UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES
I Z T A C A L A

FILOGENIA Y EVOLUCION MOLECULAR DE LAS PROTEINAS REGULATORIAS "TWO COMPONENT"

T E S I S

QUE PARA OBTENER EL TITULO DE

B I O L O G A

P R E S E N T A :

SONIA DAVILA RAMOS

1994

Este trabajo se llevó a cabo en el Centro de Investigación sobre Fijación de Nitrógeno y en el Instituto de Biotecnología bajo la asesoría del Dr. J. Enrique Morett Sánchez y con el apoyo de la D.G.A.P.A.

Agradeciendo a Irma Dueñas, Ramón V. Moreno, José Luis Sánchez, Rafael Quintanar, Xavier Soberón y Julio Collado por la revisión y discusiones de este trabajo.

Para empezar agradezco a mi *pa* y mi *ma* por haberme traido a esta familia de locos, por brindarme siempre su apoyo y su amor incondicional.

A mis hermanos por la confianza en esta chiripa, por su apoyo que no ha sido poco y por preparar a mis *pa's* para aguantarme casi todo.

A mis sobrinos por sus sonrisas, besos y abrazos y por dejarme jugar todavía con ellos.

A los cuñaos los cuales han dado más que un granito de arena para hacer crecer a esta familia.

A mis amigas Mónica (chacha) porque desde que me acuerdo esta conmigo, a Grecia por su primo, ahh! y por estos añitos que me has aguantado, A Lidia y Alicia , por esas largas platicas en la E.N.E.P. que no tenían nada que ver con Biología y a todas ellas por estar tan cerca.

A mis amigos Mauricio y Héctor por su eterna amistad.

A Brenda por adoptarme, por revisar y ayudarme con esta tesis (por fin!) y claro por su gran amistad.

A Oscar por su cariño, por las terapias que todavía no acaban y por empujarme y empujarme para ser "mejor".

A Adriana por esas comidas divertidísimas.

A Humberto y Samantha por ser unos amiguitos consentidores y claro por los meses de reventón que son difíciles de olvidar.

A Enrique Morett, por su evidente paciencia, por su amistad y su apoyo.

A Gabriel Moreno, por su gran ayuda académica, por los coscorrones, por abrir tantas ventanas y por su nueva amistad.

A mis compañeros de laboratorio: Lety, Katy, Mari, Humberto Barrios, Humberto Flores, Ricardo, Filiberto, Victor, Joel, Paquito, Cotonete, Gabriel del Río, Gabriel Moreno, Juanita y Paty, por los picotones y por hacer la vida tan agradable.

A la familia Otero Vázquez por todo su cariño.

Y por supuesto a Jesús por la felicidad y el amor que siempre me ha dado los cuales parecen inagotables, al igual que las sonrisas, lágrimas y locuras que guardamos.

Indice

Abreviaturas de los aminoácidos y de los organismos utilizados

NOMBRE	ABREVIATURA CON TRES LETRAS	ABREVIATURA CON UNA LETRA
Alanina	Ala	A
Arginina	Arg	R
Asparagina	Asn	N
Acido Aspártico	Asp	D
Cisteina	Cys	C
Acido Glutámico	Glu	E
Glutamina	Gln	Q
Glicina	Gly	G
Histidina	His	H
Isoleucina	Ile	I
Leucina	Leu	L
Lisina	Lys	K
Metionina	Met	M
Fenilalanina	Phe	F
Prolina	Pro	P
Serina	Ser	S
Treonina	Thr	T
Triptofano	Trp	W
Tirosina	Tyr	Y
Valina	Val	V

Abreviaturas de los nombres de los organismos:

aca, *Azorhizobium caulinodans*; **aeu**, *Alcaligenes eutrophus*; **asp**, *Anabaena* sp. **ath**, *Arabidopsis thaliana*; **atu**, *Agrobacterium tumefaciens*; **avi**, *Azotobacter vinelandii*; **bbr**, *Bordetella bronchiseptica*; **bja**, *Bradyrhizobium japonicum*; **bpe**, *Bordetella pertusis*; **bsp**, *Bradyrhizobium* sp.; **bsu**, *Bacillus subtilis*; **cca**, *Cyanidium caldarium*; **ccr**, *Caulobacter crescentus*; **eco**, *Escherichia coli*; **efa**, *Enterococcus faecium*; **kpn**, *Klebsiella pneumoniae*; **lin**, *Leptospira interrogans*; **mxa**, *M. xanthus*; **pae**, *Pseudomonas aeruginosa*; **pae**, *Porphyridium aerugineum*; **psy**, *Pseudomona syringae*; **pvu**, *Proteus vulgaris*; **rca**, *Rhodobacter capsulatus*; **rle**, *Rhizobium leguminosarum*; **rme**, *Rhizobium meliloti*; **sau**, *Staphylococcus aureus*; **sce**, *Sacaromyces cereviciae*; **sco**, *Streptomyces coelicolor*; **sty**, *Salmonella typhimurium*; **val**, *Vibrio alginolyticus*; **xca**, *Xanthomonas campestris*.

1. INTRODUCCION

1.1. Generalidades del sistema de dos componentes

Las bacterias viven en ambientes donde la variación en algunos parámetros físicos, como la temperatura, osmolaridad, viscosidad o luz; los cambios en la concentración de diversos compuestos químicos; la presencia de organismos hospederos y los niveles de toxinas o nutrientes, ocurren rápida e inesperadamente (Gross *et al,* 1989; Parkinson *et al,* 1992; Stock *et al,* 1990). Las bacterias perciben estos cambios y pueden responder a ellos de muy diversas formas, ya sea movilizándose hacia lugares más favorables; realizando ajustes en su estructura o fisiología, o incluso modificando su expresión genética. Todo esto les permite sobrevivir y multiplicarse (Gross *et al,* 1989; Stock *et al,* 1989).

