grupos de trabajo en un debate continuo entre teoría y métodos. Así la determinación del cenancestro a través de la distribución de algunas enzimas biosintéticas presentes en los tres dominios Arquea, Eubacteria y Eucaria han permitido delinear un perfil del cenancestro comparable al de los modernos procariontes en cuanto a su complejidad biológica, adaptabilidad ecológica y potencial evolutivo (Lazcano 1995).Sin embargo pueden hacerse interpretaciones incorrectas al momento de cuantificar e identificar sus rasgos como a continuación se describe.

Cuando Mushegian and Koonin (1996) pretendieron detectar la cantidad de DNA requerida para mantener una célula mínima, compararon los dos genomas completamente secuenciados en ese momento *Haemophilus influenzae* y *Micoplasma genitalium,* y publicaron un inventario de 256 genes encontrándose ausentes algunas rutas biosintéticas. Estas conclusiones fueron derivadas para proponer que el último ancestro común tenía un genoma de RNA sin embargo el trabajo fue rigurosamente criticado por Becerra et al (1997) pues ambos genomas son de bacterias parásitas de humanos y han perdido gran cantidad de DNA, además la ausencia en su muestra de proteínas esenciales de eucariontes y Arqueas, involucradas en la replicación, no es evidencia para afirmar que el cenancestro tuvo un genoma de RNA.

Los estilos de vida son aspectos que pueden afectar el inventario de los genes del último ancestro común por ejemplo, la pérdida de genes, adaptaciones a microambientes intracelulares, o vida libre en ambientes muy específicos de tal forma que. El estilo de vida intracelular (de parásitos obligados y simbiontes obligados) restringe la posibilidad de adquirir genes de otros organismos vía transferencia horizontal, pudiendo también perder secuencias de inserción y secuencias relacionadas con fagos que le confieren al genoma rearreglos propios y posibilidad de cambio (Steopkowski 2001).

En lo que concierne a la respuesta al oxígeno no todos los organismos con tamaño pequeño son anaerobios, lo que puede ser explicado a través de una serie compleja de adaptaciones secundarias que han guiado a la reducción polifilética de su genoma (Becerra et al; 2000; Islas et al; 2000). Sin embargo, se encontró, que en general los procariontes anaerobios, microaerofílicos y facultativos anaerobios están dotados con genomas más pequeños que los aerobios. Los organismos con genomas pequeños no son por su propio tamaño una muestra que se pueda interpretar como formas ancestrales de vida. Así, ni Micoplasmas o Rickettsias son buenos modelos de organismos ancestrales del Arqueano, y menos aún pueden ser candidatos a representar un minigenoma a partir del cual los procariontes evolucionaron por medio de duplicaciones completas del mismo. Igualmente, el intervalo relativamente pequeño del tamaño de genoma de los hipertermófilos puede revelar una tendencia a que ambientes con altas temperaturas limitan el contenido de DNA a un intervalo específico (0.5 Mb-5.10 Mb), probablemente por la reducción del tamaño promedio de sus genes (Islas et al 2003a). Se ha sugerido que los primeros microorganismos fueron anaerobios heterótrofos, y que la disponibilidad de oxígeno promovió la aparición de nuevas capacidades metabólicas (Oparin, 1938). Nuestros resultados muestran una correlación entre tamaño de genoma y la respuesta al oxígeno, donde es evidente que los organismos anaerobios obligados microaerofilicos y facultativos

anaerobios están dotados con genomas más pequeños que aquellos procariontes aerobios, aunque hay considerable variación y traslape entre ellos (Islas et al 2000). La mayor diversidad taxonómica procarionte se encuentra en el intervalo de 1.5 a 4.0 Mb, en dónde además los cuatro tipos de respuesta al oxígeno están presentes. En contraste con otros grupos tales como las Arqueas hipertermofílicas, las proteobacterias han explotado exitosamente un amplio intervalo de tamaños de genoma, mientras algunos de sus miembros como las mixobacterias han experimentado la mayor expansión de sus capacidades codificantes adaptándose a ambientes ricos en oxígeno y desarrollando ciclos de vida complejos. Otros miembros como *Buchnera* han seguido una dirección opuesta, en una reducción máxima de su genoma debido a la pérdida de cantidades considerables de DNA que conlleva la adaptación a la vida intracelular. Se puede decir que todos los demás grupos de procariontes están dentro de este amplio intervalo.

Por otra parte, los genomas más grandes (>6.5 Mb) corresponden a bacterias de vida libre con ciclos de vida complejos los cuales deben haber evolucionado una vez que significativas cantidades de oxígeno libre llegaron a estar disponibles en el ambiente Precámbrico. En ningún caso, sin embargo, hay la evidencia que apoye que el contenido de DNA de aerobios estrictos sea el resultado de duplicaciones totales del genoma. Hasta ahora no hay reportes disponibles en la literatura de genomas de arqueas con tamaños de genoma comparables a aquellos de *Stigmatella* (Casjens, 1998). No es claro, por supuesto, si esto refleja las estrategias evolutivas del dominio Arquea, y nuestra interpretación puede estar limitada por las descripciones actuales de la diversidad procarionte. De hecho, la base de datos analizada aquí está afectada por el significado médico y económico de los organismos, y no refleja en una manera exacta el biodiversidad de los procariontes. Es decir, los organismos patógenos y parásitos están claramente sobre-representados por la importancia médica así como la trascendencia económica que representan. Sin embargo, un gran número de organismos microaerofílicos y facultativos anaerobios de diferentes grupos filogenéticos en nuestra muestra probablemente reflejan la exitosa adaptación de niveles de oxígeno cada vez mas altos en la atmósfera terrestre. Por consiguiente, se puede concluir que la expansión de ambientes aeróbicos durante el Precámbrico no sólo dirigió a la diversificación de bacterias adaptadas a condiciones novedosas, sino también al desarrollo evolutivo en lo que se refiere a la conservación de genomas grandes.

Como ya se mencionó anteriormente, es notoria una tendencia en un intervalo más o menos definido de los tamaños de genoma procariontes extremófilos que va de (0.5-5.10 Mb) que corresponden respectivamente a *Nanoarchaea equitans* y *Methanosarcina acetivorans*. No obstante, este intervalo no necesariamente expresa una correlación entre el tamaños del genoma y el estilo de vida microbiano extremo (hiper)termófilos pues existen otros grupos de procariontes que comparten esos mismos tamaños de genoma siendo mesófilos.

Un rasgo evidente es el que muestra que los genomas termófilos e hipertermófilos están dotados con secuencias génicas codificantes más pequeñas (283 $\pm$ 5.8) en comparación con los procariontes mesófilos (340 $\pm$ 9.4); sin embargo, el reducido tamaño de los genes en organismos extremófilos es un rasgo polifilético ya que esta característica es encontrada también en mesófilos de diferentes grupos.

También se encuentran, secuencias simples en organismos con ambos estilos de vida mesófilos e (hiper)termófilo, pero en los hipertermófilos, excepto *Thermoplasma acidophilum,* las secuencias simples presentan gran cantidad de ácido glutámico que son estables en condiciones de pH ácido.

Es bueno recordar que la llamada raíz del árbol universal no corresponde al primer sistema vivo, solamente refleja la punta de un tronco de tamaño indeterminado en la trayectoria de una sucesión de eventos evolutivos muy antiguos como el surgimiento de familias de genes, transferencia horizontal etc. Así la posición basal de los hipertermófilos en los árboles de rRNA se puede explicar por: a) el alto impacto al que estuvo sometido el Arqueano temprano, b) como una respuesta adaptativa de las bacterias debida a la transferencia horizontal de la reverso girasa de las arqueas y c) competencia entre mesófilos más antiguos e hipertermófilos adaptados a condiciones de altas temperaturas.

Si los genes pueden moverse de un organismo a otro vía transferencia horizontal algunos genes pudieron haberse dispersado ampliamente de tal forma que ellos pudieran detectarse como parte del último ancestro común, sin percatarnos que de hecho son más recientes .

Pero aún estamos muy lejos de entender completamente el origen y los atributos de los primeros seres vivos en cuanto a que no siempre se dispone de evidencias de rutas metabólicas, datos bioquímicos, ciclos de vida, registro paleontológico etc para integrar e interpretar correctamente los pasos y eventos que siguieron los organismos procariontes durante su evolución.

Referencias

Alifano P, Fani R, Lio P, Lazcano A, Bazzicalupo M, Carlomagno MS, Bruni CB. (1996) Histidine biosynthetic pathway and genes: structure, regulation, and evolution. Microbiol Rev. 60(1):44-69

Anderson R P y Roth RJ (1977). Tandem genetic duplications in phage and bacteria. Ann. Rev. Microbiol. 31:473-505

Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV. (1998) Evidence for masive gene exchange between archaeal and bacterial hyperthermophiles Trends Genet 14:442-444

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

Becerra A, Cocho G, Delaye L, y Lazcano A. Simple sequences: it is something you have, wheter you like it or not . enviado

Bendich AJ y Drlica K (2000) Prokaryotic and eukaryotic chromosomes : What´s the difference? BioEssays 22.: 481-486

Brasier DM, Green RO, Jephcoat PA, Kleppe KA, Van Kranendonk JM, Lindsay FJ,Steele A y Grassineau VN. (2002) Questioning the evidence for Earth´s oldest fossils. Nature 416:76-81

Bresler V, Montgomery WL, Fishelson L, Pollak PE (1998) Gigantism in a bacterium Epulopiscium fishelsoni correlates with a complex patterns in arrengements quantity and segregation of DNA. J. Bacteriol 180:5601-5611.

Cantor,CR., Smith,C.L., & Mathew, K.M. (1988) Pulsed-field gel electrophoresis of very large DNA molecules. Annu Rev Biophys Chem. 17:287-304

Casjens S, Huang W. (1993) Linear Chromosomal Physical and Genetic Map of Borrelia burgdorferi, The Lyme disease agent .Mol Microbiol 8: 967-80

Eisen AJ. (2000) Horizontal transfer among microbial genomes: new insights from complete analysis. Curr Opin Genet Dev 10:606-611

Fani, R., Mori, E., Tamburini, E. and Lazcano, A. (1998). Evolution of the structure and chromosomal distribution of histidine biosynthetic genes. Origins Life Evol. Biosph. 28: 555-570

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, et al,. (1995) The minimal gene complement of *Mycoplasma genitalium* Science 270: 397-403

Haldane J B S (1932)The causes of evolution Princenton University Press. second printing p222

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

Herdman M. (1985) The Evolution of Bacterial Genomes. In Cavalier Smith (ed) The Evolution of genome size. John Wiley, London.

Holt GJ, Krieg RN, Sneath AH, Staley TJ, Williams TS. (1994). Bergey´s Manual of determinative bacteriology . Williams & Wilkins

Islas S, Castillo A, Vázquez H G, and Lazcano A. (2000).
On the role of genome duplications in the evolution of prokaryotic chromosomes
In: Chela-Flores et al. (eds), Astrobiology, 289-292.
Kluwer Academic Publishers. Netherlands.

Islas S, Velasco A M, Becerra A, Delaye L, and Lazcano A (2003 a)
Hyperthermophily and the origin and earliest evolution of life.
International Microbiology. Aceptado para el vol. de junio

Islas S, Becerra A, Luisi Luigi P, and Lazcano A. (2003 b)
Comparative genomics and the gene complement of a minimal cell.
Enviado a Origins of lifeand evolution of the biosphere. (Enviado).

Jain R, Rivera MC, Lake JA.(1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A. 96(7):3801-6

Jain R, Rivera MC, Moore JE, Lake JA. (2002) Horizontal Gene transfer in Microbial genome evolution. Theor Popul Biol. 61:489-495

Jensen R A, (1976). Enzyme Recruitment in Evolution of new function. Annu Rev Microbiol. 30: 409-25

Kim JR, Kang, BS, Ko JH, Park JS, Kim SJ, Bai GH, Chung TH, Nam KS, Choi Y K, Choi IS, Chung T W, Lee Y C, and Kim CH (1996), Genomic heterogeneity in clinical strains of *Mycobacterium tuberculosis, M. terraecomplex, M. gordonae, M. avium-intracellulare complex*, and *M. fortuitum* by pulsed-field gel electophoresis, *J. Biochem. Mol. Biol.* 29: 569-573

Kolsto AB (1999) Time for a fresh look at the bacterial chromosome. Trend in Microbiol 7(6):223-226

Lawson et al, (1996) Phylogentic analysis of CarB genes complex evolutionary history includes an internal duplication within a gene which can root the tree of life. Mol Biol Evol 13(7): 970-977

Li. WH (1997) Molecular Evolution. Sinauer Associates, Inc., Publishers. Sunderland Mass. USA.

Lin YS, Kieser HM, Hopwood DA Chen CW. (1993) The Chromosomal DNA of *Streptomyces lividans 66* is linear Mol Microbiol 10:923-33

Margulis, L. (1993). Symbiosis in Cell Evolution, 2nd Edition. Freeman, New York.

Martin-Didonet GC, Chubatsu SL, Souza ME, Kleina M, Rego MG, Rigo UL, Yates GM, and Pedroza OF. (2000) Genome Structure of Genus Azospirilum J Bacteriol 182(14):4113 -4116

Mira A,Ochman H, Moran AN (2001) Deletional bias and the evolution of bacterial genomes Trends Genet 17 (10): 589-596

Moreno E. (1998) Genome evolution within the alpha proteobacteria: why do some bacteria not possess plasmids and others exhibit more than one different chromosome? FEMS Microbiol Rev 22: 255-275

Nelson KE, Clayton RA, Gil SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA et al (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritime. Nature 399:323-329

Neumann B, Pospiech A, Schairer HU. (1992) Size and stability of the genomes of the myxobacteria *Stigmatella aurantiaca* and *Stigmatella erecta*. J Bacteriol 174:6307-10

Ng WV. Kennedy Sp, Mahairas GG, Beruist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, Swartzell S, Weir D, Hall J Dahl TA, Welti R, Goo YA, Leithauser B, Keller K, Cruz R, Danson MJ, Hough DW, Maddocks DG, Jablonski PE, Krebs MP, Angevine CM, Dale H, Isenbarger TA, Peck RF, Pohlschroder M, Spudich JL, Jung KW, Alam M, Freitas T, Hou S, Daniels CJ, Dennis PP, Omer AD, Ebhardt H, Lowe TM, Liang P, Riley M, Hood L, DasSarma S (2000) Genome sequence of *Halobacterium species NRC-1* Proc Natl Acad Sci USA 97(22): 12176-81

Ohno, S. (1970) *Evolution by gene duplication Springer Verlag*. New York

Oparin (1938) The origin of life. MacMillan, New York. USA

Petrov A D (2001). Evolution of genome size : new approach to and old problems Trends Genet. 17(1): 23-28

Romero D, and Palacios R. (1977) Gene amplification and genome plasticity in prokaryotes Annu Rev Genet. 31:91-111

Roy H P. (1999) Horizontal transfer of genes in bacteria
Microbiology today 26: 168-170.

Schopf J.W. (1993). Microfossils of the early archean apex chert: new evidence of the antiquity of life. Science 260:640-6

Shimkets, L.J. (1998) Structure and sizes of the genomes of the Archea and Bacterial. In Bacterial genomes: Physical structure and analysis . Bruijn F,Lupski J,Weinstock G (eds) New York: Chapman & Hall

Simillon C, Vandepoele K, Van Montagu C E, Zebeau M y Van de Peer Y. (2002) The hidden duplication past of Arabidopsis thaliana. Proc Natl Acad Sci. USA 99(21):13627-13632

Soltis E D & Soltis S P (1999). Poliploidy recurrent formation and genome evolution TREE.14(9):348-352

Sparrow AH y Nauman AF. (1976) Evolution of Genome Size by DNA Doublings. Science192(4239):524-529

Stepkowski T, and Legocki A. (2001). Reduction of bacterial genome size and expansion resulting from obligate intracellular lifestyle and adaptation to soil habitat. Acta Biochem Polonica. 48(2): 367-381

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

Trevors J.T. (1996) Genome size in bacteria. *Antonie van Leeuwenhoek* 69: 293-303

Trun JN (1999) Genome Ploidy In: Bacterial Genomes Physical Structure and Analysis Brujin J. F, Lupski RJ and Weinstock MG . eds New York: Chapman & Hall

Wallace D.C. and Morowitz, H. J. (1973) Genome size and evolution. *Chromosome* 40: 121-126

Wolfe KH and Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708-713

Zipkas D. and Riley M. (1975) Proposal concerning mechanism of evolution of the genome of *Escherichia coli. Proc Natl Acad Sci USA* 72(4):1354

Database of the prokaryotic diversity according with the genome size and its relations with their metabolism and some environmental properties

Bacteria: genera and specie ; **GS:** genome size expresed in Mb; **GROUP:** each one of the main bacterial categories according to the Bergey's Manual of Bacteriology; Order: taxonomic categories belonging for each organism (NCBI); **TEMP**: Categories about the optimal grow for each microorganisms (Mesophilic: 25 - 44ºC; Thermophilic: 45 - 70ºC; Hyperthermophilic: 71 - 115ºC) **LS:** Life Style, parasites obligades simbiontes obligades and free life organisms; Oxygen response (metabolism) **REF:** references in wich the genome size is reported.