Gran parte de las respuestas celulares a variaciones en el ambiente, están reguladas por proteínas que forman sistemas de dos componentes. Uno de los cuales recibe la señal específica, mientras que el otro regula la respuesta (Nixon *et al,* 1986).

Se ha demostrado que uno de los componentes tiene actividad de histidin cinasa (HPK, de su abreviatura en inglés, **H**istidin **P**rotein **K**inase) y está localizado generalmente en la membrana citoplasmática, lo que le permite recibir las señales del exterior, transducirlas hacia el citoplasma y así cumplir la función de sensor. La señal es transmitida hacia el segundo componente del sistema, el cual modula la respuesta adecuada a la señal recibida. Este segundo componente es una proteína reguladora de la respuesta (RR) (Stock *et al,* 1990).

Se ha demostrado también, que la transducción de la señal ocurre mediante una cascada de fosforilaciones y desfosforilaciones. Al recibir la señal inductora particular, la proteína HPK se autofosforila en un residuo de histidina (H). Mediante una interacción específica reconoce e interactúa con la proteína RR de su sistema, la cual se autofosforila en un residuo de ácido aspártico (D), desfosforilando a la proteína HPK, sólo en esta forma regula los genes involucrados en generar la respuesta adecuada (Stock *et al,* 1989 y 1990; Gross *et al,* 1989; Nixon *et al,* 1986; Albright *et al,* 1989).

La química de la fosfotransferencia se resume a continuación:

$$HPK + ATP \longrightarrow HPK\text{-}P + ADP$$

$$HPK\text{-}P + RR \longrightarrow HPK + RR\text{-}P$$

$$RR\text{-}P \longrightarrow RR + Pi$$

1.2. Organización estructural de las proteínas del sistema de dos componentes

Al comparar las secuencias de aminoácidos de las proteínas, tanto de la familia de las HPK como de las RR, se identificaron regiones muy conservadas que permitían subdividir a estas en distintos dominios estructurales y funcionales muy característicos. Los dominios más conservados son los que están involucrados en llevar a cabo las reacciones de fosfotransferencia (Fig-1). Otros dominios les confieren características particulares para cada sistema (Parkinson *et al,* 1992).

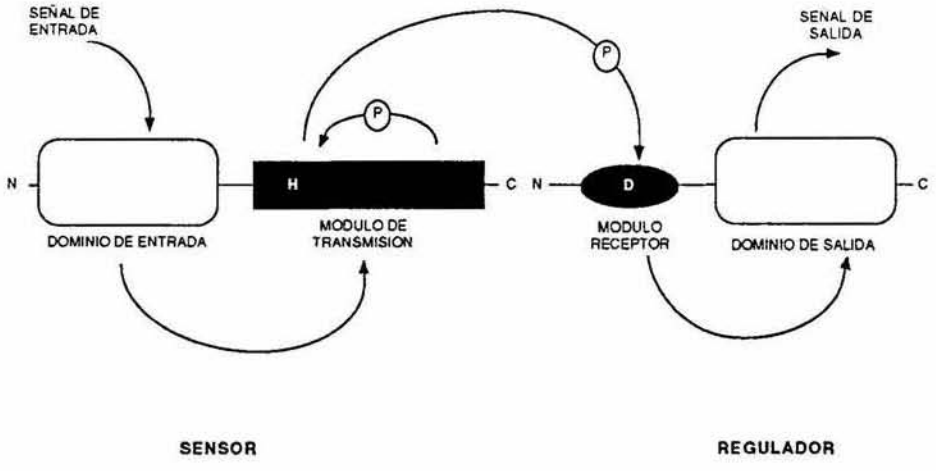
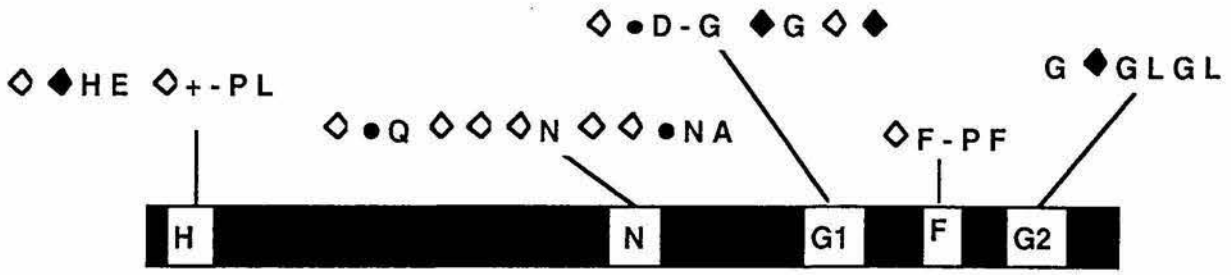


Figura 1. Transmisión de las señales mediante los dominios de comunicación de las proteínas de dos componentes. Los módulos en negro de transmisión y el receptor, están muy conservados. Los dominios en blanco de entrada y salida, son variables.

1.2.1. Proteínas sensoras o histidin cinasas (HPK)

Generalmente, las proteínas HPK tienen dos dominios: El dominio C-terminal es homólogo en toda la familia y está muy conservado. Este dominio se denomina módulo de transmisión, el cual contiene aproximadamente 250 residuos aminoácidos, formando cinco bloques estructurales conservados (Fig-2). El primer bloque es el más variable, y contiene una histidina (H) 100% conservada en la posición 48. El segundo bloque contiene algunas asparaginas muy conservadas, aunque se ignora su función. A continuación se encuentran los bloques G1 y G2, los cuales son ricos en glicina y probablemente forman dominios de unión a nucleótidos. Estos se encuentran separados por el bloque F, el cual contiene dos fenilalaninas, altamente conservadas (Parkinson *et al,* 1992).



Dominio C-terminal de las proteinas HPK

Figura 2. Características estructurales del dominio C-terminal de las proteínas sensoras ortodoxas con sus 5 motivos principales. Las letras indican los aminoácidos presentes en al menos 70% de las proteínas alineadas. Los diamantes indican que por lo menos el 50% de los aminoácidos pertenecen a la misma familia química. Los blancos son no polares (I, L, M, V); los negros son polares (A, G, P, S, T); el signo positivo son básicos (H, K, R); el signo negativo son ácidos o amídicos (D, E, N, Q). Los puntos negros representan menos del 50% de conservación.

El dominio N-terminal de las proteínas HPK es característico para cada sistema. Este dominio no homólogo es muy variable y probablemente está involucrado en recibir una señal de entrada específica; la cual puede ser, por ejemplo, variaciones en los niveles de nitrógeno o de oxígeno. Este dominio de entrada contiene generalmente dos secuencias hidrofóbicas transmembranales, las cuales conectan a la región superficial, que recibe el estímulo, con la región intracelular, que controla la actividad de cinasa en el módulo de transmisión (Ronson *et al,* 1987).

Para poder identificar el dominio y particularmente el residuo donde las proteínas HPK se fosforilan, se realizaron experimentos de marcaje en las proteínas CheA y NRII con [γ-^{32}P] ATP. En estos experimentos se demostró que el residuo marcado por el grupo $^{32}PO_4$ en ambas proteínas fue la histidina 48 (Stock *et al,* 1989).

La familia de las HPK se organiza en dos grupos (Fig-3), el primero presenta la típica topología transmembranal de las proteínas HPK, donde están incluidas VirA, BvgC, FixL, DctB, PhoR, UhpB, PhoM, CpxA, y NarX. Dentro de este grupo VirA de *Agrobacterium tumefaciens,* FixL de *Rhizobium meliloti* y BvgC de *Bordetella pertussis,* mantienen una gran similitud tanto en la región C-terminal, como en la N-terminal (Stock *et al,* 1989; Gross *et al,* 1989).

El segundo grupo de la familia consiste de proteínas citoplásmicas que carecen de dominios transmembranales obvios, este grupo incluye a las proteínas NtrB, DegS y CheA. Se ha propuesto que estas proteínas reconocen o perciben señales internas, o que incluso están moduladas por proteínas sensoras adicionales, necesarias para iniciar la transmisión de una señal. Un ejemplo de lo anterior es la proteína CheA, la cual no percibe directamente los cambios en el ambiente, pero integra las señales recibidas por otras proteínas sensoras, denominadas MCPs o quimiosensores, para posteriormente mandar la señal fosforilando a las proteínas reguladoras CheY y CheB (Ames *et al*, 1988).

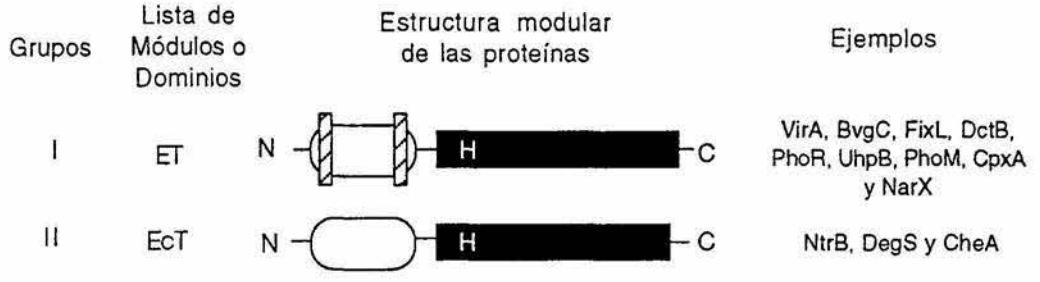


Figura 3. Estructura modular de las proteínas sensoras. En la lista de la izquierda encontramos los módulos o dominios, en orden, a partir del amino terminal. El símbolo **E** representa el dominio de entrada, **T** es el módulo de transmisión y **c** que indica que es citoplásmico. Algunos dominios de entrada tienen sitios de anclaje a la membrana representados por las barras rayadas.

1.2.2. Proteínas reguladoras de la respuesta (RR)

Los alineamientos de secuencias en las proteínas RR demuestran que éstas tienen generalmente dos dominios independientes: un dominio regulador homólogo para toda la familia y muy conservado, denominado módulo receptor, el cual recibe las señales del módulo de transmisión de las HPK, y un dominio no homólogo muy variable, denominado dominio de salida, el cual se localiza hacia el extremo C-terminal.

El módulo receptor consta de 100 a 120 aminoácidos, los residuos aspartato (D) en las posiciones 13, 14 y 57 y la lisina (K) en la posición 109 son invariables. El análisis de la estructura cristalográfica de la proteína CheY, demostró que los aspartatos están muy cerca topológicamente y forman una bolsa ácida en una zona estructurada en forma de hoja β (Fig-4). El aspartato de la posición 57 es el aminoácido que se fosforila preferencialmente, aunque los otros también llegan a fosforilarse. Mutaciones en cualquiera de los aspartatos conservados reducen la capacidad de quimiotaxis en la célula (Parkinson *et al*, 1992; Ronson *et al*, 1987; Gross *et al*, 1989; Stock *et al*, 1990).

El mecanismo de transferencia del grupo fosfato de las proteínas HPK a las RR, se demostró utilizando el sistema OmpR/EnvZ. En estos experimentos se observó que la proteína EnvZ (RR), cataliza tanto la desfosforilación del aminoácido H en OmpR (HPK), como su autofosforilación en el aminoácido D, localizado en el módulo receptor (Stock *et al,* 1989).