	Name	G.S.	GROUP	ORDER	TEMP	LS	MET.	REF
1	*Buchnera spCCE*	448	G-	g proteobacteria	M	SO	A	Gil R, et al (2002)
2	*Buchnera spCCU*	476	G-	g proteobacteria	M	SO	A	Gil R, et al (2002)
3	*Buchnera spCTU*	477	G-	g proteobacteria	M	SO	A	Gil R, et al (2002)
4	*Nanoarchaeum equitans*	500	A	Nanoarchaea	T	SO	ANA	Huber H, et al (2002)
5	*Buchnera spCHP*	508	G-	g proteobacteria	M	SO	A	Gil R, et al (2002)
6	*Buchnera spTHS*	544	G-	g proteobacteria	M	SO	A	Gil R, et al (2002)
7	*Buchnera spTCA*	565	G-	g proteobacteria	M	SO	A	Gil R, et al (2002)
8	*Mycoplasma genitalium*	573	M	Mycoplasmatales	M	PO	FA	Chung, J. SU and Baseman J.B . (1990).
9	*Mycoplasma arginini*	610	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G., et al (1989)
10	*Mycoplasma bovigenitalium*	610	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G., et al (1989)
11	*Buchnera sp*	645	G-	g proteobacteria	M	SO	A	Gil R, et al (2002)
12	*Buchnera spAPS*	657	G-	g proteobacteria	M	SO	A	Charles, H; and Ishikawa,H (1999)
13	*Buchnera sp*	669	G-	g proteobacteria	M	SO	A	Gil R, et al (2002)
14	*Wigglesworthia pallidipes*	705	G-	g proteobacteria	M	SO	FA	Akman L and Aksoy S (2001)
15	*Mycoplasma agalactiae*	710	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G., et al (1989)
16	*Mycoplasma orale*	710	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G., et al (1989)
17	*Mycoplasma salivatorium*	710	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G., et al (1989)

18 Wigglesworthia palpalis	710	G-	g proteobacteria	M	SO	FA	Akman L and Aksoy S (2001)
19 Mycoplasma arthritidis	720	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G., et al (1989)
20 Mycoplasma fermentans	720	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G., et al (1989)
21 Mycoplasma gallisepticum	730	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G.,et al (1989)
22 Mycoplasma hominis	740	M	Mycoplasmatales	M	PO	FA	Ladefoged, S.A. and Christiansen G. (1992).
23 Eperythrozoon suis	745	M	Mycoplasmatales	M	PO	FA	Messick JB,et al (2000)
24 Mycoplasma mobile	747	M	Mycoplasmatales	M	PO	FA	Bautsch, W. (1988).
25 Wigglesworthia brevipalis	755	G-	g proteobacteria	M	SO	FA	Akman L and Aksoy S (2001)
26 Mycoplasma pneumoniae	784	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G..et al (1989)
27 Mycoplasma hyorhinis	800	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G..et al (1989)
28 Mycoplasma M1	820	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G., et al (1989)
29 Mycoplasma sp. StraG -n 831-C4	860	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G., et al (1989)
30 Mycoplasma ellychniae	870	M	Mycoplasmatales	M	PO	FA	Weisburg,W.G., et al (1989)
31 Ehrlichia sennetsu	878	G-	a proteobacteria	M	PO	A	Rydkina E, Roux V, Roult D. (1999).
32 Ehrlichia risticii	880	G-	a proteobacteria	M	PO	A	Rydkina E, Roux V, Roult D. (1999).
33 Ureaplasma urealyticum	890	M	Mycoplasmatales	M	PO	MI	Weisburg,W.G., et al (1989)
34 Borrelia afzelii	890	G -	a proteobacteria	M	PO	MI	Casjens S. (1998)
35 Borrelia andersonii	910	G -	a proteobacteria	M	PO	MI	Casjens S. (1998)
36 Borrelia garinii	910	G -	a proteobacteria	M	PO	MI	Casjens S. (1998)
37 Borrelia japonica	910	G -	a proteobacteria	M	PO	MI	Casjens S. (1998)
38 Borrelia hermsii	920	G -	a proteobacteria	M	PO	MI	Casjens S. (1998)
39 Mycoplasma agalactiaePG2	945	M	Mycoplasmatales	M	PO	FA	Tola S, et al (2001)
40 Borrelia burdogferi	950	G -	a proteobacteria	M	PO	MI	Casjens S, Huang WM (1993)
41 Wolbachia wDim	950	G-	a proteobacteria	M	PO	A	Sun LV, et al (2001).
42 Mycoplasma flocculare	952	M	Mycoplasmatales	M	PO	FA	Robertson JA, et al (1990)
43 Spiroplasma monobial MQ1 T	980	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
44 Spiriplasma monobial MMG	995	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
45 Spiroplasma sp w115	1000	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
46 Helicobacter pyloriHP3	1040	G -	e proteobacteria	M	PO	MI	Takami S, et al (1993)
47 Mycoplasma hyopneumoniae	1045	M	Mycoplasmatales	M	PO	FA	Robertson JA, et al (1990)

48	*Chlamydia trachomatis serovar L2*	1050	G -	Chlamydiales/Verrucomicrobia	M	PO	MI	Birkelund, S. and Stephens, R.S. (1992).
49	*Treponema pallidum*	1070	G -	Spirochaetales	M	PO	MI	Walker, EM.,et al (1991).
50	*Spiroplasma monobial CUAS-1*	1080	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F. and Bové J. M. (1994).
51	*Spiroplasma sp LD-1*	1085	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
52	*Spiroplasma sabaudiense Ar-1343 t*	1095	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
53	*Wolbachia wBma*	1100	G-	a proteobacteria	M	SO	A	Sun LV, et al (2001).
54	*Rickettsia prowazeki*	1120	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
55	*Mycoplasma capricolum*	1122	M	Mycoplasmatales	M	PO	FA	Méndez-Alvarez et al (1995)
56	*Ureaplasma diversum*	1130	M	Mycoplasmatales	M	PO	MI	Kakulphimp J,Finch LR, Robertson JA. (1991).
57	*Spiroplasma sp EA-1*	1150	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
58	*Itt rikettsia ISTT CDC1*	1200	G-	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
59	*Spiroplasma taiwanense CT-1 t*	1220	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
60	*Rickettsia akariMK (Kaplan)*	1222	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
61	*Ehrlichia chaffensis arkansas*	1225	G-	a proteobacteria	M	PO	MI	Rydkina E, Roux V, Roult D. (1999).
62	*Anaplasma marginale*	1229	G -	a proteobacteria	M	PO	A	Alleman A.R., et al (1993)
63	*Rickettsia parkeri Maculatum20*	1230	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
64	*Rickettsia conorii Moroccan*	1235	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
65	*Ttt rickettsia TT-118*	1237	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
66	*Mycoplasma mycoides mycoides*	1238	M	Mycoplasmatales	M	PO	FA	Pyle, L.E. and Finch, L.R. (1988).
67	*Rickettsia sibirica232*	1243	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
68	*Rickettsia slovaca13-B*	1248	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
69	*Spiroplasma sp DF-1*	1250	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
70	*Rickettsia montana ATTCC VR611*	1250	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
71	*Rickettsia africaeESF-5*	1251	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
72	*Rickettsia rhipicephali3-7-6*	1256	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
73	*Ehrlichia chaffensis 91HE17*	1262	G-	a proteobacteria	M	PO	MI	Rydkina E, Roux V, Roult D. (1999).
74	*Campylobacter fetus*	1267	G -	e proteobacteria	M	FL	MI	Chang, N. and Taylor, D.E. (1990)
75	*Rickettsia australis Phillips*	1269	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
76	*Spiroplasma mirum SMCA T*	1270	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
77	*Rickettsia rickettsii Sheila Smith*	1272	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)

78	*Rickettsia japonicaYM*	1276	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
79	*Helicobacter pylori HP5*	1295	G -	e proteobacteria	M	PO	MI	Takami S, et al (1993)
80	*Spiroplasma sp DU-1*	1280	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
81	*Helicobacter hepaticus*	1300	G -	e proteobacteria	M	PO	MI	Saunders KE, McGovern KJ, Fox JG.(1997)
82	*Spiroplasma apis B31 T*	1300	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
83	*Helicobacter pyloriHP6*	1317	G -	e proteobacteria	M	PO	MI	Takami S, et al (1993)
84	*Helicobacter pyloriHP6*	1317	G -	e proteobacteria	M	PO	MI	Takami S, et al (1993)
85	*Spiroplasma culicicola AES-1 t*	1320	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
86	*Spiroplasma cantharicola CB-1 t*	1320	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
87	*Spiroplasma floricola OBMG*	1325	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
88	*Bartonella quintana*	1331	G -	g proteobacteria	M	PO	A	Roux ,V. And Raoult,D.(1995)
89	*Spiroplasma sp 1-25*	1340	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
90	*Spiroplasma sp PUP-1*	1350	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
91	*Wolbachia Me1Pop*	1360	G-	a proteobacteria	M	SO	A	Sun LV, et al (2001).
92	*Wolbachia M1*	1360	G-	a proteobacteria	M	SO	A	Sun LV, et al (2001).
93	*Wolbachia MelCS*	1360	G-	a proteobacteria	M	SO	A	Sun LV, et al (2001).
94	*Bartonella vinsonii*	1370	G -	g proteobacteria	M	PO	A	Roux ,V. And Raoult,D.(1995)
95	*Bartonella henselae*	1378	G -	g proteobacteria	M	PO	A	Roux ,V. And Raoult,D.(1995)
96	*Rickettsia masiliae Mtu 1*	1382	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
97	*Spiroplasma sp TN-1*	1390	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
98	*Rickettsia helvetica C9P9*	1397	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
99	*Chlamydia psittaci AB*	1450	G -	Chlamydiales/Verrucomicrobia	M	PO	MI	Frutos, R. et al (1989).
100	*Chlamidia psittaci 1H*	1450	G-	Chlamydiales/Verrucomicrobia	M	PO	MI	Frutos, R.et al (1989).
101	*Chlamydia psittaci 1B*	1450	G-	Chlamydiales/Verrucomicrobia	M	PO	MI	Frutos, R.et al (1989).
102	*Halomonas halmophila ATCC1971*	1450	G-	g proteobacteria	M	FL	FA	Mellado, E. et al (1998)
103	*Campylobacter laridis UA487*	1451	G-	e proteobacteria	T	FL	MI	Chang, N. and Taylor, D.E. (1990)
104	*Spiroplasma melliferum BC-3* ^{T}	1460	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
105	*Spiroplasma sp MQ4*	1480	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
106	*Ehrlichia HGE*	1494	G-	a proteobacteria	M	PO	MI	Rydkina E, Roux V, Roult D. (1999).
107	*Acholeplasma modicum*	1500	M	Acholeplasmatales	M	PO	FA	Weisburg,W.G., et al (1989)

108	*Fervidobacterium islandicum*	1535	G -	Thermotogales	T	FL	ANA	Bauman C, et al (1998)
109	*Dichelobacter nodosus*	1540	G -	g proteobacteria	M	PO	ANA	La Fontaine S, Rood JI. (1997)
110	*Acholeplasma hippikon*	1540	M	Acholeplasmatales	M	PO	FA	Neimark HC, Lange CS.(1990)
111	*Stygiolobus azoricus*	1543	A	Sulfolobales	T	FL	A	Bauman C, et al (1998)
112	*Porochlamydia buthi*	1550	G -	**	M	PO	A	Frutos, R. Et al (1989).
113	*Thermosipho africanus*	1550	G -	Thermotogales	T	FL	ANA	Bauman C, et al (1998)
114	*Helicobacter bizzozeronii HB13*	1558	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
115	*Helicobacter bizzozeronii HB9*	1570	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
116	*Bifidobacterium catenulatum*	1575	G+	Actinobacteria	M	FL	ANA	O' Riordan K, Fitzgerald FG (1997)
117	*Cowdria ruminantium*	1576	G-	a proteobacteria	M	PO	A	De Villiers EP,et al (2000)
118	*Spiroplasma kunkelii E-275T*	1580	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
119	*Spiroplasma chinese CCH t*	1580	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
120	*Spiroplasma sp 27F*	1580	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
121	*Helicobacter felis Dog*	1585	G-	e proteobacteria	M	PO	MI	Jalava K, et al (1999)
122	*Bifidobacterium breve225*	1585	G+	Actinobacteria	M	FL	ANA	O' Riordan K, Fitzgerald FG (1997)
123	*Helicobacter bizzozeronii HB16*	1595	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
124	*Acholeplasma florum*	1600	M	Acholeplasmatales	M	PO	FA	Weisburg,W.G., et al (1989)
125	*Coxiella burnetii.*	1600	G -	g proteobacteria	M	PO	A	Heinzen ,R., et al (1990)
126	*Aquifex pyrophilus*	1600	G -	Aquaficales	H	FL	A	Shao Z, Mages W, Schitt R. (1994)
127	*Bartonella bacilliformis*	1600	G -	g proteobacteria	M	PO	A	Krueger CM, Marks KL, Ihler GM. (1995).
128	*Bifidobacterium pseudocatenulatum*	1600	G+	Actinobacteria	M	FL	ANA	O' Riordan K, Fitzgerald FG (1997)
129	*Helicobacter felisCS7*	1603	G-	e proteobacteria	M	PO	MI	Jalava K,et al (1999)
130	*Helicobacter pylori763*	1608	G -	e proteobacteria	M	PO	MI	Taylor, D.E.(1992).
131	*Methanobacterium thermoautotrophicum*	1623	A	Methanobacteriales	T	FL	ANA	Stettler,R., Leisinger, T. (1992)
132	*Helicobacter pyloriHP9*	1625	G -	e proteobacteria	M	PO	MI	Takami S, et al (1993)
133	*Pyrodictium abyssi*	1627	A	Pyrodictiales	H	FL	ANA	Bauman C, et al (1998)
134	*Helicobacter felis Into*	1628	G-	e proteobacteria	M	PO	MI	Jalava K, et al (1999)
135	*Acholeplasma oculii ISM1499*	1630	M	Acholeplasmatales	M	PO	FA	Tigges E, Minion Fc. (1994).
136	*Bifidobacterium angulatum*	1635	G+	Actinobacteria	M	FL	ANA	O' Riordan K, Fitzgerald FG (1997)
137	*Helicobacter felisLoki 13*	1635	G-	e proteobacteria	M	PO	MI	Jalava K, et al (1999)

138 Helicobacter pylori830	1639	G -	e proteobacteria	M	PO	MI	Taylor, D.E. et al (1992).
139 Helicobacter pylori832	1640	G -	e proteobacteria	M	PO	MI	Taylor, D.E. et al (1992).
140 Campylobacter laridis JCM2530T	1645	G -	e proteobacteria	T	FL	MI	Matsumoto, K., Matsuda, M., and Kaneuchi, Ch. (1992).
141 Bartonella elizabethae	1648	G-	g proteobacteria	M	PO	A	Roux ,V. And Raoult,D.(1995)
142 Anaeroplasma abactoclasticum	1650	M	Acholeplasmatales	M	PO	ANA	Weisburg,W.G., et al (1989)
143 Helicobacter pylori829	1651	G -	e proteobacteria	M	PO	MI	Taylor, D.E. et al (1992).
144 Helicobacter pyloriHP8	1653	G -	e proteobacteria	M	PO	MI	Takami S, et al (1993)
145 Helicobacter pylori765	1657	G -	e proteobacteria	M	PO	MI	Taylor, D.E. et al (1992).
146 Methanococcus igneus	1658	A	Methanococcales	H	FL	ANA	Bauman C, et al (1998)
147 Helicobacter bizzozeronii HB5	1659	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
148 Spiroplasma clarkii CN-5 1	1660	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F. and Bové J. M. (1994).
149 Rickettsia bellii369L42-1	1660	G -	a proteobacteria	M	PO	A	Roux, V.Raoult, D (1993)
150 Wolbachia wRi	1660	G-	a proteobacteria	M	SO	A	Sun L.V, et al (2001).
151 Helicobacter pylori803	1666	G -	e proteobacteria	M	PO	MI	Taylor, D.E. et al (1992).
152 Helicobacter felis CSI	1670	G-	e proteobacteria	M	PO	MI	Jalava K, et al . (1999)
153 Gardnercla vaginalis	1670	G +	Actinobacteria	M	PO	FA	Lim,D.,Trivedi,H.,Nath,k. (1994)
154 Helicobacter pylori	1670	G -	e proteobacteria	M	PO	MI	Taylor, D.E. et al (1992).
155 Helicobacter Salomonis HS3	1675	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
156 Helicobacter Salomonis CCUG37845 (HS 4)	1679	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
157 Acholeplasma laidlawii	1680	M	Acholeplasmatales	M	PO	MI	Weisburg,W.G., et al (1989)
158 Taylorella equigenitalis	1682	G-	b proteobacteria	M	PO	FA	Matsuda M, et al (1994)
159 Helicobacter pylori 823	1693	G -	e proteobacteria	M	PO	MI	Taylor, D.E. et al (1992).
160 Helicobacter mustelae	1700	G -	e proteobacteria	M	PO	MI	Taylor DE, et al (1994)
161 Streptococcus thermophilus ST1	1700	G +	Bacillus Clostridium	M	FL	FA	Le Burgeois, P., Mata, M., and Ritzenthaler. (1989)
162 Campylobacter coli UA417R	1700	G -	e proteobacteria	M	FL	MI	Yan, W., and Taylor, D.E. (1991).
163 Rochalimaea quintana	1700	G-	g proteobacteria	M	PO	A	Resche D.K., Frazier M. E., Mallavia L.P. (1991).
164 Helicobacter pyloriNCTC11637	1702	G -	e proteobacteria	M	PO	MI	Takami S, et al (1993)
165 Helicobacter pyloriHP4	1705	G -	e proteobacteria	M	PO	MI	Takami S, et al (1993)
166 Helicobacter bizzozeronii HB17	1706	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
167 Pyrobaculum aerophilum	1709	A	Thermoproteales	H	FL	ANA	Bauman C, et al (1998)

168	*Helicobacter pylori800*	1710	G -	e proteobacteria	M	PO	MI	Taylor, D.E. et al (1992).
169	*Helicobacter pylori802*	1710	G -	e proteobacteria	M	PO	MI	Taylor, D.E. et al (1992).
170	*Helicobacter pylori 844*	1711	G -	e proteobacteria	M	PO	MI	Taylor, D.E.et al (1992).
171	*Bifidobacterium sp 3612*	1713	G+	Actinobacteria	M	FL	ANA	O' Riordan K, Fitzgerald FG (1997)
172	*Campylobacter coli UA578*	1714	G-	e proteobacteria	M	FL	MI	Chang, N. and Taylor, D.E. (1990)
173	*Rickettsiella melolonthae*	1717	G -	a proteobacteria	M	PO	A	Frutos, R. Et al (1989).
174	*Streptococci SF465*	1718	G+	Bacillus/Clostridium	M	PO	FA	Suvorov AN and Ferretti JJ (1997)
175	*Asteroleplasma anaerobium*	1720	M	Unclassified Molli	M	PO	ANA	Weisburg.W.G., et al (1989)
176	*Streptococci SF365 M12*	1723	G+	Bacillus/Clostridium	M	PO	FA	Suvorov AN and Ferretti JJ (1997)
177	*Thermus thermophilusHB28*	1730	G-	Thermus/Deinococcus	T	FL	A	Moreira,M.M., Da Costa,S.M.,Sá-Correira Isabel.(1997)
178	*Campilobacter jejuniNCTC11168*	1730	G-	e proteobacteria	M	FL	MI	Karlyshev V A, et al (1998).
179	*Streptococci SF451M1*	1733	G+	Bacillus/Clostridium	M	PO	FA	Suvorov AN and Ferretti JJ (1997)
180	*Helicobacter bizzozeronii HB6*	1757	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
181	*Haemophillus ducreyi 35000*	1760	G -	g proteobacteria	M	PO	FA	Hobbs MM, et al (1996)
182	*Spiroplasma insolitum M-55T*	1760	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
183	*Helicobacter pyloriHP1*	1760	G -	e proteobacteria	M	PO	MI	Takami S, et al (1993)
184	*Helicobacter pyloriHP2*	1760	G -	e proteobacteria	M	PO	MI	Takami S, et al (1993)
185	*Helicobacter bizzozeronii CCUG35545 (HB1)*	1766	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
186	*Spiroplasma sp N525*	1780	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
187	*Helicobacter bizzozeronii HB14*	1780	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
188	*Helicobacter pylori*	1780	G -	e proteobacteria	M	PO	MI	Takami S, et al (1993)
189	*Helicobacter salomonis HS6*	1781	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
190	*Streptococci SF402 M18*	1783	G+	Bacillus/Clostridium	M	PO	FA	Suvorov AN and Ferretti JJ (1997)
191	*Archaeoglobus fulgidus*	1784	A	Archaeoglobales	H	FL	ANA	Bauman C, et al (1998)
192	*Streptococcus SF403M3*	1784	G+	Bacillus/Clostridium	M	PO	FA	Suvorov AN and Ferretti JJ (1997)
193	*Helicobacter salomonisHS1*	1784	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
194	*Streptococcus thermophilus054*	1791	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
195	*Helicobacter bizzozeronii HB10*	1796	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
196	*Streptococcus thermophilus ND1-6*	1797	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
197	*Helicobacter salomonisCCUG37848 (HS 8d)*	1800	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).