Por último, el dominio de salida de las RR es el encargado de iniciar la respuesta al estímulo original. Estas proteínas tienen distintas funciones, dependiendo del sistema de que se trate, tales como la de regular la transcripción, por su unión al DNA o presentar actividad enzimática, entre otras (Parkinson *et al,* 1992).

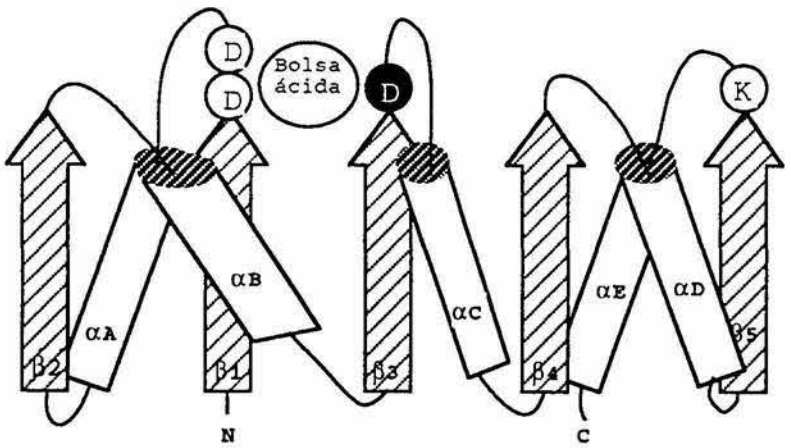


Figura 4. Características estructurales de las proteínas reguladoras. Las flechas indican las hojas β y los cilindros las hélices α. En el círculo negro se encuentra el aspartato, que es el sitio activo donde se lleva a cabo la fosforilación, los demás aminoácidos conservados se encuentran encerrados en los círculos.

Los dominios no homólogos permiten agrupar a esta familia en cuatro grupos principales (Fig-5) (Stock *et al,* 1989; Gross *et al,* 1989; Parkinson *et al,* 1992).

El primer grupo contiene a las proteínas CheY de *Salmonella typhimurium, Escherichia coli*; y Spo0F de *Bacillus subtilis,* que solo están formadas por el módulo de recepción de la señal.

El segundo grupo comprende a las proteínas activadoras de la trascripción, como OmpR, SfrA, PhoB, PhoM, en *E. coli;* OmpR en *S. typhimurium* y VirG de *A. tumefaciens* con un dominio C-terminal no conservado.

El tercer grupo contiene una región C-terminal conservada dentro de este grupo, pero diferente de la del segundo grupo. Este grupo comprende a las proteínas BvgA de *B. pertussis*; DegU de *B. subtilis*; FixJ de *R. meliloti*, NarL y UhpA de *E. coli*.

El cuarto grupo contiene a las proteínas NtrC de *Klebsiella pneumoniae*, *S. typhimurium*, *E. coli*, y *R. meliloti* y DctD de *Rhizobium leguminosarum*, y *R. meliloti*. Este grupo lo forman reguladores transcripcionales que interactúan con el factor σ^{54} mediante su dominio central el cual esta conservado en estas proteínas. Existen otras proteínas como NifA, XylR que también requieren el factor σ^{54} para activar la transcripción, estas proteínas comparten el dominio central de interacción con σ^{54}, pero no el dominio perteneciente a la familia de proteínas de dos componentes (Morett y Segovia 1992).

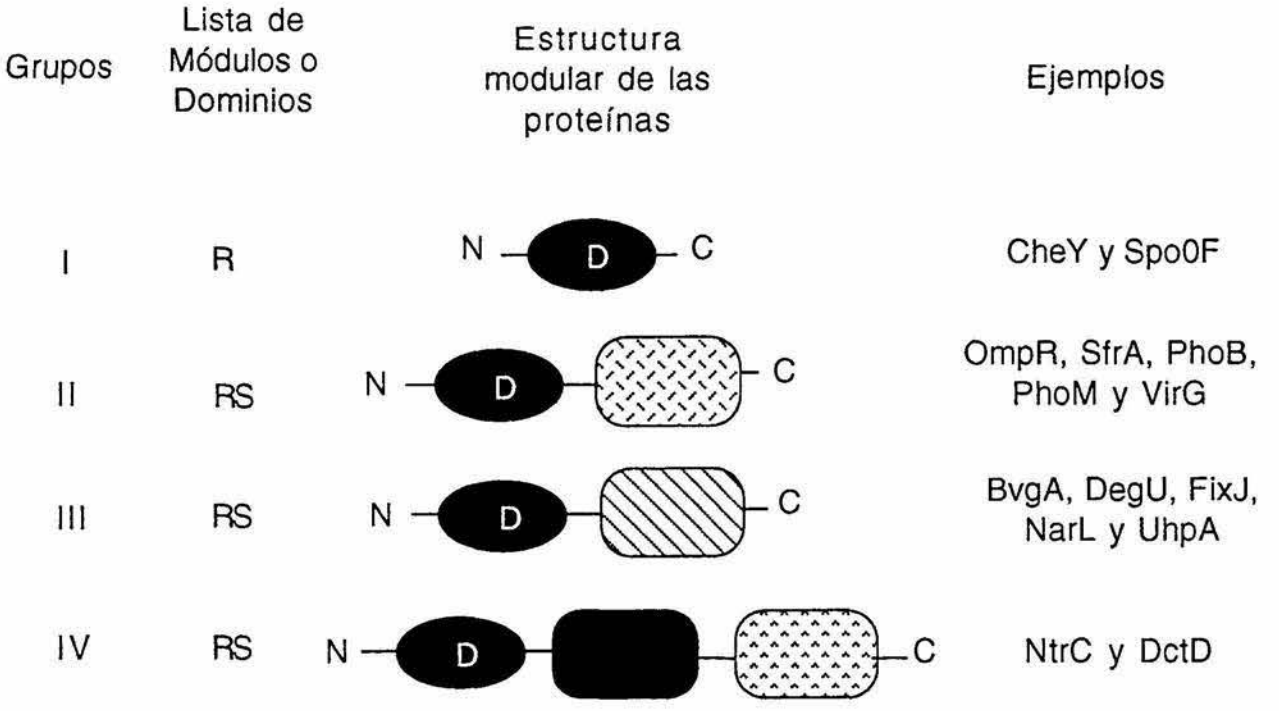


Figura 5. Estructura modular de las proteínas reguladoras. En la lista de la izquierda se observan los modulos en orden de aparición de la región amino hacia la carboxilo.

1.3. Comunicación cruzada

Como ya se mencionó, el sistema de dos componentes funciona mediante reacciones de fosforilación y desfosforilación entre las proteínas HPK y las RR. La gran similitud que existe entre las proteínas de cada familia, particularmente en los aminoácidos donde se llevan a cabo las reacciones de fosforilación y desfosforilación, demuestran que estas proteínas son homólogas, es decir, que provienen de un ancestro común. Debido a esta homología se propuso la hipótesis de que dentro de una misma célula, las proteínas de dos componentes forman dos capas; las de las cinasas y las de las reguladoras; y que cada una de las proteínas de estas capas podía comunicarse con cualquier proteína de la otra capa, formando una red de fosfotransferencia. La comunicación entre proteínas de distintos sistemas HPK-RR, se denomina comunicación cruzada (Fig-6). Sin embargo, se ha observado que cada una de las proteínas han formado distintos sistemas que responden a señales específicas y que si se llega a dar la comunicación entre sistemas diferentes, esto no tiene efectos fisiológicos importantes (Stock *et al,* 1989).

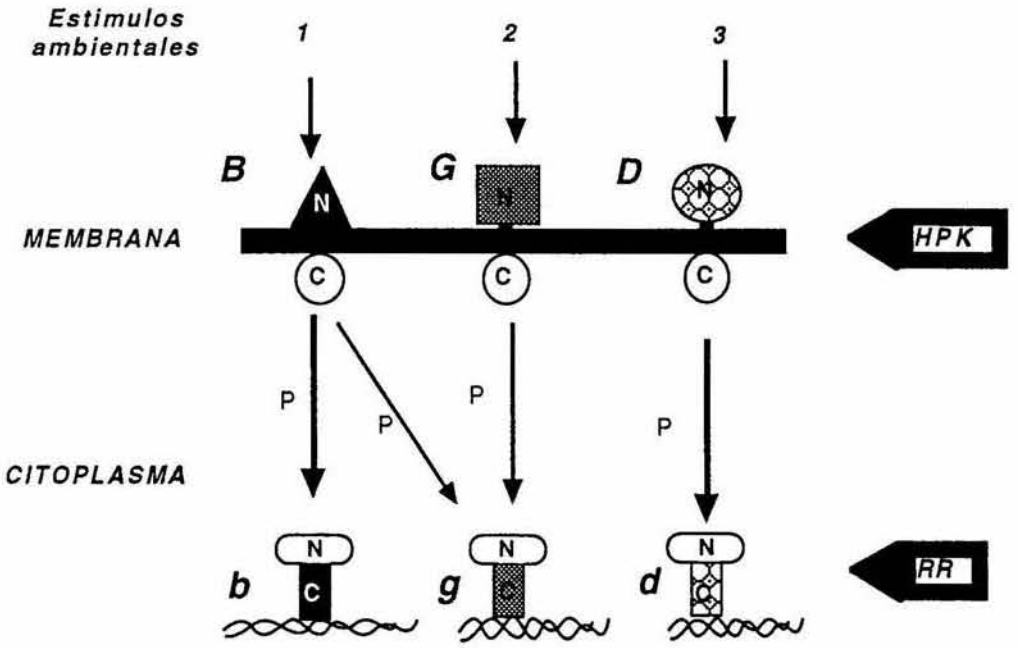


Figura 6. Representación esquemática, de lo que puede ser la "comunicación-cruzada". La proteína "**B**" que es una **HPK** (histidin-protein-cinasa) puede fosforilar tanto a "**b**" como parcialmente a "**g**" que son proteínas **RR** (reguladoras de la respuesta). La transferencia de "**B**" a "**g**" origina que "**g**" induzca la expresión genética en respuesta a un estímulo diferente del que normalmente responde. Las 3 proteínas responden a distintos estímulos ambientales (1, 2 y 3). La **C** indica la región carboxilo-terminal de las proteínas y la **N** la región amino-terminal.

La comunicación cruzada se ha observado en mutantes que carecen de una HPK determinada o que exhiben bajos niveles de expresión. Así, la respuesta de estos sistemas depende principalmente de la concentración de proteína RR que haya en la célula. Por ejemplo, cuando una célula se encuentra en un ambiente de baja concentración de fosfato, la proteína sensora PhoR activa a la proteína reguladora PhoB. Mutantes *phoB*$^-$ no activan el regulón de fosfato. Por otro lado, mutantes en el gene *phoR*, no responden a la concentración de fosfato, pero sí a la presencia de la proteína CreC. La estructura de la proteína CreC es muy parecida a la de PhoR. La expresión debida a la comunicación cruzada con CreC se da a muy bajos niveles y sólo se puede detectar en células mutantes de PhoR (Wanner 1992). Uno de los experimentos más sobresalientes que probaron la existencia de comunicación cruzada, fué el que se realizó con sistemas heterólogos *in vitro*, en donde la proteína sensora NR_{II} es capaz de fosforilar a la proteína CheY en ausencia de su activador CheA y la proteína CheA es capaz de fosforilar a la proteína NR_I también en ausencia de la proteína NR_{II} (Ninfa *et al*, 1988). Otro ejemplo de comunicación cruzada *in vivo* fué observado en sistemas heterólogos formados por las proteínas HPK, Bvg de *Bordetella* y AlgR de *Pseudomonas aeruginosa*, en donde se probó que ambas proteínas son capaces de activar a la proteína OmpR de *E.coli* (DeVault 1989).