198	*Halobacterium salinarum (halobium) NRC817*	1800	A	Halobacteriales	M	FL	A	López-García P, Amils R and Antón J. (1996).
199	*Halobacterium salinarum (halobium) CCM2090*	1800	A	Halobacteriales	M	FL	A	López-García P, Amils R and Antón J. (1996).
200	*Halobacterium salinarum (halobium) CECT396*	1800	A	Halobacteriales	M	FL	A	López-García P, Amils R and Antón J. (1996).
201	*Helicobacter pyloriNCTC11916*	1804	G -	e proteobacteria	M	PO	MI	Takami S, et al (1993)
202	*Helicobacter bizzozeronii HB3*	1804	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
203	*Streptococcus thermophilus 030*	1807	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
204	*Prochlorococcus marinus CCMP1357*	1810	G-	Cyanobacteria	M	FL	A	Strehl B, et al (1999)
205	*Campylobacter jejuni TCH9011*	1812	G-	e proteobacteria	M	FL	MI	Kim,N.W. Et al (1992)
206	*Streptococcus thermophilus 957*	1817	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
207	*Spiroplasma citri R82A T*	1820	M	Mycoplasmatales	M	PO	FA	Carle P, et al (1995)
208	*Streptococci SF 448 M22*	1823	G+	Bacillus/Clostridium	M	PO	FA	Suvorov AN and Ferretti JJ (1997)
209	*Streptococcus thermophilus 1020*	1824	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
210	*Streptococcus thermophilus 019*	1825	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
211	*Streptococcus thermophilus 013*	1826	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
212	*Streptococcus thermophilus 985*	1828	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
213	*Acidianus infernus*	1829	A	Sulfolobales	H	FL	FA	Bauman C, et al (1998)
214	*Streptococcus thermophilus CNRZ703*	1830	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
215	*Streptococcus thermophilus 958*	1833	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
216	*Haemophilus influenzae*	1834	G -	g proteobacteria	M	PO	FA	Lee, J.J Smith HO (1988)
217	*Streptococcus thermophilus 956*	1835	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
218	*Streptococcus thermophilus 959*	1835	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
219	*Spiroplasma phoeniceum P-40 T*	1840	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
220	*Streptococcus thermophilus 026*	1841	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
221	*Streptococcus thermophilus CNRZ385*	1841	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
222	*Lactobacillus sakei23K*	1845	G+	Bacillus/Clostridium	M	FL	FA	Dudez AM, et al (2002)
223	*Streotococcus thermophilus 4ML*	1854	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
224	*Lactobacillus acidophilus*	1846	G +	Bacillus/Clostridium	M	FL	FA	Roussel Y, et al (1993)
225	*Acidianus ambivalens*	1850	A	Sulfolobales	T	FL	FA	Bauman C, et al (1998)
226	*Oenococcus oeni PSU-1*	1855	G+	Bacillus/Clostridium	M	FL	FA	Ze-Ze L, et al (1998)
227	*Metallosphaera prunae*	1857	A	Sulfolobales	H	FL	A	Bauman C, et al (1998)

228	*Streptococcus thermophilus 780*	1868	G+	Bacillus/Clostridium	T	FL	FA	O'Sullivan TF, Fitzgerald GF, (1998)
229	*Bifidobacterium infantis2255*	1870	G+	Actinobacteria	M	FL	ANA	O' Riordan K, Fitzgerald FG (1997)
230	*Acidianus brierleyi*	1879	A	Sulfolobales	T	FL	FA	Bauman C, et al (1998)
231	*Metallosphaera sedula*	1880	A	Sulfolobales	T	FL	A	Bauman C, et al (1998)
232	*Kozo Todo, Takatsugu Goto, Kazuaki Miyamoto, Shig*	1880	A	Thermococcales	H	FL	ANA	Bauman C, et al (1998)
233	*Bifidobacterium bifidum 8810*	1880	G+	Actinobacteria	M	FL	ANA	O' Riordan K, Fitzgerald FG (1997)
234	*Chromohalobacter marismortuiA-492*	1885	G-	g proteobacteria	M	FL	FA	Mellado,E. Et al (1998)
235	*Halomonas halodurans ATCC2986*	1886	G-	g proteobacteria	M	FL	FA	Mellado, E (1998)
236	*Thermococcus celer Vu13*	1890	A	Thermococcales	H	FL	ANA	Noll.M.K. 1989
237	*Archaeoglobus lithotrophicus*	1891	A	Archaeoglobales	H	FL	ANA	Bauman C, et al (1998)
238	*Campylobacter jejuni UA580*	1890	G-	e proteobacteria	M	FL	MI	Chang, N. and Taylor, D.E. (1990)
239	*Leptospirilum ferroxidans ATCC49879*	1891	G -	Nitrospira Group	M	FL	A	Amils,R., et al (1998).
240	*Streptococci SF372M1*	1893	G+	Bacillus/Clostridium	M	PO	FA	Suvorov AN and Ferretti JJ (1997)
241	*Methanococcus voltae*	1900	A	Methanococcales	M	FL	ANA	Sitzman J. and Klein A.(1991)
242	*Streptococci SF265M49*	1916	G+	Bacillus/Clostridium	M	FL	FA	Suvorov AN and Ferretti JJ (1997)
243	*Thermus thermophilusHB27*	1920	G -	Thermus/Deinococcus	T	FL	A	Tabata, K., Kosuge, T., Nakahara, T., and Hoshino, T. (1993)
244	*Streptococcus pyogenes*	1925	G+	Bacillus/Clostridium	M	PO	FA	Suvorov AN, Ferreti JJ.1996.
245	*LactobacillushelveticusCNRZ241*	1930	G +	Bacillus/Clostridium	M	FL	FA	Lortal, S., et al (1997)
246	*Lactobacillus helveticus CNRZ303*	1932	G+	Bacillus/Clostridium	M	FL	FA	Lortal, S.et al (1997)
247	*Streptococci SF370M1*	1933	G+	Bacillus/Clostridium	M	PO	FA	Suvorov AN and Ferretti JJ (1997)
248	*Sulfolobus methallicus*	1940	A	Sulfolobales	T	FL	A	Bauman C, et al (1998)
249	*Moraxella catarrhalisATCC25238*	1940	G -	g proteobacteria	M	PO	A	Furihata K, Sato K, Matsumoto H. (1995)
250	*Helicobacter bizzozeronii CCUG35046 (HB2)*	1941	G -	e proteobacteria	M	PO	MI	Hannien M and Hirvi U. (1999).
251	*Lactobacillus helveticus CIPH57.15*	1953	G+	Bacillus/Clostridium	M	FL	FA	Lortal, S., et al (1997)
252	*Thermus aquaticus*	1994	G -	Thermus/Deinococcus	T	FL	FA	Moreira,M.M., Da Costa,S.M.,Sá-Correira Isabel.(1997)
253	*Pediococcus acidilactici*	1995	G +	Bacillus/Clostridium	M	FL	FA	Casjens S. (1998)
254	*Bifidobacterium adolescentis*	1995	G+	Actinobacteria	M	FL	ANA	O' Riordan K, Fitzgerald FG (1997)
255	*Campylobacter upsaliensis*	2000	G -	e proteobacteria	M	FL	MI	Bourke B, et al (1995)
256	*Halorubrum sodomenseATCC33755*	2000	A	Halobacteriales	M	FL	A	López-García P, Amils R and Antón J. (1996).
257	*Natronobacterium pharaonisATCC35678*	2000	A	Halobacteriales	M	FL	A	López-García P, Amils R and Antón J. (1996).

258	*Lactobacillus gasseri*	2020	G +	Bacillus/Clostridium	M	FL	FA	Roussel Y, et al (1993)
259	*Spiroplasma sp LB-12*	2020	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F, and Bové J. M. (1994).
260	*Pyrococcus kodakaraensis KOD1*	2036	A	Thermococcales	H	FL	ANA	Fujiwara S, Takagi M, Imanaka T, (1998)
261	*Alteromonas nigrifaciens*	2040	G-	g proteobacteria	M	FL	A	Suzuki, S,. Kita-Tsukamoto, K., and Fukagawa, T. (1994)
262	*Thermus filiformis*	2054	G -	Thermus/Deinococcus	T	FL	A	Moreira,M.M., Da Costa,S.M.,Sá-Correira Isabel.(1997)
263	*Bifidobacterium breve CIP6469*	2058	G+	Actinobacteria	M	FL	ANA	Bourget, N., Sominet, J-M and Decaris, B. (1993)
264	*Sodalis glossinidius*	2066	G-	g proteobacteria	M	SO	MI	Akman L, et al (2001)
265	*Zymomonas mobilis*	2080	G -	a proteobacteria	M	FL	FA	Kang HL Kang HS,(1998)
266	*Bifidobacterium breve CIP64~0*	2085	G +	Actinobacteria	M	FL	ANA	Bourget, N., Sominet, J-M and Decaris, B. (1993)
267	*Rickettsiella grylli*	2087	G-	g proteobacteria	M	PO	A	Frutos, R. Et al (1989).
268	*Bartonella vinsonii*	2100	G-	g proteobacteria	M	PO	A	Resche D.K., Frazier M. E., Mallavia L.P. (1991).
269	*Bifidobacterium breve CIP6466*	2100	G+	Actinobacteria	M	FL	ANA	Bourget, N., Sominet, J-M and Decaris, B. (1993)
270	*Bifidobacterium breve ATCC156998*	2101	G+	Actinobacteria	M	FL	ANA	Bourget, N., Sominet, J-M and Decaris, B. (1993)
271	*Streptococcus mutans GS-5*	2101	G +	Bacillus Clostridium	M	PO	FA	Hantman, J.M. Et al (1993)
272	*Thermus oshimai*	2120	G-	Thermus/Deinococcus	T	FL	A	Moreira,M.M., Da Costa,S.M.,Sá-Correira Isabel.(1997)
273	*Chlorobium tepidum*	2137	G-	Green Sulfur Bacteria	M	FL	ANA	Naterstad K, Kolsto AB, Sirevag R.(1995)
274	*Bifidobacterium breveCIP6468*	2150	G+	Actinobacteria	M	FL	ANA	Bourget, N., Sominet, J-M and Decaris, B. (1993)
275	*Chromohalobacter marismortuiA-100*	2162	G-	g proteobacteria	M	FL	FA	Mellado, E. et al (1998)
276	*Thermus scotoductus1'i-~a*	2166	G-	Thermus/Deinococcus	T	FL	A	Moreira,M.M., Da Costa,S.M.,Sá-Correira Isabel.(1997)
277	*Streptococcus agalactiae*	2200	G +	Bacillus/Clostridium	M	PO	FA	Dmitriev A, Suvorov A, Totolian A (1998)
278	*Pseudomonas aeruginosaATCC33348*	2200	G-	g proteobacteria	M	FL	A	Trevors J.T.1996
279	*Pseudomonas aeruginosaATCC33361*	2200	G-	g proteobacteria	M	FL	A	Trevors J.T.1996
280	*PAO*	2200	G-	g proteobacteria	M	FL	A	Trevors J.T.1996
281	*Acetobacter xylinum*	2200	G-	a proteobacteria	M	FL	A	Dempsey JA, York J, Cannon JG (1993)
282	*Halomonas eurihalinaATCC49336*	2214	G-	g proteobacteria	M	FL	FA	Mellado, E. et al (1998)
283	*Thermus scotoductus NH*	2216	G-	Thermus/Deinococcus	T	FL	A	Moreira,M.M., Da Costa,S.M.,Sá-Correira Isabel.(1997)
284	*Spiroplasma ixodetis Y32 T*	2220	M	Mycoplasmatales	M	PO	FA	Ye, F., Laigret F. and Bové J. M. (1994).
285	*Alteromonas sp. M-1*	2240	G -	g proteobacteria	M	FL	A	Suzuki, S,. Kita-Tsukamoto, K., and Fukagawa, T. (1994)
286	*Chromohalobacter marismortuiATCC1~056*	2252	G-	g proteobacteria	M	FL	FA	Mellado,E. et al (1998)
287	*Thermus scotoductus ITI-252*	2268	G-	Thermus/Deinococcus	T	FL	A	Moreira,M.M., Da Costa,S.M.,Sá-Correira Isabel.(1997)

288	*Streptococcus pneumoniae*	2270	G +	Bacillus/Clostridium	M	PO	FA	Gasc, A. M., et al (1991)
289	*Actinobacillus pleuropneumoniae 5b L20*	2283	G-	g proteobacteria	M	PO	FA	Chevalier,B et al (1998)
290	*Halomonas eurihalinaF2-12*	2289	G-	g proteobacteria	M	FL	FA	Mellado, E. et al (1998)
291	*Neisseria gonorrheae MS11-N198*	2300	G -	b proteobacteria	M	PO	A	Bihlmaier, A., et al (1991)
292	*Neisseria meningitidis B1940*	2300	G-	b proteobacteria	M	PO	A	Bautsch W. (1993)
293	*Streptococcus sanguisBM145154*	2300	G +	Bacillus/Clostridium	M	PO	FA	Le Burgeois, P., Mata, M., and Ritzenthaler. (1989)
294	*Actinobacillus actinomycetemcomitans*	2300	G -	g proteobacteria	M	PO	FA	Valcarcel J, et al 1997
295	*Pseudomonas aeruginosaATCC33353*	2300	G-	g proteobacteria	M	FL	A	Trevors J.T.1996
296	*Pseudomonas aeruginosa ATCC33360*	2300	G-	g proteobacteria	M	FL	A	Trevors J.T.1996
297	*Lactobacillus lactis LA12301*	2300	G+	Bacillus/Clostridium	M	FL	A	Le Burgeois, P., Mata, M., and Ritzenthaler. (1989)
298	*Actinobacillus pleuropneumoniae 3 S1421*	2312	G-	g proteobacteria	M	PO	FA	Chevalier,Bet al (1998)
299	*Propionibacterium freudenreichii*	2327	G+	Actinobacteria	M	FL	FA	Méndez-Alvarez et al (1995)
300	*Haemophilus parainfluenza*	2340	G -	g proteobacteria	M	PO	FA	Kaue, L. And Goodgal, SH. (1989)
301	*Actinobacillus pleuropneumoniae 5a K17*	2352	G-	g proteobacteria	M	PO	FA	Chevalier,B et al (1998)
302	*Actinobacillus pleuropneumoniae 12 8329*	2357	G-	g proteobacteria	M	PO	FA	Chevalier,B et al (1998)
303	*Vibrio costicola*	2382	G-	g proteobacteria	M	FL	FA	Mellado Eet al (1997)
304	*Shewanella putrefaciens*	2383	G -	g proteobacteria	M	FL	ANA	Suzuki, S,. Kita-Tsukamoto, K., and Fukagawa, T. (1994)
305	*Actinobacillus pleuropneumoniae 10 13039*	2392	G-	g proteobacteria	M	PO	FA	Chevalier,B et al (1998)
306	*Actinobacillus pleuropneumoniae 11 56153*	2392	G-	g proteobacteria	M	PO	FA	Chevalier,B et al (1998)
307	*Thermus brockianus 15038*	2394	G -	Thermus/Deinococcus	T	FL	A	Moreira,M.M., Da Costa,S.M.,Sá-Correira Isabel.(1997)
308	*Actinobacillus pleuropneumoniae 2 S 1536*	2397	G-	g proteobacteria	M	PO	FA	Chevalier,B et al (1998)
309	*Halobacterium halobium NRC-1*	2400	A	Halobacteriales	M	FL	A	Bobovnikova Y, et al (1994)
310	*Fusobacterium nucleatum*	2400	G -	Fusobacteria	M	FL	ANA	Bolstad , A.I. (1994)
311	*Pseudomonas aeruginosa ATCC33349*	2400	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
312	*Pseudomonas aeruginosa ATCC33350*	2400	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
313	*Pseudomonas aeruginosaATCC33351*	2400	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
314	*Pseudomonas aeruginosa ATCC33354*	2400	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
315	*Pseudomonas aeruginosaATCC33355*	2400	G-	g proteobacteria	M	FL	A	Trevors J.T.(1996)
316	*Pseudomonas aeruginosaATCC33357*	2400	G-	g proteobacteria	M	FL	A	Trevors J.T.1996
317	*Psudomonas aeruginosaATCC33358*	2400	G-	g proteobacteria	M	FL	A	Trevors J.T.1996

318	*Actinobacillus pleuropneumoniae 4M62*	2401	G-	g proteobacteria	M	PO	FA	Chevalier,B et al (1998)
319	*Actinobacillus pleuropneumoniae 1/S 4074*	2404	G-	g proteobacteria	M	PO	FA	Chevalier,B et al (1998)
320	*Actinobacillus pleuropneumoniae 6/Fem f*	2407	G-	g proteobacteria	M	PO	FA	Chevalier,Bet al (1998)
321	*Actinobacillus pleuropneumoniae 7WF83*	2408	G-	g proteobacteria	M	PO	FA	Chevalier,B et al (1998)
322	*Actinobacillus pleuropneumoniae 8/405*	2409	G-	g proteobacteria	M	PO	FA	Chevalier,B et al (1998)
323	*Actinobacillus pleuropneumoniae 9/CVJ13261*	2416	G-	g proteobacteria	M	PO	FA	Chevalier,B et al (1998)
324	*Halomonas israelensis ATCC43985*	2490	G-	g proteobacteria	M	FL	FA	Mellado, E et al (1998)
325	*Halomonas subglaciescolaUQM2927*	2492	G-	g proteobacteria	M	FL	FA	Mellado, E. et al (1998)
326	*Halomonas elongataATCC33173*	2497	G-	g proteobacteria	M	FL	FA	Mellado, E et al (1998)
327	*Clostridium tyrobutyricum DSM1460*	2500	G+	Bacillus/Clostridium	M	FL	ANA	Young M and Cole (1993)
328	*Lactococcus lactis MG1363*	2500	G+	Bacillus/Clostridium	M	FL	A	Le Burgeois, P., Mata, M., and Ritzenthaler. (1989)
329	*Lactococcus lactis F7 2*	2500	G+	Bacillus/Clostridium	M	FL	A	Le Burgeois, P., Mata, M., and Ritzenthaler. (1989)
330	*Lactococcus lactis C2*	2500	G+	Bacillus/Clostridium	M	FL	A	Le Burgeois, P., Mata, M., and Ritzenthaler. (1989)
331	*Lactococcus lactis F166*	2500	G+	Bacillus/Clostridium	M	FL	A	Le Burgeois, P., Mata, M., and Ritzenthaler. (1989)
332	*Enterococcus faecium ATTCC35667*	2550	G+	Bacillus/Clostridium	M	FL	FA	Oana K, et al (2002)
333	*Lactococcus lactis lactis DL11*	2580	G+	Bacillus/Clostridium	M	FL	FA	Tulloch, D.L. Et al (1991)
334	*Staphylococcus carnosus*	2590	G+	Bacillus/Clostridium	M	PO	FA	Wagner E, Doskar J, Gotz F.(1998).
335	*Enterococcus faecalis JH2*	2600	G+	Bacillus/Clostridium	M	FL	FA	Le Burgeois, P., Mata, M., and Ritzenthaler. (1989)
336	*Lactococcus cremori 187*	2600	G+	Bacillus/Clostridium	M	FL	FA	Le Burgeois, P., Mata, M., and Ritzenthaler. (1989)
337	*Lactobacillus cremorisBK5*	2600	G+	Bacillus/Clostridium	M	FL	FA	Le Burgeois, P., Mata, M., and Ritzenthaler. (1989)
338	*Lactobacillus cremoris H2*	2600	G+	Bacillus/Clostridium	M	FL	FA	Le Burgeois, P., Mata, M., and Ritzenthaler. (1989)
339	*Psedomonas lemoignei ATCC17989*	2600	G-	b proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
340	*Halomonas elongata ATCC33174*	2615	G-	g proteobacteria	M	FL	FA	Mellado, E et al (1998)
341	*Enterococcus faecium ATTCC19434*	2635	G+	Bacillus/Clostridium	M	FL	FA	Oana K, et al (2002)
342	*Porochlamydia chironomi*	2650	G-	g proteobacteria	M	FL	A	Frutos, R. Et al (1989).
343	*Porochlamydia duronomi*	2650	G-	e proteobacteria	M	FL	A	Méndez-Alvarez et al (1995)
344	*Chlorobium limicola*	2650	G-	Green Sulfur Bacteria	M	FL	ANA	Méndez-Alvarez S. et al (1995)
345	*Pseudomonas aeruginosa ATCC33356*	2700	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
346	*Synechococcus sp.PCC7002*	2700	G-	Cyanobacteria	M	FL	A	Chen, X. and Widger,W, R,. (1993)
347	*Synechococcus sp.PCC6301*	2700	G-	Cyanobacteria	M	FL	A	Kaneko T, et al (1996)