La magnitud y el tipo de la conexión entre estos sistemas de dos componentes depende de muchos factores, como la concentración de las HPK, su actividad de autofosforilación, el grado de fosfotransferencia de la HPK-fosforilada a la proteína RR, la concentración de la proteína RR y la competencia de la HPK por una proteína RR dada. Todo esto juega un papel importante, tanto en la funcionalidad del sistema mismo, como en la probabilidad de que los eventos de comunicación cruzada se lleven a cabo. Es necesario mencionar que ciertas cinasas no mandan su señal a un solo regulador, un ejemplo es la proteína CheA, que es capaz de fosforilar tanto a CheB como a CheY, sin embargo, esto no se considera un evento de comunicación cruzada ya que las dos proteínas RR responden al mismo estímulo de quimiotaxis (Stock *et al*, 1990).

La comunicación que han tenido las proteínas del sistema de dos componentes a lo largo del tiempo, y la formación de módulos de transmisión de la señal relativamente específicos, nos inducen a pensar que las proteínas de las dos familias han mantenido una historia evolutiva común, no como proteínas independientes, sino como un sistema que permite tanto sensar, como regular las respuestas a los cambios en el ambiente.

Para poder entender la historia evolutiva de estos sistemas, es necesario revisar algunos conceptos evolutivos, así como las herramientas más frecuentes para realizar este tipo de estudios.

1.4. Evolución molecular

1.4.1. Paleontología, taxonomía y evolución molecular

Las preguntas centrales en los estudios de evolución están enfocadas en conocer cuál o cuáles han sido los mecanismos por los que ocurre y cómo ha sido la historia evolutiva de los organismos. Diversas disciplinas científicas han tratado de resolver estas preguntas mediante investigaciones que van desde la clasificación taxonómica de algunos organismos actuales, hasta la reconstrucción filogenética de sus ancestros mediante técnicas de biología molecular (Futuyma 1986).

La primera evidencia tangible de que los organismos han ido cambiando a lo largo del tiempo son los restos fósiles. Se han propuesto teorías acerca de la secuencia evolutiva de algunos linajes, las cuales se basan tanto en el tiempo geológico en donde se localizan los fósiles, como en las características morfológicas y taxonómicas que muestran. Sin embargo, no se puede ni podrá tener un registro fósil completo, ya que no están representadas todas las especies, ni todos los intermediarios evolutivos de cada linaje. En otras palabras, el registro fósil sólo nos puede indicar la evolución de los organismos de manera muy general y fragmentada, por lo que es necesario complementar esta información con otro tipo de estudios (Futuyma 1986; Li *et al,* 1991).

A finales del siglo pasado y principios del actual, muchos microbiólogos se interesaron en realizar estudios de evolución tratando de determinar la genealogía de los microorganismos mediante parámetros morfológicos, al igual que los botánicos y zoólogos determinaban la de plantas y animales. Sin embargo, los microbiólogos se encontraron con el problema de que algunos microorganismos, como las bacterias, son morfológicamente muy simples y esto les impedía realizar clasificaciones o inferencias evolutivas certeras (Woese 1987). Con el tiempo se comenzaron a utilizar características fisiológicas para relacionar a los microorganismos, aunque algunas de ellas se perdían o podían ser obtenidas por transferencia horizontal, por lo que resultaron no ser suficientes para los estudios de evolución.

Debido a estas complicaciones en los estudios de clasificación y evolución de las bacterias, tuvieron que pasar muchos años para poder diferenciar, clasificar y además comparar a las bacterias con otros organismos de manera más certera. Esto solo ha sido posible gracias al desarrollo de la biología molecular. Mediante las técnicas generadas en esta disciplina, ha sido posible obtener una gran cantidad de información a partir de algunas moléculas, ácidos nucleicos o proteínas, las cuales se consideran los documentos históricos de las células, ya que guardan en su secuencia su historia evolutiva (Zuckerkandl *et al,* 1965).

La comparación de las moléculas de distintos organismos, permiten proponer las relaciones filogenéticas que existen entre ellos y con esto inferir el camino evolutivo que han seguido a lo largo del tiempo. En los últimos 10 años el número de genes secuenciados se ha incrementado de manera sorprendente. Actualmente, se conocen alrededor de 32000 secuencias y siguen aumentando. Es por esto que ha sido necesaria la creación de bancos de datos, como los de EMBL, GenBank y SwissProt, los cuales archivan las secuencias de manera organizada (Doolittle 1990). Estos bancos de datos sirven de plataforma de investigación para el estudio y análisis de grupos de secuencias específicos. Incluídos los análisis filogenéticos mediante la comparación de proteínas o genes homólogos.

Las primeras moléculas que se utilizaron en los estudios de la evolución molecular de los organismos fueron el citocromo-C y las globinas. De la comparación de la secuencia primaria de estas proteínas se pudieron inferir árboles filogenéticos, los que sorprendentemente coincidían en términos generales con los árboles obtenidos por otros criterios, como los morfológicos y los bioquímicos (Woese 1987).

Los estudios evolutivos llevados a cabo antes de incluir la metodología en biología molecular estaban restringidos al estudio de los organismos superiores. Por lo que solo permitían conocer la historia evolutiva de los últimos 340 millones de años. Los estudios de restos fósiles en bacterias parecidas a las Cyanobacterias actuales, han permitido determinar que las bacterias existen sobre la tierra desde hace por lo menos 3400 millones de años. Al incluir a las bacterias en los estudios de evolución, se amplía tanto el espectro de tiempo, como la posibilidad de conocer algunas de las características del ancestro común de todos los organismos existentes (Nei 1987).