348	*Pseudomonas aeruginosa ATCC33359*	2700	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
349	*Pseudomonas aeruginosa ATCC33362*	2700	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
350	*Pseudomonas aeruginosa ATCC33363*	2700	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
351	*Pseudomonas aeruginosa ATCC33364*	2700	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
352	*Sulfolobus solfataricus Ron 12/III*	2705	A	Sulfolobales	T	FL	A	Bauman C, et al (1998)
353	*Vibrio chloreae 569B reubicar*	2760	G -	g proteobacteria	M	FL	FA	Majumder, R., et al (1996)
354	*Caldocellum saccharolyticum*	2780	G+	Bacillus/Clostridium	H	FL	ANA	Borges KM, Bergquist PL. (1993)
355	*Sulfolobus acidocaldarius*	2795	A	Sulfolobales	H	FL	A	Bauman C, et al (1998)
356	*Sulfolobus solfataricus*	2800	A	Sulfolobales	T	FL	A	Bauman C,et al (1998)
357	*Pseudomonas andropogonis NCPPB934*	2800	G -	e proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
358	*Pseudomonas corrugata ATCC13525*	2800	G -	e proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
359	*Mycobacterium leprae*	2800	G +	Actinobacteria	M	PO	A	Eiglmeier K, et al (1993)
360	*Pseudomonas aeruginosa ATCC33352*	2822	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
361	*Staphylococcus aureus*	2900	G +	Bacillus Clostridium	M	PO	FA	Méndez-Alvarez et al (1995)
362	*Haloferax mediterranei*	2900	A	Halobacteriales	M	FL	A	Méndez-Alvarez et al (1995)
363	*Haloferax mediterranei ATCC33500*	2900	A	Halobacteriales	M	FL	A	López-García P. Amils R and Antón J. (1996).
364	*Halofera volcanii NCMB2012*	2900	A	Halobacteriales	M	FL	A	López-García P. Amils R and Antón J. (1996).
365	*Haloferax gibbonsii ATCC33959*	2900	A	Halobacteriales	M	FL	A	López-García P. Amils R and Antón J. (1996).
366	*Enterococcus faeciumS3*	2975	G+	Bacillus Clostridium	M	FL	FA	Oana K, et al (2002)
367	*Enterococcus faeciumM1*	2995	G+	Bacillus/Clostridium	M	FL	FA	Oana K, et al (2002)
368	*Corynebacterium glutamicum*	3000	G +	Actinobacteria	M	PO	FA	Hermann T. Et al (1998)
369	*Clostridium stereorarium NCIM13117 54*	3000	G+	Bacillus/Clostridium	M	FL	ANA	Young M and Cole (1993)
370	*Haloarcula hispanica ATCC33960*	3000	A	Halobacteriales	M	FL	A	López-García P. Amils R and Antón J. (1996).
371	*Haloarcula vallismortis ATCC29715*	3000	A	Halobacteriales	M	FL	A	López-García P, Amils R and Antón J. (1996).
372	*Enterococcus faecalis ATCC29212*	3000	G+	Bacillus/Clostridium	M	FL	FA	Oana K, et al (2002)
373	*Sulfolobus shibatae*	3015	A	Sulfolobales	H	FL	A	Bauman C, Judex M, Huber H, Wirth R. (1998)
374	*Enterococcus faecalis ATCC19433*	3020	G+	Bacillus/Clostridium	M	FL	FA	Oana K, et al (2002)
375	*Brevibacterium linens*	3035	G+	Actinobacteria	M	FL	A	Correira A., Martin, J.F., Castro, J.M. (1994)
376	*Treponema denticola*	3052	G-	Spirochaetales	M	PO	MI	Méndez-Alvarez et al (1995)
377	*Enterococcus faecalisS2*	3055	G+	Bacillus/Clostridium	M	FL	FA	Oana K, et al (2002)

378	*Brevibacterium lactofermentum ATTC13869*	3070	G +	Actinobacteria	M	FL	A	Correira A., Martin, J.F., Castro, J.M. (1994)
379	*Enterococcus avium ATCC19432*	3070	G+	Bacillus/Clostridium	M	FL	FA	Oana K, et al (2002)
380	*Lactobacillus plantarum LP85-2*[a]	3074	G+	Bacillus/Clostridium	M	FL	FA	Chevalier B, Hubert J C and Kammerer (1994)
381	*Clostridium perfringens*	3100	G +	Bacillus/Clostridium	M	FL	ANA	Canard, B. And Cole, S.T. (1989)
382	*Pseudomonas vesicularisATCC11426*	3100	G -	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
383	*Desulfovibrio desulfuricans*	3100	G +	d proteobacteria	M	FL	ANA	Devereux R,Willis SG, Hines ME (1997).
384	*Sulpholobus acidocalcarius*	3100	A	Sulfolobales	H	FL	A	Méndez-Alvarez et al (1995)
385	*Listeria monocytogenes*	3130	G +	Bacillus/Clostridium	M	FL	FA	Michel, E. And Cossart, P. (1992)
386	*Brucella abortus*	3150	G -	a proteobacteria	M	PO	A	Michaux-Characon S. et al (1997)
387	*Mycoplana dimorfa ATCC4279*	3150	G-	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
388	*Brucella ovis*	3150	G -	a proteobacteria	M	PO	A	Michaux-Characon S. et al (1997)
389	*Serpulina hyodysenteriae*	3180	G -	Spirochaetales	M	PO	ANA	Zuerner RL, Stanton TB. (1994)
390	*Carnobacterium divergens*	3200	G +	Bacillus/Clostridium	M	FL	ANA	Casjens S. (1998)
391	*Brucella suis*	3200	G -	a proteobacteria	M	PO	A	Michaux-Characon S. et al (1997)
392	*Brucella neotomae*	3220	G -	a proteobacteria	M	PO	A	Michaux-Characon S. et al (1997)
393	*Clostridium difficile*	3200	G+	Bacillus/Clostridium	M	FL	ANA	Young M and Cole (1993)
394	*Brucella melitensis16M*	3200	G -	a proteobacteria	M	PO	A	Allardet-Servent, A. et al (1991)
395	*Pseudomonas stutzeriDSM50227*	3220	G -	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
396	*Enterococcus faecalisH4*	3250	G+	Bacillus Clostridium	M	FL	FA	Oana K, et al (2002)
397	*Brevibacterium linensATCC19391*	3262	G+	Actinobacteria	M	FL	A	Lima T P, Correia M A (2000)
398	*Brucella canis*	3300	G -	a proteobacteria	M	PO	A	Michaux-Characon S. et al (1997)
399	*Pseudomonas diminutaDSM1635*	3320	G -	a proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
400	*Lactobacillus plantarum CCM1904*	3357	G+	Bacillus/Clostridium	M	FL	FA	Chevalier B, Hubert J C and Kammerer (1994)
401	*Lactobacillus plantarum AM1021*	3357	G+	Bacillus/Clostridium	M	FL	FA	Chevalier B, Hubert J C and Kammerer (1994)
402	*Lactobacillus plantarum AM1223*	3357	G+	Bacillus/Clostridium	M	FL	FA	Chevalier B, Hubert J C and Kammerer (1994)
403	*Enterococcus aviumATCC14025*	3445	G+	Bacillus/Clostridium	M	FL	FA	Oana K, et al (2002)
404	*Pseudomonas putida 16*	3500	G-	g proteobacteria	M	FL	A	Trevors J.T.(1996)
405	*Clostridium thermocellun ATCC27401*	3500	G+	Bacillus/Clostridium	M	FL	ANA	Young M and Cole (1993)
406	*Synechocystis sp PCC6803*	3500	G -	Cyanobacteria	M	FL	A	Kaneko T. et al. (1996)
407	*Deinococcus radiodurans*	3500	G +	Thermus/Deinococcus	T	FL	A	Grimsley JK, et al (1991)

408 *Brevibacterium linenesCCUG12168*	3547	G+	Actinobacteria	M	FL	A	Lima T P, Correia M A (2000)
409 *Brevibacterium linenesCCUG23896*	3547	G+	Actinobacteria	M	FL	A	Lima T P, Correia M A (2000)
410 *Fibrobacter succinogenes*	3573	G -	Fibrobacterium-Acidobacterium group	M	PO	ANA	Ogata K, et al (1997)
411 *Desulfovibrio vulgaris*	3580	G -	d proteobacteria	M	FL	ANA	Devereux R,Willis SG, Hines ME (1997).
412 *Clostridium botulinum 2B A*	3588	G+	Bacillus Clostridium	M	FL	ANA	Hielm S. et al (1998)
413 *Clostridium botulinum 1¯B A*	3588	G+	Bacillus Clostridium	M	FL	ANA	Hielm S. et al (1998)
414 *MV-2*	3600	G-	a proteobacteria	M	FL	MI	Dean,J.A. and Bazylinski (1999).
415 *Pseudomonas putida 1.18*	3600	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
416 *Chromatium vinosum*	3600	G -	g proteobacteria	M	FL	ANA	Gaju N. et al (1996)
41¯ *Clostridium botulinum 1461B*	3606	G+	Bacillus Clostridium	M	FL	ANA	Hielm S, et al (1998)
418 *Desulfovibrio propionicus*	3610	G -	d proteobacteria	M	FL	ANA	Devereux R,Willis SG, Hines ME (1997).
419 *Thiomonas cuprina*	3670	G -	b proteobacteria	M	FL	A	Amils,R. Et al (1998).
420 *Lactobacillus plantarum CST11031*	3671	G+	Bacillus Clostridium	M	FL	FA	Chevalier B, Hubert J C and Kammerer (1994)
421 *MV-1*	3700	G-	a proteobacteria	M	FL	MI	Dean,J.A. and Bazylinski (1999).
422 *Pseudomonas putida 4*	3700	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
423 *Rhodobacter capsulatus*	3800	G -	a proteobacteria	M	FL	ANA	Casjens S. (1998)
424 *Bordetella pertussis*	3734	G -	b proteobacteria	M	PO	A	Stibitz,S.and Garletts,T.L. (1992)
425 *Pseudomonas stutzeriDSM150238*	3750	G-	g proteobacteria	M	FL	A	Ginard,M. Et al (1997)
426 *Acinetobacter sp ADP1*	3750	G -	g proteobacteria	M	FL	A	Gralton EM, Campbell AL, Neidle EL. (1997)
42¯ *Clostridium botulinum 31-25¯0E*	3767	G+	Bacillus Clostridium	M	FL	ANA	Hielm S. et al (1998)
428 *Pseudomonas saccharophilaDSM654*	3780	G -	Purple non Sulfur B	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
429 *Clostridium botulinum ¯06B*	3783	G+	Bacillus Clostridium	M	FL	ANA	Hielm S, et al (1998)
430 *Pseudomonas syringaeDSM150302*	3800	G -	g proteobacteria	M	PO	A	Grothues, D. , and Tümmler, B., (1991)
431 *Caulobacter crescentus*	3800	G -	a proteobacteria	M	FL	A	Ely, B. and Gerardot, C. J. (1993)
432 *Pseudomonas fluorescens (1)*	3800	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
433 *Pseudomonas putida 5*	3800	G-	g proteobacteria	M	FL	A	Trevors J.T. (1996)
434 *Clostridium botulinum C-51E c*	3806	G+	Bacillus/Clostridium	M	FL	ANA	Hielm S, et al (1998)
435 *Clostridium botulinum C-60Ec*	3806	G+	Bacillus/Clostridium	M	FL	ANA	Hielm S, et al (1998)
436 *Clostridium botulinum C-94Ec*	3806	G+	Bacillus/Clostridium	M	FL	ANA	Hielm S, et al (1998)
437 *Legionella pneumophila*	3820	G-	g proteobacteria	M	PO	A	Méndez-Alvarez et al (1995)

438	*Brevibacterium linens CCUG23846*	3823	G+	Actinobacteria	M	FL	A	Lima T P, Correia M A (2000)
439	*Clostridium pasteurianum ATCC603*	3840	G+	Bacillus/Clostridium	M	FL	ANA	Young M and Cole (1993)
440	*Clostridium botulinum KA-2E*	3863	G+	Bacillus/Clostridium	M	FL	ANA	Hielm S, et al (1998)
441	*Clostridium botulinum 4062E*	3882	G+	Bacillus/Clostridium	M	FL	ANA	Hielm S, et al (1998)
442	*Rhodobacter capsulatus SB1003*	3900	G-	a proteobacteria	M	FL	ANA	Jumas-Bilak,E. et al (1998)
443	*Rhodobacter capsulatus ATCC11166*	3900	G-	a proteobacteria	M	FL	ANA	Jumas-Bilak,E. et al (1998)
444	*Lactobacillus casei pseudoplantarum CST11019*	3908	G+	Bacillus/Clostridium	M	FL	FA	Chevalier B, Hubert J C and Kammerer (1994)
445	*Brevibacterium linens ATCC9172t*	3924	G+	Actinobacteria	M	FL	A	Lima T P, Correia M A (2000)
446	*Rhodobacter sphaeroides**	3934	G-	a proteobacteria	M	FL	ANA	Méndez-Alvarez et al (1995)
447	*Pseudomonas viridiflava DMS5033*	3950	G-	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
448	*Rhodococcus fascians*	3964	G+	Actinobacteria	M	FL	A	Crespi M. et al (1992)
449	*Clostridium botulinum Beluga E*	3988	G+	Bacillus Clostridium	M	FL	ANA	Hielm S, Bjorkroth J, Hyytia E, Korkeala H (1998)
450	*Clostridium botulinum RS-1*	3988	G+	Bacillus Clostridium	M	FL	ANA	Hielm S, et al (1998)
451	*Clostridium botulinum R-908 E*	3988	G+	Bacillus Clostridium	M	FL	ANA	Hielm S, et al (1998)
452	*Clostridium botulinum 202F*	3996	G+	Bacillus Clostridium	M	FL	ANA	Hielm S, et al (1998)
453	*Bacillus firmus*	4000	G+	Bacillus Clostridium	M	FL	A	Gronstad A,et al (1998).
454	*Clostridium botulinum 62A*	4000	G+	Bacillus Clostridium	M	FL	ANA	Lin,W.J., Johnson, E.A., (1995)
455	*Pseudomonas stutzeri*	4000	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
456	*Clostridium botulinum 610B8-6F*	4016	G+	Bacillus Clostridium	M	FL	ANA	Hielm S, et al (1998)
457	*Clostridium botulinum FT10F*	4016	G+	Bacillus Clostridium	M	FL	ANA	Hielm S, et al (1998)
458	*Lactobacillus plantarum CST11023*	4022	G+	Bacillus Clostridium	M	FL	FA	Chevalier B, Hubert J C and Kammerer (1994)
459	*Pseudomonas stutzeri DNSP21*	4030	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
460	*Clostridium botulinum R-90E*	4038	G+	Bacillus Clostridium	M	FL	ANA	Hielm S, et al (1998)
461	*Pseudomonas stutzeri SP1402*	4039	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
462	*Micrococcus sp Y-1*	4050	G+	Actinobacteria	M	FL	A	Park,JH., et al (1994)
463	*Pseudomonas stutzeri CH88*	4060	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
464	*Pseudomonas fluorescens (2)*	4061	G-	g proteobacteria	M	FL	A	Trevors J.T.1996
465	*Yersinia pestis*	4400	G-	g proteobacteria	M	PO	FA	Llucier TS, Brubaker RR. (1992)
466	*Pseudomonas putida 8-14*	4080	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
467	*Pseudomonas stutzeri LS401*	4102	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)

468	*Haloferax volcanii DS2*	4130	A	Halobacteriales	T	FL	A	Charlebois, L..R., et al (1991)
469	*Pseudomonas stutzeri DSM50227*	4130	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
470	*Clostridium botulinum 36208E*	4136	G+	Bacillus/Clostridium	M	FL	ANA	Hielm S, et al (1998)
471	*Bacillus subtilis 168*	4140	G +	Bacillus/Clostridium	M	FL	FA	Amjad M, et al (1991)
472	*Clostridium botulinum250E*	4149	G+	Bacillus/Clostridium	M	FL	ANA	Hielm S, et al (1998)
473	*Pseudomonas alcaligenes DSM50342*	4190	G -	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
474	*Pseudomonas pickettii ATTC27511*	4200	G -	b proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
475	*Proteus mirabilis*	4200	G -	g proteobacteria	M	FL	FA	Grothues, D. , and Tümmler, B., (1991)
476	*Pseudomonas stutzeri ST27MN3*	4200	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
477	*Pseudomonas testosterom*	4200	G -	b proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
478	*Pseudomonas pseudoalcaligenes DSM50188*	4200	G-	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
479	*Pseudomonas stutzeri AN11*	4200	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
480	*Pseudomonas setariae*	4270	G -	**	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
481	*Pseudomonas stutzeri ATCC17589*	4290	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
482	*Pseudomonas stutzeri CCUG11256*	4290	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
483	*Pseudomonas stutzeri 19SMN4*	4300	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
484	*Mycobacterium bovis*	4310	G +	Actinobacteria	M	PO	A	Philipp WJ, et al (1996)
485	*Rhodobacter sphaeroides*	4330	G -	a proteobacteria	M	FL	ANA	Jumas-Bilak,E., et al (1998)
486	*Mycobacterium tuberculosis*	4340	G +	Actinobacteria	M	PO	A	Philipp WJ., et al. (1996)
487	*Bacteroides eggerthii*	4350	G -	CFB	M	FL	MI	Shaheduzzaman SM, et al (1997)
488	*Pseudomonas putida DSM50291*	4350	G -	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
489	*Clostridium acetobutylicum ATCC824*	4400	G+	Bacillus/Clostridium	M	FL	ANA	Young M and Cole (1993)
490	*Pseudomonas stutzeri ATCC17591*	4400	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
491	*ZoBell*	4400	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
492	*Pseudomonas stutzeri ST27MN2*	4400	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
493	*Mycobacterium bovis BCG*	4400	G+	Actinobacteria	M	PO	A	Philipp WJ, et al (1998)
494	*Pseudomonas stutzeri ATCC17587*	4420	G-	g proteobacteria	M	FL	A	Ginard, M., et al (1997)
495	*Pseudomonas marginalis DSM50275*	4440	G -	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
496	*Erwinia amylovora*	4450	G -	g proteobacteria	M	PO	FA	Zhang Y, Geider K. (1997)
497	*MC-1*	4500	G-	a proteobacteria	M	FL	MI	Dean,J.A. and Bazylinski (1999).

498	*Pseudomonas stutzeri AN10*	4500	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
499	*Pseudomonas stutzeri LMN2*	4500	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
500	*Shigella flexneri*	4500	G -	g proteobacteria	M	FL	FA	Okada,N., et al (1991)
501	*Salmonella thyphi*	4500	G-	g proteobacteria	M	FL	FA	Thong KL, Puthucheary SD,Pang T.(1997)
502	*Xantomonas axonopodis vesicatoria*	4500	G-	g proteobacteria	M	PO	A	Hacioglu E., Basim H., Stall R. (1996)
503	*Aeromonas hydrophila JMP636*	4500	G-	g proteobacteria	M	FL	FA	Dodd HN, Pemberton JM. (1998)
504	*Pseudomonas stutzeri B2SMN1*	4528	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
505	*Yersinia ruckeri*	4570	G -	g proteobacteria	M	PO	FA	Romalde J.L., Iteman I, Carniel E. (1991)
506	*Escherichia coli ECOR13*	4590	G-	g proteobacteria	M	PO	FA	Bergthorsson U. and Ochman H. (1995)
507	*Bacteroides uniformis*	4592	G -	CFB	M	FL	MI	Shaheduzzaman SM. et al (1997)
508	*Pseudomonas mendocina DSM50017*	4592	G -	g proteobacteria	M	FL	A	Grothues, D . and Tümmler, B., (1991)
509	*Salmonella paratyphi*	4600	G -	g proteobacteria	M	FL	FA	Liu SL, Sanderson KE (1995)
510	*Pseudomonas stutzeri SIMN1*	4600	G-	g proteobacteria	M	FL	A	Ginard,M., (1997)
511	*Burkholderia cepaica SW3*	4600	G-	b proteobacteria	M	FL	A	http: www.apsnet.org online feature Burkholderiacepaica repli
512	*Pseudomonas stutzeri BISMN1*	4600	G-	g proteobacteria	M	FL	A	Ginard,M., et al (1997)
513	*Bacillus megaterium*	4600	G +	Bacillus Clostridium	M	FL	FA	Vary P. (1993)
514	*Escherichia coli ECOR4*	4640	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
515	*Escherichia coli K12*	4670	G -	g proteobacteria	M	FL	FA	Smith, L. C., (1987)
516	*Azotobacter vinelandii*	4676	G -	g proteobacteria	M	FL	A	Maldonado R, Jimenez J, Casadesus J. (1994)
517	*Pseudomonas (X) maltophila DSM50170*	4700	G -	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
518	*Burkholderia cepaica ATCC17760(383)*	4700	G-	b proteobacteria	M	FL	A	http: www.apsnet.org online feature Burkholderiacepaica repli
519	*Escherichia coli K12EMG2*	4700	G-	g proteobacteria	M	FL	FA	Trevors J.T.1996
520	*Ochrobactrum anthropi LMG3301*	4700	G-	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
521	*Salmonella enteriditisLT2*	4700	G -	g proteobacteria	M	FL	FA	Liu SL, Hessel A, Sanderson KE. (1993)
522	*Leptospira interrogans serovar canicola*	4703	G -	Spirochaetales	M	PO	A	Taylor, A.,Barbour,G. And Thomas,D. 1991
523	*Bordetella parapertussis*	4746	G -	b proteobacteria	M	PO	A	Casjens S. (1998)
524	*Bacteroides thetaiotaomicron*	4800	G -	CFB	M	FL	MI	Shaheduzzaman SM. et al (1997)
525	*Bacteroides distasonis*	4800	G -	CFB	M	FL	MI	Shaheduzzaman SM. et al (1997)
526	*Ochrobactrum anthropi ATCC49188*	4800	G-	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
527	*Mycobacterium microti*	4800	G+	Actinobacteria	M	PO	A	Philipp WJ, et al (1998)

528	*Azospirillum irakense*	4800	G-	a proteobacteria	M	FL	A	Martín-Didonet, et al (2000)
529	*Escherichia coli ECOR37*	4850	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
530	*Escherichia coli ECOR15*	4877	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
531	*Salmonella enterica serovar Pullorum*	4930	G-	g proteobacteria	M	PO	FA	Liu GR, et al (2002)
532	*Escherichia coli ECOR63*	4933	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
533	*Escherichia coli ECOR71*	4936	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
534	*Agrobacterium tumefaciens CFPB2607*	4964	G -	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
535	*Escherichia coli ECOR28*	4980	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
536	*Pseudomonas acidovoransDSM50251*	4983	G -	b proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
537	*Pseudomonas aureofaciensDSM50082*	5000	G -	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
538	*Pseudomonas campestrisDSM1049*	5000	G -	**	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
539	*Pseudomonas delafieldiDSM64*	5000	G -	b proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
540	*Pseudomonas flavaDSM1619*	5000	G -	b proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
541	*Pseudomonas fluorescensATCC13525*	5000	G -	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
542	*Leptospira biflexa serovar patoc 1*	5000	G -	Spirochaetales	M	PO	A	Taylor, A.,Barbour,G. And Thomas,D. 1991
543	*Mycobacterium fortuitum*	5000	G+	Actinobacteria	M	PO	A	Kim JR, et al (1996).
544	*Xanthomonas campestris pvglycines*	5000	G-	g proteobacteria	M	PO	A	Widjaja R, Suwanto A, Tjahjono B. (1999).
545	*Mycobacterium intracellulare*	5016	G+	Actinobacteria	M	PO	A	Kim JR, et al (1996).
546	*Escherichia coli 62*	5061	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
547	*Methanosarcina acetivoran spC2A*	5100	A	Methanosarcinales	T M	FL	ANA	Sowers, K. and Gunsalus, P.R. (1988)
548	*Bacteroides vulgatus*	5100	G -	CFB	M	FL	MI	Shaheduzzaman SM, et al (1997)
549	*Agrobacterium tumefaciens C58*	5100	G -	a proteobacteria	M	FL	A	Allardet-Servent, A., et al (1991)
550	*Vibrio parahaemolyticus*	5100	G -	g proteobacteria	M	FL	FA	Yamaichi Y, et al (1999)
551	*Escherichia coliECOR14*	5110	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
552	*Escherichia coli ECOR29*	5121	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
553	*Escherichia coli ECOR68*	5131	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
554	*Pseudomonas solanacearum DSM50905*	5200	G -	b proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
555	*Planctomyces limnophilus*	5200	G -	Planctomycetales	M	FL	FA	Ward-Rainey N, et al (1996)
556	*Escherichia coli ECOR38*	5280	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
557	*Pseudomonas facilis*	5300	G -	b proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)

558	*Pseudomonas palleronii*	5300	G -	b proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
559	*Bacteroides fragilis*	5300	G -	CFB	M	FL	MI	Shaheduzzaman SM. et al (1997)
560	*Bacteroides fragilisYCH46*	5300	G-	CFB	M	FL	ANA	Kuwahara T, et al (2002)
561	*Escherichia coli ECOR51*	5330	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
562	*Clostridium acetobutylicum NCP262*	5337	G+	Bacillus-Clostridium	M	FL	ANA	Keis S, Sullivan T J, Jones T D, (2001)
563	*Escherichia coli ECOR40*	5340	G-	g proteobacteria	M	FL	FA	Bergthorsson U. and Ochman H. (1995)
564	*Phyllobacterium myrsinacearum ATCC43590*	5372	G-	a proteobacteria	M	PO	A	Jumas-Bilak,E., et al (1998)
565	*Bacillus thuringiensis*	5400	G +	Bacillus Clostridium	M	FL	FA	Carlson CR, Kolsto AB. (1993)
566	*Burkholderia cepaica 6?-46*	5400	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica/repli
567	*Pseudomonas cichorii DSM50259*	5400	G-	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
568	*Pseudomonas syringae pathovar rubicola NCPPB 963*	5550	G-	g proteobacteria	M	FL	A	Charnock C. (1998)
569	*Agobacterium tumefaciens bv.CFBP2?21*	5550	G -	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
570	*Bacillus cereus**	5700	G +	Bacillus Clostridium	M	PO	FA	Kolsto, A. Gronstad A., Oppegaard, H. (1990)
571	*Burkholderia cepaica CEP521*	5700	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica/repli
572	*Burkholderia cepaica LMG14293(FC366)*	5700	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica/repli
573	*Agrobacterium rubi ATCC13335*	5735	G -	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
574	*Agrobacterium radiobacter bv.1CFBP2414*	5780	G -	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
575	*Pseudomonas chlororaphis DSM50083*	5800	G -	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
576	*Mycobacterium avium*	5838	G+	Actinobacteria	M	PO	A	Kim JR, et al (1996).
577	*Rizhobium galegae*	5892	G-	a proteobacteria	M	FL	A	Huber I, Slenka-Pobell (1994)
578	*Azospirillum halopraeferens*	5900	G-	a proteobacteria	M	FL	A	Martín-Didonet, et al (2000)
579	*Pseudomonas aeruginosa DSM1?0?*	5900	G -	g proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
580	*Pseudomonas aeruginosa PAO*	5900	G-	g proteobacteria	M	FL	A	Romiling,U. and Tummler,B. (1991)
581	*Agrobacterium radiobacter bv.1 ATCC23308*	5900	G -	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
582	*Agrobacterium radiobacter C58*	5900	G -	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
583	*Achromobacter ruhlandii DSM653*	6000	G-	b proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
584	*Pseudomonas putida KT2440*	6000	G-	g proteobacteria	M	FL	A	Ramos-Díaz MA and Ramos LJ. (1998)
585	*Burkholderia C3430*	6100	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica/repli
586	*Burkholderia cepaica LMG14280(FC365)*	6300	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica/repli
587	*Anabaena sp.PCC?120*	6400	G -	Cyanobacteria	M	FL	A	Kuritz; T., et al (1993)

588	*Rhodococcus sp R 312*	6400	G +	Actinobacteria	M	FL	A	Bigey F, et al (1995)
589	*Burkholderia cepaica 542*	6400	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica /repli
590	*Rhizobium leguminosarum bv.phaseoliATCC14482*	6435	G -	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
591	*Clostridium acetobutylicum NCIMB8052*	6500	G+	Bacillus/Clostridium	M	FL	ANA	Young M and Cole (1993)
592	*Burkholderia pseudomallei*	6500	G -	b proteobacteria	M	FL	A	Songssivila S, Dharakul T. (2000)
593	*Strptomyces ambofaciens ATCC15154*	5000	G+	Bacillus/Clostridium	M	FL	A	Leblond,Pet al .(1990).
594	*Rhizobium meliloti 1021 and 2011*	6500	G -	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
595	*Rhizobium fredii ATCC35423*	6600	G -	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
596	*Burkholderia cepaica CEP040*	6700	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica /repli
597	*Azospirillum brasilense FP2*	6700	G-	a proteobacteria	M	FL	A	Martín-Didonet, et al (2000)
598	*Pseudomonas cepaica DSM150180*	6700	G -	b proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
599	*Rhizobium leguminosarum bv trifolii ATCC14480*	6800	G -	a proteobacteria	M	FL	A	Jumas-Bilak,E.,et al (1998)
600	*Pseudomonas glathei DSM150014*	6800	G -	b proteobacteria	M	FL	A	Grothues, D. , and Tümmler, B., (1991)
601	*Burkholderia cepaica C5568*	6800	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica /repli
602	*Azospirillum brasilense SP7*	6800	G-	a proteobacteria	M	FL	A	Martín-Didonet, et al (2000)
603	*Bacteroides ovatus*	6900	G -	CFB	M	FL	MI	Shaheduzzaman SM, et al (1997)
604	*Azospirillum brasiliense Cd*	6900	G-	a proteobacteria	M	FL	A	Martín-Didonet, et al (2000)
605	*Burkholderia cepaica C527-4*	7000	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica /repli
606	*Burkholderia cepaica CEP024*	7000	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica /repli
607	*Burkholderia cepaica BC11*	7000	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica /repli
608	*Burkholderia cepaica ATCC29424(DB01)*	7100	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica /repli
609	*Ralstonia eutropha H16*	7100	G-	b proteobacteria	M	FL	ANA	Schwartz E, Friedrich B. (2001)
610	*Azospirillum brasiliense Sp245*	7100	G-	a proteobacteria	M	FL	A	Martín-Didonet, et al (2000)
611	*Burkholderia cepaica CEP511*	7200	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica /repli
612	*Azospirillum amazonenseY6*	7200	G-	a proteobacteria	M	FL	A	Martín-Didonet, et al (2000)
613	*Agrobacterium rhizogenes bv2ATCC11325*	7235	G -	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
614	*Agrobacterium rhizogenes K84*	7265	G -	a proteobacteria	M	FL	A	Jumas-Bilak,E., et al (1998)
615	*Azospirillum amazonenseY2*	7300	G-	a proteobacteria	M	FL	A	Martín-Didonet, et al (2000)
616	*Mycobacterium gordonae*	7395	G+	Actinobacteria	M	PO	A	Kim JR, et al (1996).
617	*Burkholderia cepaica BcF*	7400	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature/Burkholderiacepaica /repli

618	*Burkholderia cepaica C5424*	7500	G-	b proteobacteria	M	FL	A	http://www.apsnet.org/online/feature Burkholderiacepaica /repli
619	*Burkholderia cepaica LMG10929 (FC369)*	7500	G-	b proteobacteria	M	FL	A	http: /www.apsnet.org online feature Burkholderiacepaica /repli
620	*Streptomyces ambofaciens DSM40697*	7700	G +	Bacillus/Clostridium	M	FL	A	Leblond,P. Et al (1990).
621	*Burkholderia cepaica ATCC25416(CEP031)*	7800	G-	b proteobacteria	M	FL	A	http: /www.apsnet.org online feature Burkholderiacepaica /repli
622	*Burkholderia cepaica LMG12615(FC364)*	7900	G-	b proteobacteria	M	FL	A	http: /www.apsnet.org online feature Burkholderiacepaica repli
623	*Streptomyces griseus*	7900	G +	Bacillus Clostridium	M	FL	A	Lezhava A, et al 1995.
624	*Burkholderia cepaica ATCC17759(CEP080)*	7900	G-	b proteobacteria	M	FL	A	http: /www.apsnet.org online feature Burkholderiacepaica repli
625	*Burkholderia cepaica C4455*	7900	G-	b proteobacteria	M	FL	A	http: /www.apsnet.org online feature Burkholderiacepaica repli
626	*Azospirillum lipoferum JA25*	7900	G-	a proteobacteria	M	FL	A	Martín-Didonet, (2000)
627	*Burkholderiacepaica LMG12614(FC363)*	8000	G-	b proteobacteria	M	FL	A	http: /www.apsnet.org online feature Burkholderiacepaica repli
628	*Burkholderia cepaica CRE7*	8000	G-	b proteobacteria	M	FL	A	http: /www.apsnet.org online feature Burkholderiacepaica repli
629	*Streptomyces coelicolor A3(2)*	8000	G +	Bacillus Clostridium	M	FL	A	Kieser,M.H., Kieser,Tand Hopwood,D.(1992)
630	*Streptomyces lividans*	8000	G +	Bacillus Clostridium	M	FL	A	Leblobd P. Redenbach M.Cullum J (1993)
631	*Burkholderiacepaica C6435*	8100	G-	b proteobacteria	M	FL	A	http: /www.apsnet.org online feature Burkholderiacepaica repli
632	*Streptomyces ambofaciens ETH9427*	8200	G +	Bacillus Clostridium	M	FL	A	Leblond,P. Et al (1990).
633	*Streptomyces ambofaciens ETH1317*	8200	G +	Bacillus Clostridium	M	FL	A	Leblond,P. Et al (1990).
634	*Burkholderiacepaica LMG14294(FC367)*	8200	G-	b proteobacteria	M	FL	A	http: /www.apsnet.org online feature Burkholderiacepaica repli
635	*Burkholderia cepaica FC362*	8600	G-	b proteobacteria	M	FL	A	http: /www.apsnet.org online feature Burkholderiacepaica repli
636	*Bradyrhizobium japonicum 110*	8700	G -	a proteobacteria	M	FL	A	Kundig,Ch., Hennecke, H., and Gottfert,M.(1993).
637	*Burkholderia cepaica ATTC53617*	8900	G-	b proteobacteria	M	FL	A	http: /www.apsnet.org online feature Burkholderiacepaica repli
638	*Stigmatella aurantiaca DW4 3.1 chromosome*	9350	G -	d proteobacteria	M	FL	A	Neuman,B., Pospiech, A., and Urich Schairer, H.(1993)
639	*Myxococcus xanthus*	9427	G -	d proteobacteria	M	FL	A	Chen,H., Keseler I.M., Shimkets,l.j. (1990)
640	*Stigmatella erecta*	9550	G-	d proteobacteria	M	FL	A	Neumann B, Pospiech A, Schairer HU. (1992)
641	*Azospirillum lipoferumSp59b*	9700	G-	a proteobacteria	M	FL	A	Martín-Didonet, et al (2000)

Apéndice 1b Referencias de la base de datos

Akman L, & Aksoy S. (2001). A novel application of gene arrays: *Escherichia coli* array provides insight into the biology of the obligate endosimbiont of tsetse flies. PNAS **98** (13): 7546-51

Akman L, Río RV, Beard CB, Aksoy S. (2001). Genome size determination and coding capacity of *Sodalis glossinidius* an enteric symbiont of tsetse flies, as revealed by hybridization to Escherichia coli gene arrays. J Bacteriol. **183** (15): 4517-4525.

Allardet-Servent A, Carles-Nurit MJ, Bourg G, Michaux S, and Ramuz, (1991). M. Physical Map of the *Brucella mellitensis 16M* Chromosome. J Bacteriol. **173** (7) 2219-2224

Allardet-Servent A, Michaux-Characon S, Jumas Bilak E, Karayan L, and Ramuz M. (1991). Presence of One Linear and One Circular Chromosome in the *Agrobacterium tumefaciens C58* Genome. J Bacteriol. **175** (24): 7869-7874

Alleman AR, Kamper SM, Viseshakul N, Barbet AF. (1993) Analysis of the *Anaplasma marginale* genome by pulsed-field electrophoresis. J Gen Microbiol. **139** (pt10): 2439-44

Amils R, Irazabal N, Moreira D, Abad JP, Marin I.(1998). Genomic organization analysis of acidophilic chemolithotrophic bacteria using field gel electrophoretic techniques. Biochimie. **80** (11): 911-21

Amjad M, Castro JM, Sandoval H, Wu JJ, Yang M, Henner DJ, Piggot PJ. (1991). An SifI restriction map of the *Bacillus subtilis 168* genome. Gene. **101** (1): 15-21

Bauman C, Judex M, Huber H, Wirth R. (1998). Estimation of genome sizes of hyperthermophiles. Extremophiles. **2** (2): 101-8

Bautsch W. (1993). A NheI macrorestriction map of the *Neisseria meningitidis B1940* genome. FEMS Microbiol Lett. **107** (2-3): 191-7

Bautsch W. (1988). Rapid Physical Mapping of the *Mycoplasma mobile* genome by two-dimensional field invertion gel electrophoresis techniques. Nucleic Acids Research. **16** (24): 11461-11467

Bergthorsson U. and Ochman H. (1995). Heterogeneity of Genome sizes among Natural isolates of *Escherichia coli.* J Bacteriol. **177** (20):5784-5789

Bigey F, Janbon G, Arnaud A, Galzy P. (1995). Sizing of the *Rhodococcus sp R312* Genome by pulsed-field gel electrophoresis. Localization of genes involved in nitrile degradation. Antonie Van Leeuwenhoek. **68**:173-79

Bihlmaier A, Römling U, Meyer TF, Tümmler and Gibbs CP. (1991). Physical and Genetic Map of the *Neisseria gonorrhoeae* strain MS11-N198 chromosome. Mol Microbiol. **5:** (10): 2529-2539.

Birkelund S. and Stephens RS. (1992) Construction of Physical and Genetic Maps of *Chlamydia trachomatis Serovar L2* by Pulsed-Field Gel Electorphoresis. J Bacteriol. **174** (9): 2742-2747.

Bobovnikova Y, Ng WL, Dassrma S, Hackett N. (1994) Restriction mapping the genome of *Halobacterium halobium strain NRC-1.* Syst Appl Microbiol. **16:** 597-604

Bolstad A I. (1994). Sizing the *Fusobacterium nucleatum* genome by pulsed-field gel electrophoresis. FEMS Microbiol Lett. **123:** (1-2)145-51

Borges KM, Bergquist PL, (1993). Pulsed Field gel electophoresis study of the genome of *Caldocellum saccharolyticum.* Curr Microbiol. **27:**15-19

Bourget N, Sominet J-M, and Decaris B. (1993). Analysis of the Genome of the five *Bifidobacterium breve* strains: plasmid content, pulsed-field gel electrophoresis genome size estimation and rrr loci numeber. FEMS Microbiol Letters. **110:** 11-20

Bourke B, Sherman P, Louie H, Hani E, Islur P, Chan VL. (1995). Physical and genetic map of the genome of *Campylobacter upsaliensis.* Microbiology. **141:**2417-24

Canard B. and Cole ST. (1989). Genome organization of the anaerobic pathogen *Clostridium perfringens.* PNAS USA. **86:** 6676- 6680

Carle P, Laigret F, Tully JG, and Bové (1995). Heterogeneity of genomes Sizes within the Genus *Spiroplasma.* Int J Syst Bacteriol. **45** (1): 178-181

Carlson CR, Kolsto AB. (1993). A complete physical map of a *Bacillus thuringiensis* chromosome. J Bacteriol. **175:**1053-60

Casjens S, Huang WM. (1993). Linear chromosomal physical and genetic map of *Borrelia burgdorferi*, the lyme disease agent. Mol Microbiol. **8** (5): 967-80

Casjens S. (1998). The Diverse and dynamic structure of bacterial genomes. Annu Rev Genet. **32:** 339-77

Correira A, Martin JF, Castro JM. (1994). Pulsed-field gel electrophoresis analysis of the genome of aminoacid producing corynebacteria : chromosome sizes and diversity of restriction patterns. Microbiology. **140** (Pt10): 2841-7

Crespi M, Messens E, Caplan AB, van Montagu M, Desomer J. (1992). Fasciation induction by the phytopathogen *Rhodococcus fascians* depends upon a linear plasmid encoding cytokinin synthase gene. EMBO J. **11:** 795-804

Chang N. and Taylor DE. (1990). Use of Pulsed Field Agarose Gel Electrophoresis to Size Genomes of *Campilobacter* Species and to Construct a SalI Map of *Campylobacter jejuni UA580.* J Bacteriol. **172** (9): 5211-5217.