Sin embargo, para poder incluir a las bacterias en un estudio general, fué necesario encontrar una molécula común a todos los organismos y que cumpliera con ciertos criterios, como el de ser homóloga y el de tener diferentes velocidades de cambio en distintas regiones de su secuencia. De esta manera, las regiones más conservadas permitirían hacer comparaciones de la molécula entre organismos que se encuentran muy alejados evolutivamente, mientras que las regiones variables permitirían diferenciar a los organismos que divirgieron recientemente (Olsen *et al*, 1986).

Un grupo de moléculas que cumple con estos requisitos son los RNA ribosomales (rRNA). Estas moléculas son universales, los procariotes tienen tres rRNA de distintos tamaños, el 5S, el 16S y el 23S. El 5S, a pesar de cumplir con todas las características anteriores, carece de suficiente información como para realizar estudios de evolución, pues sólo tiene 120 nucleótidos. En cambio, el 16S (1500 nucleótidos) y 23S (2900 nucleótidos) sí contienen suficiente información (Olsen *et al*, 1986).

Woese (1987), utilizando secuencias parciales de RNA ribosomal 16S (Apéndice-3), concluyó que: a) todos los organismos actuales provienen de 3 líneas principales: los eucariontes, las eubacterias y las arqueobacterias y b) los organelos en las células eucariontes tienen un origen procarionte, lo cual apoya la teoría endosimbiótica del origen de los organismos eucariontes propuesta por Margulis (1970).

Esta teoría propone que los eucariontes se generaron de la fusión de cosimbiontes bacterianos, ya que la información genética en organelos, como las mitocondrias y los cloroplastos, es diferente a la que había en los cromosomas. Estos resultados han sido corroborados con algunos datos obtenidos de investigaciones morfológicas y bioquímicas (Margulis 1970; Woese 1987).

Actualmente, la evolución a nivel molecular comprende dos áreas de estudio principales (Li 1991):

La primera trata de la evolución de las macromoléculas; esta estudia tanto la velocidad como el tipo de cambios que ocurren en el material genético (DNA) y en sus productos (proteínas), durante un tiempo determinado, es decir, el estudio de las causas y efectos de los cambios evolutivos de las moléculas y los mecanismos responsables de estos cambios.

La segunda área, conocida como filogenia molecular, utiliza a las moléculas como instrumento para reconstruir la historia o relaciones evolutivas de las macromoléculas y la de los organismos que las contienen relacionandolas con sus ancestros.

1.4.2. Sustituciones en nucleótidos y el reloj molecular

El número de cambios que sufren las secuencias con respecto al tiempo permite calcular la velocidad de evolución de las moléculas. En los estudios comparativos entre las secuencias de hemoglobina y de citocromo-C de diferentes especies, se observó que la velocidad de sustitución de aminoácidos en estas proteínas es aproximadamente el mismo en varias lineas de mamíferos (Woese 1987), es decir, que el número de sustituciones se incrementa de manera lineal con respecto al tiempo. A este fenómeno se le denominó reloj molecular y esto nos permite fechar algunos eventos evolutivos que han sucedido a lo largo del tiempo (Zuckerkandl y Pauling 1962; Margoliash 1963).

Poco tiempo después, se llegó a la conclusión de que cualquier proteína podía ser un reloj molecular. Este descubrimiento ha tenido importantes implicaciones para el estudio de la evolución; puede ser usado no sólo para estimar a *grosso modo* los tiempos de evolución de varios organismos, sino también para construir árboles filogenéticos.

Estudios recientes indican que los relojes evolutivos no corren tan regularmente como se pensó originalmente, ya que las moléculas pueden tener distintas velocidades de cambio. Sin embargo, esto no afecta seriamente la utilidad de los datos moleculares para realizar estudios evolutivos, porque aunque no se conozca con exactitud el tiempo que ha pasado desde que las moléculas divirgieron, si se pueden obtener las relaciones evolutivas entre las moléculas en estudio (Nei 1987).

1.4.3. Teorías evolutivas y los mecanismos de la Evolución

En "El origen de las especies" (1859), Darwin propuso su teoría de la evolución por selección natural. En este libro concluye que las especies han cambiado a lo largo del tiempo y que los organismos son el resultado, tanto de la acumulación, como de la preservación de variaciones sucesivas favorables. Cuando las leyes de Mendel fueron redescubiertas y se determinó que las mutaciones son la fuente de la variación genética, se unificó el principio en biología, estructurándose la teoría sintética de la evolución o neo-darwinismo (1930-1940).

En su tiempo el neo-darwinismo fue aceptado como dogma de la biología evolutiva. La selección natural era considerada como la única fuerza capaz de dirigir el proceso de evolución (seleccionismo), ya que se pensaba que otros factores, como la mutación y la deriva genética al azar, no contribuían de manera importante. De acuerdo con los seleccionistas, para poder entender el proceso evolutivo, era necesario entender que la fijación de sustituciones genéticas en una población ocurre como consecuencia de la selección de mutaciones ventajosas, y que el polimorfismo genético es mantenido por un balance en la selección de las mutaciones (Nei 1987). Esta teoría tuvo mucho auge a finales de los 50's y a principios de los 60's, fue dominada por el punto de vista de que la velocidad y la dirección de la evolución estan determinadas casi completamente por la selección natural, donde las mutaciones juegan un pequeño papel. Muchos evolucionistas de ese tiempo creían que las mutaciones neutrales, es decir las mutaciones que no confieren ventajas ni desventajas, o no son frecuentes, o no existen.