Charlebois LR, Schalkwyk CL, Hoffman DJ, and Doolitle WF. (1991). Detailed Physical Map and of Overlapping Clones Covering the Genome of the *Archaebacterium Haloferax volcanii DS2.* J Mol Biol. **222:** 509-524

Charles H, Ishikawa H. (1999). Physical and Genetic Map of the Genome of *Buchnera*. The Prymary Endosymbiont of the Pea Aphid *Acytrosiphon psium* J Mol Evol. **48:**142-150

Charnock C. (1998). Characterization of phytopathogen *Pseudomonas siringae pathovar ribicola NCPPB 963.* Microbios. **94** (377): 23-34

Chevalier B, Dugourd D, Tarasiuk, Harel J, Gotlschalk M, et al. (1998). Chromosome sizes and phylogenetic relationships between serotypes of *Actinobacillus pleuropneumoniae.* FEMS Microbiol Lett. **160** (2): 209-16

Chevalier B, Hubert JC, and Kammerer. (1994). Determination of Cromosome size and number rrn loci in *Lactobacillus plantarum* by pulsed-field gel electrophoresis. FEMS Microbiology Letters. **120:** 51-56

Chen X. and Widger WR. (1993). Physical Genome Map of the Unicellular Cyanobacterium *Synechococcus sp. Strain PCC 7002.* J Bacteriol. **175** (16): 5106-5116

Chen H, Keseler IM, Shimkets, IJ. (1990). Genome size of Myxococcus xanthus determined by pulsed-field gel electrophoresis. J Bacteriol. **172** (8):4206-13

Chung JSU, and Baseman JB. (1990). Genome Size of *Mycoplasma genitallium.* J Bacteriol. **172** (8): 4705-4707.

Dean JA, and Bazylinski. (1999). Genome Analysis of Several Marine Magnetotactic Bacterial Strains by Pulsed Field Gel Electrophoresis. Curr Microbiol. **39:** 219-225

Dempsey J A, York J, Cannon JG. (1993). Characterization of the genome of *Acetobacter xylinum* by pulsed field gel electrophoresis. absH-64, p 201 In Abstacts of the 93 rd. General meeting of American Society for Microbiology. American Society for Microbiology, Washington, D.C,

Devereux R, Willis SG, Hines ME (1997). Genome sizes of *Desulfovibrio desulfuricans*, *Desulfovibrio vulgaris*, and *Desulfovibrio propionicus* estimated by pulsed-field gel electrophoresis of linearized chromosomal DNA. Curr Microbiol. **34** (6): 337-9

Dmitriev A, Suvorov A, Totolian A. (1998). Physical and genetic chromosomal maps of *Strptococcus agalactiae*, serotypes II and III rRNA operon organization. FEMS Microbiol Lett. **167**(1): 33-9

Dodd HN, Pemberton JM. (1998). Construction of a physical and preliminary genetic map of *Aeromonas hydrophyla JMP636.* Microbiology. **144** (Pt 11): 3087-3096

Dudez AM, Chaillou S, Hissler L, Stentz R, Champomier-Verges MC, Alpert, CA, Zagorec, M. (2002). Physical and genetic map of the *Lactobacillus sakei 23K* Chromosome. Microbiology **148** (Pt 2): 421-31

Eiglmeier K, Honore N, Woods SA, Caudron B, Cole ST. (1993). Use of an ordered cosmid library to deduce the genome organization of *Mycobacterium leprae*. Mol Microbiol **7:** 197-206

Ely B. and Gerardot CJ. (1993). Use of Pulsed- Field Gel Electrophoretic to construct a physical map of the *Caulobacter crescentus* genome. Gene. **68:** 323-333

Frutos R. Pages M, Bellis M, Roizes G,and Bergoin M. (1989). Pulsed-Field Gel Electrophoresis Determination of the Genome Size of Obligate Intracellular Bacteria Belonging to the genera *Chlamydia*, *Rickettsiella*, and *Porochlamydia*. J Bacteriol. **171** (8): 4511-4513.

Fujiwara S, Takagi M, Imanaka , (1998). Archaeon *Pyrococcus kodakaraensis KOD1*: Application and Evolution Biotechnol. Annu. Rev. 4: 259-84

Furihata K, Sato K, Matsumoto H. (1995) Construction of a combined Not I/SamaI Physical and genetic map of *Moraxella* (*Branhamella*) *catarrhalis strain ATCC25238*. Microbiol Immunol. **39** (10): 745-51

Gaju N, Pavon V, Marin I, Esteve I, Guerrero R, Amils R. (1995). Chromosome map of the phototrophic bacterium *Chromatium vinosum*. FEMS Microbiol. **126** : 241-248

Gasc AM, Kauc L, Barraillé P, Sicard M. and Goodgal S. (1991) Gene localization size, and physical map of the chromosome of *Streptococcus pneumoniae*. J Bacteriol. **173:** 7361-7367.

Gil R, Sabater-Muñoz B, Latorre A, Silva JF and Moya A. (2002). Extreme Genome reduction in *Buchnera spp*: Toward the minimal genome needed for symbiotic life. PNAS. **99** (7): 4454-4458

Ginard M, Lalucat J, Tummler B, and Romling U. (1997). Genome Organization of *Pseudomonas stutzeri* and Resulting Taxonomic and Evolutionary Considerations. Int J Syst Bacteriol. **47**(1): 132-143.

Gralton EM, Campbell AL, Neidle EL. (1997). Directed introduction of DNA cleavage sites to produce a high resolution

genetic and physical map of the *Acinetobacter sp strain ADP 1 (BD413UE)* chromosome. Microbiology. **143:** 1345-57

Grimsley JK, Masters CI, Clark EP, Minton KW. (1991) Analysis by pulsed-field gel electrophoresis of DNA double strand breakge and repair in *Deinococcus radiodurans* and radiosensitive mutant. Int J Radiat Biol. **60:** 613-26

Gronstad A, Jaroszewicz E, Ito M, Sturr Mg, Kruiwich TA, Kolsto AB. (1998). Physical map of alkaliphilic *Bacillus firmus OF4* and detection of a large endogenous plasmid. Extremophiles. **2** (4): 447-53

Grothues D, and Tümmler B, (1991). New approaches in genome analysis by pulsed field gel electrophoresis application to the analysis of *Pseudomonas species*. Mol Microbiol. 5 (11): 2763-2776.

Hacioglu E, Basim H, Stall R. (1996). Rarely cutting restriction endonucleases useful for determining genome size and physical map of the chromosome of *Xanthomonas axonopodis pv vesicatoria*. Phytopathol. **86:**S 77-S78

Hannien M, and Hirvi U. (1999). Genetic diversity of canine gastric helicobacters, *Helicobacter bizzozeronii* and *H. salomonis* studied by pulsed-field gel electrophoresis. J Med Microbiol. **48**(4): 341-347

Hantman JM, Sun S, Piggot PJ, and Daneo-Moore L. (1993). Chromosome Organization of *Streptococcus mutans GS-5*. Journal of Gen Microbiol. **139:** 67-77

Heinzen R, Stiegler GL, Whithing LL, Schmittt SA, Mallavia LP, Fraizer ME. (1990). Use of pulsed field gel electrophoresis to differentiate *Coxiella burnetii* strains. Ann N Y Acad Sci. **590:** 504-13

Hermann T, Wersch G, Uhlemann Em, Schmid R, Burkovski A (1998). Mapping and identification of *Corynebacterium glutamicum* protein by two-dimensional gel electrophoresis and microsequencing. Electrophoresis. **19** (18): 3217-21

Hielm S, Bjorkroth J, Hyytia E, Korkeala H. (1998). Genomic analysis of *Clostridium botulinum* group 11 by pulsed-field yeld electrophoresis. Appl Envirom Microbiol. **64** (2): 703-708

Hobbs MM, Leonardi MJ, Zaretzky FR, Wang TH, Kawula TH. (1996). Organization of the *Haemophilus ducreyi 35000* chromosome. Microbiology **142:** 2587-94

http://www.apsnet.org/online/feature/Burkholderiacepaica /replicon.htm

Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature **417**(6884): 63-7

Huber I, Slenka-Pobell. (1994). Pulsed-field gel electrophoresis-fingerprinting, genome size estimation and rrn loci number of *Rhizobium galgae*. J.Appl Bacteriol. **77:** 528-533

Jalava K, De Ungria MC, O´Rourke J, Lee A, Hirvi U, and Hannien M-L. (1999). Characterization of *Helicobacter felis* by Pulsed-Field Gel Electrophoresis, Plasmid Profiling and Ribotyping. Helicobacter. **4**(1): 17-27

Jumas-Bilak E, Michaux-Charachon, S, Bourg G, Ramuz M, and Alardent-Servent. (1998). Unconventional genomic organization in the alpha subgroup of the proteobacteria. J Bacteriol. **180** (10): 2749-2755

Kakulphimp J, Finch LR, Robertson JA. (1991). Genome sizes of mammalian and avian Ureaplasmas. Int. J. Syst. Bacteriol **41:** 326-27

Kaneko T, Matasubayashi T, Sugita M, Sugiura M. (1996). Physical and gene maps of the unicellular cyanobacterium *Synechococcus sp. Strain PCC6301* genome. Plant Mol. Biol. **31:**193-201

Kaneko T. Sato S, Kotani H, Tanaka A, Asamisu E, et al. (1996). Sequence analysis of the genome of the unicelular *Synechocystis sp strain PCC6803II*. Sequence determination of the genome and assignment of potential protein-coding regions. DNA Res. **3:**109-36

Kang HL, Kang HS. (1998). A physical map of the genome of ethanol fermentative bacterium *Zymomonas mobilis ZM4* and localization of genes on the map Gene **206**(2): 223-8

Karlyshev V A, Henderson J, Ketley M J, and Wren W B. (1998). An improved physical and genetic map of *Campylobacter jejuniNCTC11168 (UA580)*. Microbiology. **144:** 503-508

Kauc L, And Goodgal SH. (1989). The size and physical map of the chromosomes of *Haemophilus parainfluenzae*. Gene. **83:** 377-380

Keis S, Sullivan TJ, Jones TD. (2001). Physical and genetic map of the *Clostridium Saccharobutylicum* (formerly *Clostridium acetobutylicum) NCP262* Chromosome. Microbiology. **147:** 1909-1922

Kieser MH, Kieser T and Hopwood D.(1992). A Combined Genetic and Physical Map of the *Streptomyces coelicolor A3(2)* Chromosom3. J Bacteriol. **174**(17): 5496-507

Kim JR, Kang BS, Ko JH, Park JS, Kim SJ, Bai GH Chung TH, Nam KS, Choi YK, Choi IS, Chung TW, Lee YC, Kim CH (1996). Genomic heterogeneity in clinical strains of *Mycobacterium tuberculosis*, *M terraecomplex*, *M gordonae*, *M avium-intracellulare* complex and *M fortuitum* by pulsed-field gel electrophoresis. J Biochem Mol Biol. **29:** 569-573.

Kim NW, Bingham H, Khawaja R, Louie H, Hani E, Neote K, Chan VL. (1992). Physical Map of *Campylobacter jejuni TGH9011* and localization of 10 genetic markers by use of pulsed-field gel electrophoresis. J Bacteriol **174** (11): 3494-8

Kolsto A. Gronstad A, Oppegaard H. (1990). Physical Map of the *Bacillus cereus* Chromosome. J Bacteriol **172:** 3821-3825

Krueger CM, Marks KL, Ihler GM. (1995). Physical map of the *Bartonella bacilliformis* genome J Bacteriol. **177:** 7271-74

Kundig Ch, Hennecke H, and Gottfert M.(1993). Correlated physical and genetic map of the *Bradyrhizobium japonicum 110 genome*. J Bacteriol. **175** (3): 613-622

Kuritz T, Ernest A, Black AT, and Wolk CP. (1993). High-resolution mapping of genetic loci of *Anabaena PCC 7120* required for photosynthesis and nitrogen fixation. Mol Microbiol. **8** (1): 101-110

Kuwahara T, Sarker MR, Ugai H, Akimoto S, Shaheduzzaman SM, Nakayama H, Miki T, Ohnishi Y. (2002). Physical and genetic map of the *Bacteroides fragilis YCH46* Chromosome. FEMS Microbiol Lett **207** (2): 193-7

La Fontaine S, Rood JI. (1997). Physical and genetic map of the chromosome of *Dichelobacter nodosus strain A198*. Gene. **184:** 291-98

Ladefoged SA, and Christiansen G. (1992). Physical and Genetic Mapping of the Genomes of Five *Mycoplasma hominis* Strains by Pulsed-Field Gel Electorphoresis. J Bacteriol. **174** (7): 2199-2207.

Le Burgeois P, Mata M, and Ritzenthaler. (1989). Genome Comparation of *Lactococcus* strains by pulsed field gel electrophoresis. FEMS Microbiol Letters **59:** 65-70

Leblobd P, Redenbach M, Cullum J. (1993). Physical of *Streptomyces lividans 66* genome and comparison with that of the related strain Steptomyces coelicolor A3 (2). J Bacteriol. **175** (11): 3422-9

Leblond P, Francou FX, Simonet JM, Decaris B.(1990).Pulsed-field gel electrophoresis analysis of the genome of *Streptomyces ambofaciens* strains. FEMS Microbiol Lett. **60** (1-2): 79-88

Lee JJ, Smith HO. (1988). Sizing of the *Haemophilus influenzae Rd* genome by pulsed-field agarose gel electrophoresis. J Bacteriol 170(9): 4402-5

Lezhava A, Mizukami T, Kajitani T, Kameoka D, Redenbach M, et al. (1995). Physical map of the linear chromosome of *Streptomyces griseus*. J Bacteriol.**177:** 6492-98

Lim D, Trivedi H, Nath k. (1994). Determination of *Gardnerella vaginalis* genome size by pulsed-field gel electrophoresis. DNA Res. **1**(3):115-22

Lima TP, Correia MA. (2000). Genetic fingerprinting of *Brevibacterium linens* by pulsed-field gel electrophoresis and ribotyping. Curr Microbiol. **41:** 50-55

Lin WJ, Johnson EA. (1995). Genome analysis of *Clostridium botulinum TypeA* by pulsed-field gel electrophoresis. Appl Environ Microbiol. **61** (12): 4441-7

Liu SL, Hessel A, Sanderson KE. (1993). The XbaI-BInI-CeuI genome cleavage map of *Salmonella enteriditis* shows an inversion relative to Salmonella thyphimurium LT2. Mol Microbiol. **10:** 655-64

Liu SL, Sanderson KE. (1995). The chromosome of *Salmonella paratyphi A* is inverted by recombination between rrnH and rrnG. J Bacteriol. **177** (22): 6585-92

Liu Hessel A. And Sanderson. (1993). Genomic mapping wiyh I-Ceu I, an intron-encoded endonuclease specific for genes for ribosomal RNA, in *Salmonella spp.*, *Escherichia coli*, and other bacteria. PNAS. USA **90:** 6874-6878

Liu GR, Rahn A, Liu WQ, Sanderson KE, Johnston RN, Liu SL (2002). The evolving genome of *Salmonella enterica serovar Pullorum* J Bacteriol **184** (10):2626-33

López-García P, Amils R. and Antón J. (1996). Sizing chromosomes and megaplasmids in haloarchaea. Microbiology. **142:** 1423-1428

Lortal S, Rouault A, Guezenec S, Gautier M. (1997). *Lactobacillus helveticus*: strain typing and genome size estimation by pulsed field gel electrophoresis. Curr Microbiol. **34:**180-185

Llucier TS, Brubaker RR. (1992) Determination of genome size, macrorestriction pattern polymorphism, and nonpigmentation-specific deletion in *Yersinia pestis* by pulsed-field gel electrophoresis. J Bacteriol. **174:** 2078-86

Majumder R, Sengupta S, Khetawat G, Bhadra RK, Roychoudhury S, Das J. (1996)Physical Map of the genome of *Vibrio cholerae 569B* and localization of genetic markers. J Bacteriol. **178** (4): 1105-12

Maldonado R, Jimenez J, Casadesus J. (1994) Changes of ploidy during the *Azotobacter vinelandii* growth cycle. J Bacteriol. **176:** 3911-19

Martín-Didonet, Chubatsu S L, Souza M E, Kleina M, Rego GM, Rigo UL, Yates GM, Pedrosa O. (2000). Genome Structure of the Genus *Azospirillum* .J Bacteriol .**182** (14): 4113-4116

Matsuda M, Asami Y, Miyazawa T, Sugawara T, Kumano M, Isayama Y, Honda M. (1994). Estimation of the genome of

Taylorella equigenitalis by crossed fiel gel electrophoresis. Vet. Res Comm. **18:** 99-102

Matsumoto K, Matsuda M, and Kaneuchi Ch. (1992). Analysis of Chromosome-sized DNA form the Bacterial Genome of Thermophilic *Campilobacter laridis* by Pulsed-Field Gel Electorphoresis and Physical Maping. Microbios. **71:** 7-14.

Mellado E, García MT, Nieto JJ, Kaplan S, Ventosa A. (1997) Analysis of the genome of *Vibrio costicola*: Pulsed field gel electrophoretic analysis of the genome size and plasmid content. Syst Appl Microbiol **20:** 20-26

Mellado E, García MT, Roldán E, Nieto JJ, Ventosa A. (1998) Analysis of the genome of the gram-negative moderate halophiles *Halomonas* and *Chromohalobacter* by using pulsed-field electrophoresis. Extremophiles. **2:** 435-438.

Méndez-Alvarez et al. (1995). Microbiología. SEM **11:** 323-336

Méndez-Alvarez S, Pavon V, Esteve I, Guerrero R, Gaju N. (1995). Genomic heterogeneity in *Chlorobium limicola*: Chromosomic and plasmidic differences among strains. FEMS Microbiol. Lett.**134:** 279-85

Messick JB, Smith G, Berent L, Cooper S. (2000). Genome size of *Eperythrozoon suis* and hybridization with 16S rRNA gene. Can J Microbiol. **46** (11): 1082-1086

Michaux-Characon S, Bourg G, JumasBilak E, Guigue-Talet P, Allardet-Servent A, et al. (1997). Genome structure and phylogeny in the genus *Brucella*. J Bacteriol. **179:** 3244-49

Michel E. and Cossart P. (1992). Physical map of the *Listeria monocytogenes* chromosome. J Bacteriol. **174:** 7098-7103.

Moreira MM, Da Costa, SM, Sá-Correira Isabel. (1997). Comparative genomic analysis of isolates beloning to the six species of the genus *Thermus* using pulsed-field gel electrophoresis and ribotyping. Arch Microbiol. **168** (2): 92-101

Naterstad K, Kolsto AB, Sirevag R. (1995). Physical map of the genome of the green phototrophic bacterium *Chlorobium tepidum*. J Bacteriol. **177:** 5480-84

Neimark HC, Lange CS. (1990). Pulsed field electrophoresis indicates full length Mycoplasma chromosomes range widely in size. Nucleic Acids Res. **18:** 5443-48

Neuman B, Pospiech A, and Urich Schairer, H. (1993). A physical and genetic map of the *Stigmatella aurantiaca DW4/3.1* crhomosome. Mol Microbiol. **10** (5):1087-1099

Neumann B, Pospiech A, Schairer HU. (1992). Size and stability of the genomes of the myxobacteria *Stigmatella aurantiaca* and *Stigmatella erecta.* J Bacteriol. **174:** 6307-10

Noll MK, (1989). Chromosome Map of thermophilic Archaebacterium *Thermococcus celer.* J Bacteriol **171** (12): 6720-6725

Oana K, Okimura Y, Kawakami Y, Hayashida N, Shimosaka M, Okazaki M, Hayashi T, Ohnishi M. (2002). Physical and genetic map of the *Enterococcus faecium ATCC19434* and demonstration of intra- and interspecific genomic diversity in enterococci. FEMS Microbiol Lett. **207** (2): 133-9

Ogata K, Aminov RI, Nagamine T, Sugiura M, Tajima K, Mitsumori M, Sekizaki T, Kudo H, Minato H, Benno Y. (1997) Construction of a *Fibrobacter succinogenes* genomic map and demostration of diversity at the genomic level. Curr Microbiol. **35** (1):22-7

Okada N, Sasakawa C, Tobe T, Talukder KA, Komatsu K. And Yoshikawa M. (1991) Construction of a physical map of the chromosome of *Shigella flexneri 2a* and direct assigment of nine-virulence associated loci identified by Tn5 insertions. Mol Microbiol. **5:** 2171-80

O´ Riordan K, Fitzgerald FG. (1997). Determination of Genetic diversity within the genus *Bifidobacterium* and estimation of chromosomal size. FEMS Microbiol Letters.**156:** 259-264

O`Sullivan TF, Fitzgerald GF, (1998). Comparison of *Streptococcus thermophilus* stains by pulsed field geld electrophoresis of genomic DNA. FEMS Microbiol Lett. **168**(2): 213-9

Park JH, Song JC, Kim MH, Lee DS, Kim CH. (1994). Determination of genome size and preliminary physical map of an extreme alkaliphile, *Micrococcus sp. Y-1*, by pulsed-field gel electrophoresis. Microbiology. **140** (Pt9):2247-50

Philipp WJ, Gordon S, Telenti A and Cole T S. (1998) Pulsed field gel electrophoresis for Mycobacteria. Methods Mol Biol. **101:** 51-63

Philipp WJ, Nair S, Guglielmi G, Laggranderie M, Gicquel B, Cole ST. (1996). Physical mapping of *Mycobacterium bovis BCG* pasteur reveals differences from genome map of *Mycobacterium tuberculosis H37Rv* and M. Bovis. Microbiology. **142:** 315-45

Philipp WJ, Poulet S, Eiglmeier K, Pascopella L, Balasubramanian V, et al. (1996). A integrated map of the genome of tubercle bacillus, Mycobacterium tuberculosis H37Rv, and comparison with Mycobacterium leprae. PNAS USA **93:** 3132-37

Pyle LE, and Finch LR. (1988). A Physical Map of the Genome of *Mycoplasma mycoides subespecies mycoides Y* with some functional moci. Nucliec Acids Research. **16** (13): 6027-6039.

Ramos-Díaz MA, and Ramos LJ. (1998). Combined physical and genetic map of the *Pseudomonas putida KT2440* Chromosome. J Bacteriol. **180** (23): 6352-6363

Resche DK, Frazier ME, Mallavia LP. (1991). Transformation and genomic restriction mapping of *Rochalimaea spp.* Acta Virol. **35** (6): 519-25

Robertson JA, Pyle LE, Stemke GW, Finch LR. (1990). Human Ureaplasmas show diverse genome sizes by pulsed-field electrophoresis. Nucleic Acids Res. **18:** 1451-55

Romalde JL, Iteman I, Carniel E. (1991). Use of pulsed field gel electrophoresis to size the chromosome of the bacterial fish pathogen *Yersinia ruckeri.* FEMS Microbiol lett. **68** (2): 217-25

Romiling U. and Tummler B. (1991). The Impact of two-dimensional pulsed-field gel electrophoresis techniques for the consistent and complete mapping of bacterial genomes: refined physical map of *Pseudomonas aeruginosa PAO.* Nucleic Acids Research. **19** (12): 3199-3206

Roussel Y, Colmin C, Simonet JM, Decaris B. (1993). Strain characterization, genome size and plasmid content in the *Lactobacillus acidophilus* group (Hansen and Macquot). J Appl Bacteriol. **74** (5): 549-56

Roux V. And Raoult D.(1995). Inter-and intraspecies identification of Bartonella (Rochalimaea) Species. J Clin Microbiol. **33** (6): 1573-1579

Roux V. Raoult D. (1993). Genotypic identification and phylogenetic analysis of the spotted group rickettsiae by pulsed-field gel electrophoresis. J Bacteriol. **175** (15): 4895-904

Rydkina E, Roux V, Roult D. (1999). Determination of genome size of *Ehrlichia ssp.*, using pulsed field gel electrophoresis . FEMS Microbiol Lett. **176** (1): 73-8

Saunders KE, McGovern KJ, Fox JG. (1997). Use of pulsed-field gel electrophoresis to determine genomic diversity in strains of *Helicobacter hepaticus* from geographically distant locations. J Clin Microbiol. **35** (11): 2859-63

Schwartz E, Friedrich B. (2001). A Physical map of the megaplasmid pHG1, one of the three genomic replicons in *Ralstonia eutropha H16.* FEMS Microbiol Lett. **201** (2): 213-219

Shaheduzzaman SM, Akimoto S, Kuwahara T, Kinouchi T, Ohnishi Y. (1997). Genome analysis of Bacteroides by pulsed-field gel electrophoresis: chromosome sizes and restriction patterns. DNA Res. **4** (1): 19-25

Shao Z, Mages W, Schitt R. (1994). A physical map of the hyperthermophilic bacterium *Aquifex pyrophilus* chromosome. J Bacteriol. **176:** 6776-80

Sitzman J, and Klein A. (1991). Physical and genetic map of the *Methanococcus voltae* chromosome. Mol. Microbiol. **5:** 505-513

Smith LC, Econome,G.J., Schutt,A., Klco,S., Cantor,R.Ch. (1987). A Physical Map of the *Escherichia coli K12* Genome . Science **236:**1448-14553

Songssivila S, Dharakul T. (2000). Multiple replicons constitute the 6.5-megabase genome of *Burholderia pseudomallei.* Acta Trop. **74** (2-3): 9-79

Sowers K. and Gunsalus PR. (1988). Plasmid DNA from Acetotrophic Methanogen *Methanosarcina acetivorans.* J Bacteriol. **170** (10): 4979-4982

Stettler R, Leisinger T. (1992).Physical Map of the *Methanobacterium termoautotrophicum Marburg* chromosome. J Bacteriol. **174** (22): 7227 34

Stibitz S, and Garletts TL (1992). Derivation of a Physical Map of the Chromosome of *Bordetella pertisis Tohama I.* J Bacteriol. **174** (23): 7770-7

Strehl B, Holtzendorff J, Partensky F, Hess RW. (1999). A small and compact genome in the marine cyanobacterium *Prochlorococcus marinus CCMP 1375*: lack of an intron in the gene for tRNA (Leu) uaa and single copy of rRNA operon. FEMS Microbiol Letters. **181:** 261-266

Sun LV, Foster JM, Tzertzinis G, Ono M, Bandi C, Slatko BE, O´ Neill SL (2001). Determination of Wolbachia genome size by pulsed-field gel electrophoresis *Wolbachia strains*. J Bacteriol. **183** (7): 2219-2225.

Suvorov AN, Ferreti JJ. (1996). Physical and genetic chromosomal map of an M type 1 strain of *Streptococcus pyogenes*. J Bacteriol. **178:** 5546-49

Suvorov AN, and Ferretti JJ. (1997). Chromosomal Analysis of group a Streptococci by pulsed field gel electrophoresis. Adv. Exp. Med. Biol. **418:** 979-981

Suzuki S,. Kita-Tsukamoto K, and FukagawaT. (1994). The 16S rRNA sequence and genome sizing of tributyltin resistant marine bacterium, strain M-1. Microbios. **77:** 101-109.

Tabata K. Kosuge T, Nakahara T, and Hoshino T. (1993). Physical Map of the extremely thermophilic bacterium *Thermus thermophillus HB27* chromosome. FEBS **331** (1,2): 81-85.

Takami S, Hayashi T, Tonokatsu Y, Shimoyama T, and Tamura T. (1993). Chromosomal Heterogeneity of *Helicobacter pylori* Isolates by Pulsed-Field Gel Electrophoresis. Zbl. Bakt. **280:** 120-127

Taylor A, Barbour G, and Thomas D. (1991) Pulsed-field gel electrophoretic analysis of Leptospiral DNA. Infection and Immunity. **59** (1): 323-329

Taylor DE, Chang N, Taylor NS, Fox JG. (1994). Genome conservation in *Helicobacter mustelae* as determined by pulsed field gel electrophoresis. FEMS Microbiol. Lett. **118:** 31-36

Taylor DE, Eaton M, Chang N, and Salama SM. (1992). Construction of a *Helicobacter pylori* Genome Map and Demostration of Diversity at the Genome level. J Bacteriol. **174** (21): 6800-6806

Thong KL, Puthucheary SD, Pang T. (1997). Genome size variation among recent human isolates of *Salmonella typhi.* Res Microbiol. **148** (3): 229-35

Tigges E, Minion F. (1994). Physical map of the genome of *Acholeplasma oculi ISM1499* and construction of a Tn 40001 derivative for macrorestriction chromosomal mapping. J.Bacteriol. **176:** 1180-83

Tola S, Idini G, Rocchigiani MA, Rocca S, Manunta D, Leori G (2001). A physical map of the *Mycoplasma agalactiae strain PG2.* Veterinary Microbiology. **80:** 121-130

Trevors JT. (1996) Genome size in bacteria. Antonie Van Leeuwenhoek. **69:** 293-303

Tulloch DL, Finch LR, Hillier AJ, and Davidson Barrie E. (1991). Physical Map of the Chromosome of *Lactococcus lactis subsp. lactis DL11* and localization of Six Putative rRNA Operons. J Bacteriol. **173** (9) : 2768-2775.

Valcarcel J, Allardet-Servent A, Bourg G, O Callaghan D, Michailesco P, Ramuz M. (1997) Investigation of the *Actinobacillus actinomycetemcomitans* genome by pulsed field gel electrophoresis. Oral Microbiol Immunol **12** (1): 33-9

Vary P. (1993). The genetic map of Bacilus megaterium.The genetic map of *Bacillus megaterium* 475-81 In :Sonenshein A,

Hoch J,Losick R, eds. (1993). *Bacillus subtilis* and Other gram positive bacteria. Washington, DC: Am.Soc. Microbiol.

De Villiers EP, Brayton KA, Zweygarth E, Allsopp BA. (2000). Genome size and genetic map of *Cowdria ruminantium.* Microbiology. **146** (pt10) 2627-2634

Wagner E, Doskar J. Gotz F. (1998).Physical and genetic map of the genome of *Staphylococcus carnosus TM300*. Microbiology. **144:** 509-17

Walker EM, Arnett JK, Heath JD, Norris SJ. (1991). *Treponema pallidum subs pallidumhas* a single, circular chromosome with a size of approximately 900 kilobase pairs. Infect Immun. **59** (7): 2476-9

Ward-Rainey N, Rainey FA, Wellington EM, Stackebrandt E (1996) Physical map of the genome of *Planctomyces limnophilus*, a representative of the phylogenetically distinct Planctomycete linage. J Bacteriol. **178** (7): 1908-13

Weisburg WG, Tully JG. Rose DL. Petzel PJ. Oyaizu H. Yang D. Mandelco L, Sechrest J, Lawrwnce GT. Van Etten J, Maniloff J. and Woese CR. (1989). A Phylogenetic Analisysis of the Mycoplasmas: Basis for their Clasification. J Bacteriol. **171** (12): 6455-6467

Widjaja R, Suwanto A, Tjahjono B. (1999). Genome size and macrorestriction map of *Xantomonas campestris pv. Glycines YR32* chromosome. FEMS Microbiol Lett. **175** (1): 59-68

Yamaichi Y, Iida T, Park KS, Yamamoto K, Honda T. (1999). Physical and genetic map of the genome of *Vibrio parahaemoliticus*: presence of two chromosomes in Vibrio species. Mol Microbiol. **31** (5): 1513-21

Yan W, and Taylor DE. (1991). Sizing and Mapping of the Genome of *Campylobacter coli strain UA417R* using pulsed field gel electrophoresis. Gene. **101:** 117-120.

Ye F, Laigret F, and Bové JM. (1994). A physical map of the prokaryote *Spiroplasma melliferum* and its comparison with *Spiroplasma citri* map C.R Acad Sci **317**: 392-398.

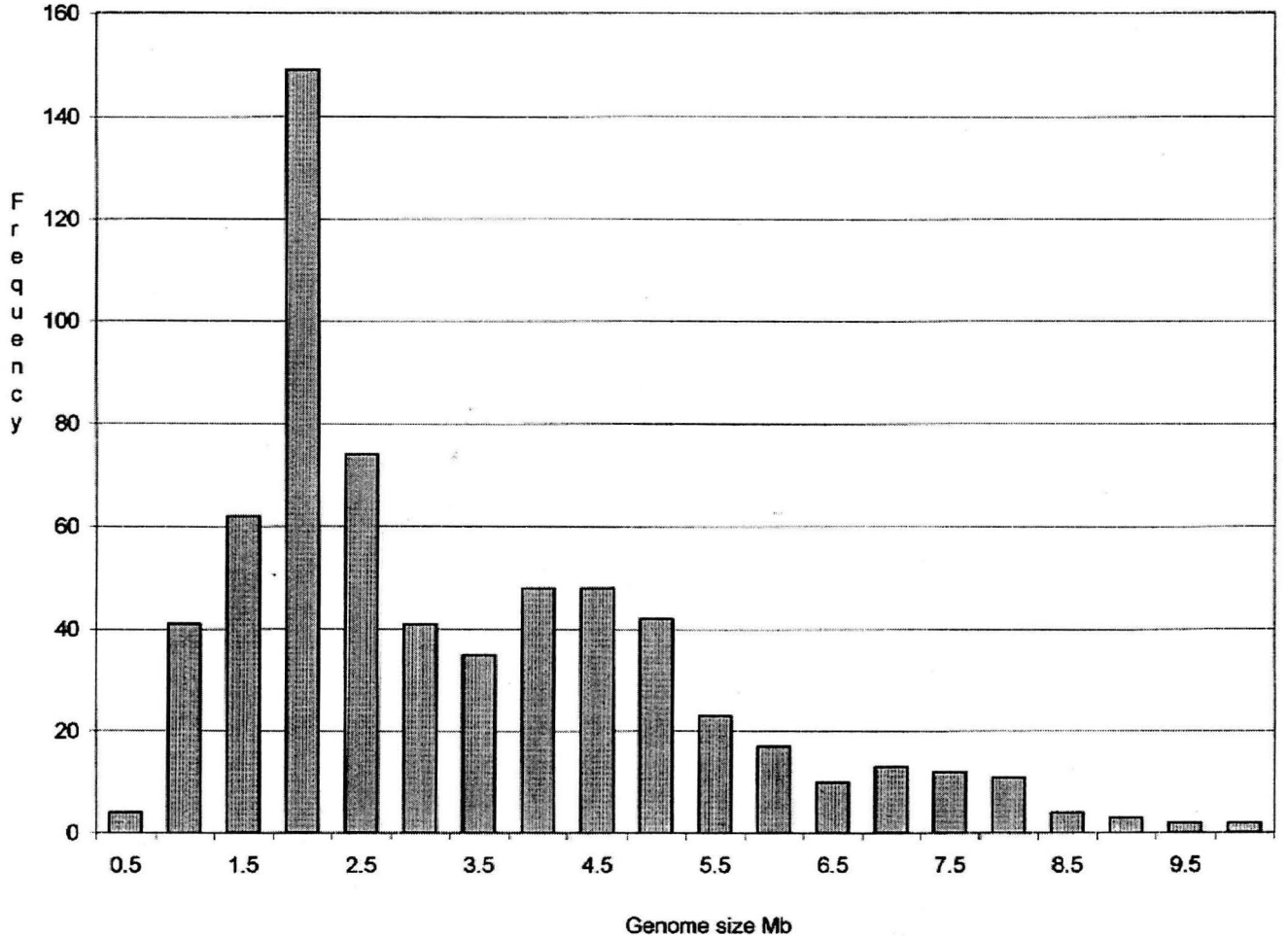
Young M. and Cole In: Abraham L. Sonenshein, Hoch, Loseck (eds) (1993)*Bacillus subtilis* and other Gram positive Bacteria. Biochemestry, Physiology, and Molecular Genetics Am Soc Microbiol. Washington D.C.

Ze-Ze L, Tenreiro R, Brito L, Santos MA, Paveia H. (1998) Physical map of the genome of *Oenococcus oeni PSU-1* and localization of genetic markers. Microbiology **144** (Pt 5): 1145-56

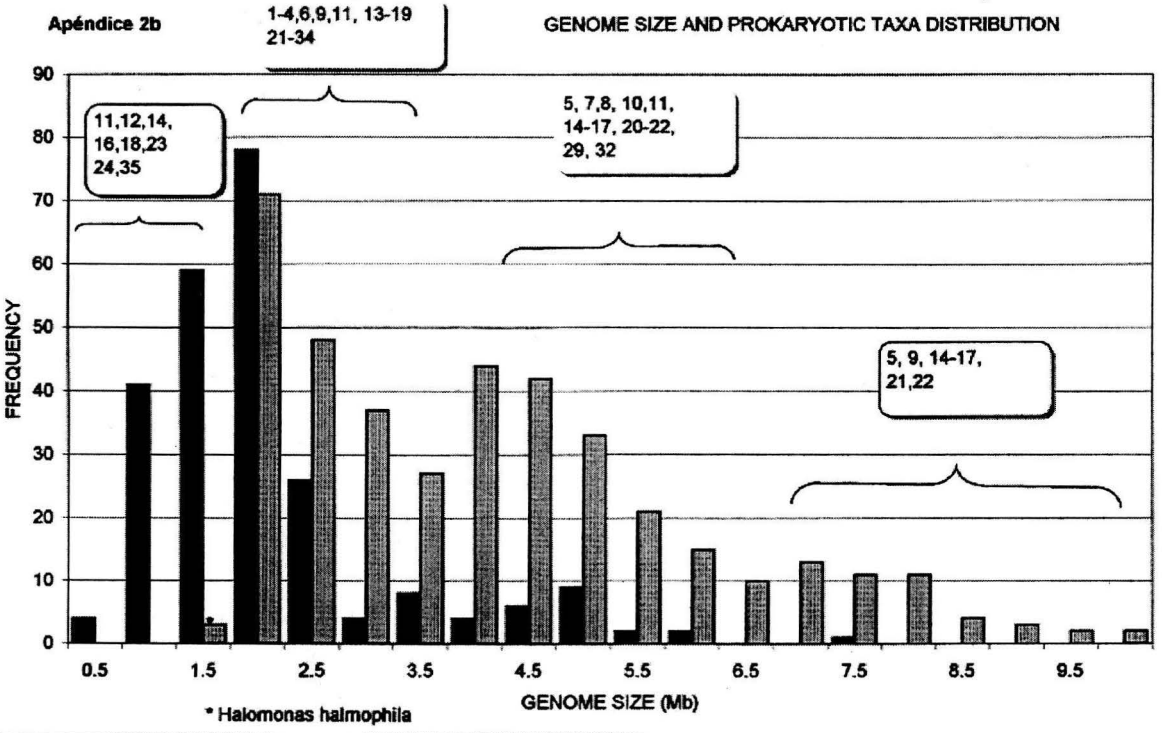
Zhang Y, Geider K. (1997). Differentiation of Erwinia amylovora strains by pulsed-field gel electrophoresis. Appl Environ Microbiol. **63** (11): 4421-6

Zuerner RL, Stanton TB. (1994) Physical and genetic map of the *Serpulina hyodysenteriae B78T* chromosome. J Bacteriol. 176: 1087-92

Apéndice 2a


160
140
120
100
80
60
40
20
0
Frequency
0.5
1.5
2.5
3.5
4.5
5.5
6.5
7.5
8.5
9.5
Genome size Mb

Apéndice 2b



1 Thermophilic Oxygen reducers
2 Thermotogales
3 Thermus/Deinococcus
4 Leptospirilum-Nitrospira
5 CFB
6 Green Sulphur Bacteria
7 Planctomycetales
8 Purple non Sulphur Bacteria
9 Cyanobacteria
10 Fibrobacter and Acidobacterium

11 Spirochetes
12 Chlamydiales/Verrucomicrobia
13 Fusobacteria
14 a proteobacteria
15 b proteobacteria
16 g proteobacteria
17 d proteobacteria
18 e proteobacteria
19 Porochlamydia
20 Unclasifipseudomonas

21 Actinobacteria
22 Bacillus/Clostridium
23 Mycoplasmatales
24 Acholeplasmatales
25 Unclasif Mollicutes
26 Pyrodictiales
27 Thermoproteales
28 Sulfolobales
29 Halobacteriales
30 Methanobacteriales

31 Methanococcales
32 Methanosarcinales
33 Archaeoglobales
34 Thermococcales
35 Unkownarchaea

J Mol Evol (1997) 45:115–118

Apéndice 3[a]

JOURNAL OF MOLECULAR EVOLUTION

Point Counter Point

Polyphyletic Gene Losses Can Bias Backtrack Characterizations of the Cenancestor

Mushegian and Koonin (1996) have recently published the results of a detailed comparison of the complete genomes of *Haemophilus influenzae* and *Mycoplasma genitalium* in conjunction with the fragmentary data from other organisms available as of March 1996. Once parasite-specific sequences were discarded, the final outcome was an inventory of 256 genes that may resemble, not only the genetic complement of the ancestor of Gram-positive and Gram-negative bacteria, but probably also the amount of DNA required today to sustain a minimal cell. Since most of these sequences have eukaryotic and/or archaeal homologs, Mushegian and Koonin discuss how this figure may be reduced to describe the genome of the last common ancestor (LCA) of the Bacteria, Archaea, and Eucarya, that is, the cenancestor, and suggest how insights on even earlier stages of evolution can be achieved. Given the rapid pace at which more and more cellular genomes are being completely mapped and sequenced, the assumptions and strategies used in such approaches merit considerable attention. As argued here, important pitfalls can be avoided if the polyphyletic gene losses that have taken place in widely separated lineages are properly acknowledged.

The Cenancestor Probably Had a DNA Genome

The backtrack methodology proposed by Mushegian and Koonin (1996) is quite straightforward, and partly based on the idea that genes that are not found in both bacteria and eucarya, or in bacteria and archaea, were probably absent from the cenancestor. The nonstated assumption is that the archaea and eucarya are sister groups, an evolutionary relationship supported by an increasingly larger amount of molecular data. However, such an approach can inadvertently miss nuclear-encoded genes which may have been part of the LCA but lost independently in both the bacterial and archeal domains, or not present in the prokaryotic genomes of a given data set. For instance, the absence in their sample of eukaryotic or archaeal homologs of several key proteins involved in DNA replication led Mushegian and Koonin to speculate that the cenancestor may have had an RNA genome. Several objections can be raised against this conclusion: (1) Sequence similarities shared by many ancient, large proteins found in all three domains suggest that considerable fidelity already existed in the operative genetic system of their common ancestor, but such fidelity is unlikely to be found in RNA-based genetic systems. (2) Sequence analysis and biochemical characterization of a ribonucleotide reductase from the archaeon *Pyrococcus furiosus* has shown that this enzyme shares considerable similarities with both its eubacterial and eukaryotic counterparts (Riera et al. 1997). (3) As underlined by Mushegian and Koonin (1996), their analysis was performed before any complete archaeal or eucaryal genomes became available in the public data bases, and should thus be considered preliminary. Indeed, release of the entire *Methanococcus jannaschii* genome has allowed the identification of one archaeal DNA polymerase exhibiting sequence similarity and three conserved motifs with the eubacterial DNA polymerase II, and with the eukaryotic α, γ, and ε polymerases (Bult et al. 1996). Taken together, these results suggest that DNA genomes and polymerases with proofreading and synthesizing functions evolved prior to the divergence of the three primary kingdoms.

To Salvage or Not to Salvage

Until a more complete data set is available, backtrack inferences on the nature of the cenancestor should be considered as preliminary and perhaps biased by the reduced genomic content of parasites, many of which have undergone multiple secondary losses. For instance, the de novo purine nucleotide biosynthesis is probably one of the oldest metabolic pathways, but it is also one of the

most easily lost by a wide range of obligate symbionts and parasites. Failure to recognize such polyphyletic streamlining processes, which have taken place in *H. influenzae* and at an even greater degree in *M. genitalium*, can lead to some misunderstanding. It would be tempting, for instance, to interpret the absence of purine biosynthesis in the minimal set defined by Mushegian and Koonin (1996) as evidence that the growth and reproduction of the first life-forms depended on the heterotrophic uptake of nucleotides present in the primitive soup (see, for instance, Pennisi 1996). However, such conclusions would be at odds with the problems associated with the chemical synthesis and accumulation under primitive conditions not only of ribose, but also of purine and pyrimidine ribosides, which suggest that none of them are truly prebiotic compounds (cf. Lazcano and Miller 1996).

The phylogenetic distribution of purine nucleotide salvage enzymes can also lead to some confusion regarding the cenancestor's metabolic capabilities. Based on their data set, Mushegian and Koonin (1996) conclude that their minimal cell had the complete nucleotide salvage pathways for all bases except thymine. Adenine deaminase (ADA), which catalizes the hydrolytic deamination of adenine into hypoxanthine, is absent in both *H. influenzae* and *M. genitalium*, and, therefore, was not included in such inventory. However, since the ADA gene is found in other nonpathogenic Gram-positive and Gram-negative bacteria, it may have been part of the LCA genome. The same is probably true of the GMP reductase *guaC* gene. Since GMP reductase is not found in *H. influenzae*, *M. genitalium*, *M. jannaschii*, and *Saccharomyces cerevisiae*, it could be argued that the cenancestor lacked *guaC*. Such conclusion is not supported by the presence of GMP reductase in a group of widely separated species that includes *Escherichia coli*, *Tritrichomonas foetus*, *Trypanosoma cruzi*, *Leishmania mexicana*, and humans (Berens et al. 1995). Even organisms with close phylogenetic affinities can differ in their salvage abilities. Hypoxanthine- and guanine phosphoribosyltransferase activities have been found in cell extracts of the euryarchaeota *Methanococcus voltae* (Bowen et al. 1996), but the corresponding genes appear to be absent in the closely related *M. jannaschii*, where the only recognizable purine phosphoribosyltransferase gene is that of adenine PRTase (Bult et al. 1996).

Molecular Phylogenies Are Not Rooted in the Origin of Life

The pioneering work of Mushegian and Koonin (1996) is an important improvement over previous attempts to characterize the LCA (Lazcano 1995 and references therein), but it can be improved by systematic efforts to identify streamlining processes that have led to polyphyletic gene losses in widely separated species. This may be particularly significant given the choice of model organisms whose entire DNA is being sequenced, some of which have been selected because of their relatively small, compact genomes. It is expected that in few years larger volumes of genomic data reflecting a broader cross-section of biological diversity will become available. This will allow not only more precise descriptions of the gene complements of ancestral states, but also an understanding of the effects of parasitism on genomes and the dynamics of gene losses.

Genome sequencing and analysis is rapidly becoming a key element in our understanding of early biological evolution, but it is difficult to see how its applicability can be extended beyond a threshold that corresponds to a period of evolution in which protein biosynthesis was already in operation. Older stages are not yet amenable to this type of analysis. The first life-forms were probably simpler than any cell now alive, and may have lacked not only familiar traits like protein catalysts, but perhaps even genetic macromolecules with ribose–phosphate backbones (Lazcano and Miller 1996). Given the huge gap in our understanding of the evolutionary transition between the prebiotic synthesis of organic compounds and the cenacestor, the temptation to describe the nature of the very first living systems based solely on molecular cladistics should be carefully avoided.

Acknowledgments. We thank Dr. Marc Fontecave and his coworkers for providing us with their results prior to publication. The work of J.I.L. has been supported by the Consejo Superior de Investigaciones Científicas (CSIC, Madrid, Spain). A.L. is an Affiliate of the NSCORT (NASA Specialized Center for Research and Training) in Exobiology at the University of California, San Diego.

References

Berens RL, Krug EC, Marr J (1995) Purine and pyrimide metabolism. In: Marr JJ, Müller M (eds) Biochemistry and molecular biology of parasites. Academic Press, London pp 89–117

Bowen TL, Lin WC, Whitman WB (1996) Characterization of guanine and hypoxanthine phosphoribosyltransferase activities in *Methanococcus voltae*. J Bacteriol 178:2521–2556

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Relch CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geohagen NSM, Weldman JF, Fuhrmann JL, Nguyen D, Utterback TR, Kelley JM, Peterson JD, Sadow PW, Hanna MC, Cotton MD, Roberts KM, Hurst MA, Kaine BP, Borodovsky M, Klenk MP, Fraser CM, Smith HO, Woese CR, Venter JC (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073

Lazcano A (1995) Cellular evolution during the early Archean: what happened between the progenote and the cenancestor? Microbiologia SEM 11:1–13

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

Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 93:10268–10273

Pennisi E (1996) Seeking life's bare (genetic) necessities. Science 272: 1098–1099

Riera J, Robb FT, Weiss R, 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

Arturo Becerra
Sara Islas
José Ignacio Leguina
Ervin Silva
Antonio Lazcano
Facultad de Ciencias
Universidad Nacional Autónoma de México
Apartado Postal 70-407
Cd. Universitaria
04510 México
D. F. México

FALTAN PAGINAS 132-134

Apéndice 3b

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

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

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

1. Introduction

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

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

J. Chela-Flores et al. (eds.), Astrobiology, 135–150.

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

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

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

2. On the nature of eukaryotic cells

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

A major achievement of this approach has been the evolutionary comparison of small subunit ribosomal RNA (rRNA) sequences, which has allowed the construction of a trifurcated, unrooted tree in which all known organisms can be grouped in one of three major (apparently) monophyletic cell lineages: the eubacteria, the archaebacteria, and the eukaryotic nucleocytoplasm, now referred to as new taxonomic categories, i.e.,

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

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

While molecular phylogenies have confirmed the endosymbiotic origin of plastids and mitochondria, a number of trees also suggest that a major portion of the eukaryotic nucleocytoplasm originated from an archaebacteria-like cell whose descendants form the monophyletic eucaryal branch (Gogarten-Boekels and Gogarten, 1994). As asserted by Woese and his collaborators, although the presence of endosymbionts is of critical importance to the eukaryotes, it is undeniable that the latter "have a unique, meaningful phylogeny" (Wheelis et al., 1992). While such view assumes an absolute continuity between the nucleocytoplasm and its direct ancestor, the holistic arguments advocated by Margulis and Guerrero (1991), Cavalier-Smith (1992), and others, emphasize the evolutionary emergence of an novel type of cell as a result of endosymbiotic events. According to the latter, the key transitional event leading to eukaryosis was the evolutionary acquisition of heritable intracellular symbionts, and the eucaryal branch does not represent eukaryotic cells as a whole, any more than fungal hyphae or

phycobionts like the *Trebouxia* algal cells exhibit, by themselves, all the phenotypic and genetic characteristics of a lichen thallus.

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

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

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

From an evolutionary point of view it is reasonable to assume that at some point in time the ancestors of all forms of life must have been less complex than even the simpler extant cells, but our current knowledge of the characteristics shared between the three lines has shown that the conclusion that the last common ancestor was a

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

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

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

This approach was used independently a few years ago by Iwabe et al (1989) and Gogarten et al (1989), who analyzed paralogous genes encoding (a) the two elongation factors (EF-G and EF-Tu) that assist in protein biosynthesis; and (b) the alpha and beta hydrophilic subunits of F-type ATP synthetases. Using different tree-constructing algorithms, both teams independently placed the root of the universal trees between the eubacteria, on the one side, and archaebacteria and eukaryotes on the other. Their results imply that eubacteria are the oldest recognizable cellular phenotype, and imply that specific phylogenetic affinities exist between the archaea and the eucarya.

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

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

Minor differences in the basic molecular processes of the three main cell lines can be distinguished, but all known organisms, including the oldest ones, share the same essential features of genome replication, gene expression, basic anabolic reactions, and membrane-associated ATPase mediated energy production. The molecular details of these universal processes not only provide direct evidence of the monophyletic origin of all extant forms of life, but also imply that the sets of genes encoding the components of these complex traits were frozen a long time ago, i. e., major changes in them are very strongly selected against and are lethal. Biological evolution prior to the divergence of the three domains was not a continuous, unbroken chain of progressive transformation steadily proceeding towards the cenancestor. However, no evolutionary intermediate stages or ancient simplified version of the basic biological processes have been discovered in extant organisms.

Nevertheless, clues to the genetic organization and biochemical complexity of the earlier entities from which the cenancestor evolved may be derived from the analysis of paralogous sequences. Their presence in the three cell lineages implies not only that their last common ancestor was a complex cell already endowed, among others, with

pairs of homologous genes encoding two elongation factors, two ATPase hydrophilic subunits, two sets of glutamate dehydrogenases, and the A and B DNA polymerases, but also that the cenancestor itself must have been preceded by simpler cells in which only one copy of each of these genes existed. In other words, Archean paralogous genes provide evidence of the existence of ancient organisms in which ATPases lacked the regulatory properties of its alpha subunit, protein synthesis took place with only one elongation factor, and the enzymatic machinery involved in the replication and repair of DNA genomes had only one polymerase ancestral to the *E. coli* DNA polymerase I and II.

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

Thus, prior to the early duplication events that led to what may be a rather large number of cenancestral paralogous sequences, simpler living systems existed which lacked the large sets of enzymes and the sophisticated regulatory abilites of contemporary cells. Although lateral transfer of coding sequences may be almost as old as life itself, gene duplication followed by divergence probably played a dominant tole in the accretion of complex genomes, and may have led to a rapid rate of microbial evolution. If its is assumed that the rate of gene duplicative expansion of ancient cells was comparable to today's preser values, which are of 10^{-5} to 10^{-3} gene duplications per gene per cell generation (Stark and Wahl, 1984), the maximum time required to go from an hypothetical 100-gene organism to one endowed with a filamentous cyanobacterial-like genome of approximately 7000 genes would be less than ten million years (Lazcano and Miller, 1994).

Although there are no published data on the rate of formation of new enzymatic activities resulting from gene duplication events under either neutral or positive selection conditions, the role of duplicates in the generation of evolutionary novelties is

well stablished. Once a gene duplicates, one of the copies may be free to accumulate non-lethal mutations and acquire new additional properties, which could lead into its specialization or recruitment into new role. Data summarized here supports the idea that primitive biosynthetic pathways were mediated by small, inefficient enzymes of broad substrate specificity (Jensen, 1976). Larger substrate ranges may had not been a disadvantage, since relatively unspecific enzymes may have helped ancestral cells with reduced genomes overcome their limited coding abilities (Ycas, 1974).

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

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

The recognition that enzyme recruitment may have played a major role in metabolic evolution leads, however, to assume some caution in phylogenetic inferences. Although in some cases metabolic pathways may be sucessfully used to assess the phylogenetic relationship of prokaryotes (DeLey, 1968; Margulis, 1993), the possibility that some of the enzymes of archaic pathways may have survived in unusual organisms (Keefe et al., 1994), or that important portions of extant metabolic routes may have been assembled by a patchwork process (Jensen, 1976), suggest that considerable prudence should be exerted when attempting to describe the physiology of truly primordial organisms by simple direct back extrapolation of extant metabolism.

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

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

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

Indeed, we are still very far from understanding the origin and attributes of the first living beings, which may have lacked even the most familiar features in extant cells. For instance, protein synthesis is such an essential characteristic of cells, that it is frequently argued that its origin should be considered synonymous with the emergence of life itself.

However, the discovery of the catalytic activities of RNA molecules has led considerable support to the possibility that during early stages of biological evolution living systems were endowed with a primitive replicating and catalytic apparatus devoid of both DNA and proteins The scheme may be even more complex, since RNA itself may have been preceded by simpler genetic macromolecules lacking not only the familiar 3',5' phosphodiester backbones of nucleic acids, but perhaps even today's bases (Lazcano and Miller, 1996).

Although molecular cladistics may provide clues to some late steps in the development of the genetic code, it is difficult to see how the applicability of this approach can be extended beyond a threshold that corresponds to a period of cellular evolution in which protein biosynthesis was already in operation. Older stages are not yet amenable to molecular phylogenetic analysis. Although there have been considerable advances in the understanding of chemical processes that may have taken place before the emergence of the first living systems, life's beginnings are still shrouded in mystery. A cladistic approach to this problem is not feasible, since all possible intermediates that may have once existed have long since vanished. The

temptation to do otherwise is best resisted. Given the huge gap existing in current descriptions of the evolutionary transition between the prebiotic synthesis of biochemical compounds and the cenancestor (Lazcano, 1994), it is naive to attempt to describe the origin of life and the nature of the first living systems from the available rooted phylogenetic trees.

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

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

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

In fact, hyperthermophiles not only share the same basic features of the molecular machinery of all other forms of life: they also require a number of specific biochemical adaptations. Any theory on the hot origin of life must address the question of how such traits, or their evolutionary precedessors, arose spontaneously in the prebiotic environment. Such adaptations may include histone-like proteins, RNA modificating enzymes, and reverse gyrase, a peculiar ATP-dependent enzyme that twists DNA into a positive supercoiled conformation (Confalonieri et al., 1993). Clues to the origin of

hyperthermophily may be hidden in this list, and its evolutionary analysis may contribute to the understanding of the rather surprising phylogenetic distribution of the immediate mesophilic descendants of heat-loving prokaryotes, which shows that at least five independent abandonments events of hyperthermophilic traits took place in widely separated branches of universal trees, one of which corresponds to the eukaryotic nucleocytoplasm (García-Meza et al., 1995).

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

5. Final remarks

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

As summarized by Patterson (1988), the theoretical foundations of molecular cladistics have been based on a number of central concepts, most of which were inherited from older disciplines, such as physiology, anatomy, and neodarwinism. Homology, which is one of the key concepts in evolutionary theory, was originally used by Wolfgang Goethe, Ettiene Geoffroy Saint-Hilaire, Richard Owen, and others, to describe structural resemblance to an archetype (Donoghue, 1992). In recent years it has not only been repeatedly confused with sequence similarity (Reeck et al., 1988), but is also used to describe a wider range of possible evolutionary relationships that include species- or gene-phylogeny. In fact, some classes of homology that describe

phenomena at the molecular genetic level may have no exact equivalent in orthodox evolutionary analysis of morphological traits. One such case is paralogy, a term coined by Fitch (1970) to describe the diversification of genes following duplication events.

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

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

The same is true, of course, with universal phylogenetic trees derived from elongation factors (Iwabe et al, 1989). In this case neither set is older than its homologue. In this case, the reconstruction of ancestral character states from dichotomously varying paralogous genes does not comes from the analysis of the outgroup, but may be inferred from the realization that the root of the tree must have been preceded by an even older, more primitive condition in which only one copy of the gene existed, prior to the paralogous duplication. Recognition of this fact is likely to play a central role in future understanding of enzyme evolution during the early Archean. Although it is true that the raw material for molecular cladistic analysis is restricted to sequences derived from living organisms (or from fossil samples from which ancient preserved DNA can be retrieved) and cannot be applied to extinct groups of organisms, the construction of trees derived from archaic paralogous sequences may allow us to infer evolution prior to the ealiest detectable nodes.

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

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

But for all its foibles, the relationship between molecular biology and evolutionary theory has opened new, unsuspected avenues of intellectual exploration. Never before has such a wealth of methodological approaches and empirical data been available to the students of life's phenomena. In part because of this prosperity, systematics and evolutionary biology, two of the most broadly oriented fields of life sciences, are now in a state of intellectual agitation. The symptoms are manifold; it is possible that the traditional species concept may not apply to prokaryotes, time-cherished concepts like that of the existence of kingdoms are under fire, the origin and taxonomic position of genetic mobile elements is unknown. There is an increased awareness that the understanding of the processes underlying the generation of evolutionary novelties and the origin of ontogenic patterns cannot be restricted by classical neodarwinian explanations. We are living in the midst of hectic times in which epoch-making debates are reshaping the future of the life sciences, and the development of a more integrated molecular biology may be a never-ending story. It is said that to wish someone to live in an interesting time is one of the most terrible of all Chinese curses. Whatever the

outcome of current discussions and debates. for biology the putative Oriental curse may turn out to be nothing less than an intellectual blessing.

Acknowledgments

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

6. References

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

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

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

Confalonieri, F., Elie, C., Nadal, M., Bouthier de la Tour, C., Forterre, P., and Duguet, M. (1993) Reverse gyrase, a helicase-like domain and a type I topoisomerase in the same polypeptide, *Proc. Natl. Acad. Sci. USA* **90**, 4753-4758

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

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

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

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

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

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

Forterre, P., Benachenhou-Lahfa, N., Confalonieri, F., Duguet, M., Elie, Ch., Labedan, B. (1993) The nature of the last universal ancestor and the root of the tree of life, still open questions, *BioSystems* **28**, 15-32

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Sogin, M. L. (1994) The origin of eukaryotes and evolution into major kingdoms, in S. Bengtson (ed), *Early Life on Earth, Nobel Symposium No. 84*, Columbia University Press, New York, pp.181-192

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

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

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

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

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

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

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

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

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

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

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

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

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

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