Sin embargo, al comparar la evolución desde los puntos de vista fenotípico y molecular, se observan diferencias en la velocidad y en la manera en que los cambios ocurren. En la evolución molecular la velocidad de cambio es más o menos constante, en tanto que la fenotípica presenta una velocidad irregular y sucede de manera oportunista y a saltos. Esto puede entenderse si se asume que la evolución fenotípica esta dada principalmente por selección natural positiva. En tanto que la evolución molecular se ve afectada por la fijación azarosa de mutaciones neutrales o casi neutrales (Kimura 1974; 1981).

En contraste con la teoría de la selección natural, la teoría neutral asegura que la gran mayoría de los cambios evolutivos a nivel molecular son causados por la fijación de cambios neutrales que incluso pueden llegar a favorecer la selección de ciertas mutaciones ventajosas. Esta teoría no niega el papel de la selección natural para determinar el curso de la evolución adaptativa, aunque asume que sólo una pequeña fracción de los cambios en el DNA (o RNA) son adaptativos y que la mayoría de los alelos polimórficos se mantienen en las especies por el balance que existe entre la mutación y el grado de extinción. En otras palabras, la teoría apoya la idea de que el polimorfismo de las moléculas es una fase transitoria de la evolución de las mismas, rechazando la idea de que el polimorfismo es el resultado de adaptaciones mantenidas por balances de selección.

Algunos años atrás Kimura y Ohta (1974), explicaron la evolución molecular mediante los siguientes puntos:

-La evolución medida o dada por sustituciones de aminoácidos o nucleótidos es aproximadamente constante por sitio, por año y en varios linajes.

-Las partes funcionalmente menos importantes de una molécula evolucionan más rápidamente que las más importantes, ya que las mutaciones neutrales se fijan con mayor probabilidad en moléculas o parte de moléculas menos comprometidas con la función o la estructura. Es decir, la evolución molecular tiende a ser conservativa.

-Los genes con una nueva función generalmente provienen de una duplicación de una parte del genoma, aunque los genes duplicados pueden sufrir mutaciones deletereas.

-Debido a su tamaño, el genoma de los organismos superiores es más flexible que el de otros organismos y el número de copias de una secuencia de DNA puede aumentar o disminuir rápidamente bajo ciertas condiciones, lo que los hace más polimórficos.

-La variabilidad genética es mayor en poblaciones grandes que en pequeñas.

1.4.4. Alineamientos de secuencias

Para poder determinar las relaciones filogenéticas a nivel molecular entre diversos organismos, es necesario identificar y comparar genes o proteínas (completos o parte de ellos) que sean homólogos, pues sólo las moléculas o parte de ellas que lo sean, contendrán información importante para poder establecer relaciones evolutivas, ya que provienen de un ancestro común.

Generalmente las regiones homólogas están separadas por regiones no homólogas, que representan eventos de inserción o deleción que han ocurrido en algunas de las secuencias desde que divirgieron. Para conocer cual de las secuencias sufrió el cambio es necesario compararlas con otras (Devereux *et al*, 1984).

El primer paso en tales comparaciones consiste en el alineamiento de secuencias (Fig-7). Cuando se alinean moléculas de DNA, se utiliza una matriz que permite evaluar las diferentes sustituciones que puede sufrir un nucleótido, como son: transversiones (de purinas a pirimidinas y viceversa); transiciones (de purinas a purinas y de pirimidinas a pirimidinas), las cuales ocurren con mayor frecuencia que las primeras y por último las inserciones o deleciones (indeles). A cada uno de estos cambios se les asignan valores diferentes (Kimura 1981). Los alineamientos con proteínas son más complicados ya que debe tomarse en cuenta que cada aminoácido puede ser sustituido por otros 19 y aún por si mismo.

Existen diversas tablas que asignan un cierto valor cuando las sustituciones se llevan a cabo entre aminoácidos equivalentes en grupos funcionales o en carga, y un valor diferente cuando no lo son. Otras tablas se basan en el código genético, determinando el número de cambios necesarios en un triplete para modificar el aminoácido que se tenía. La más utilizada es la tabla de Dayhoff, o matriz **PAM250** (**P**orcentaje de **M**utaciones **A**ceptadas) (Apendice-2). Esta tabla se generó de un análisis empírico de la frecuencia de sustitución de los aminoácidos, la cual asigna valores positivos a las sustituciones que conservan el aminoácido o a aquellas sustituciones más frecuentes, las que generalmente son las que afectan menos la funcionalidad o la estructura de una proteína, y valores negativos a los que causan un cambio drástico. Por ejemplo, en aminoácidos como cisteína o triptófano, al mantenerlos conservados se les asignan valores positivos muy altos, debido a que estos aminoácidos están implicados tanto en la estructura como en la función de las proteínas. Si alguno de ellos se sustituyera por otro aminoácido, la proteína sufriría un cambio importante y a estos se les asignan valores negativos (Doolittle 1990).

Los alineamientos pueden ser locales o globales. Los locales encuentran el segmento en dos secuencias donde existe la mayor similitud, sin tomar en cuenta el resto de la misma (Smith y Waterman, cita 25), con la desventaja de que no utilizan la información de los sitios o estructuras menos conservadas. Los alineamientos globales, en cambio, aparean dos secuencias completas tratando de aumentar la similitud entre ellas, con el mínimo número de inserciones o deleciones (Needleman y Wunsch, cita 25). Este último método se utiliza para realizar alineamientos múltiples, ya que permite agrupar varias secuencias y así descubrir las regiones conservadas y enfatizarlas.

Otro método para realizar alineamientos múltiples es el de Feng y Doolitle, el cual primero elige las dos secuencias más parecidas entre sí y las agrupa, la tercer secuencia más parecida a este par se alinea con el par anterior formando un nuevo grupo; a estas tres se añade la siguiente secuencia más parecida y así sucesivamente hasta obtener el alineamiento completo.

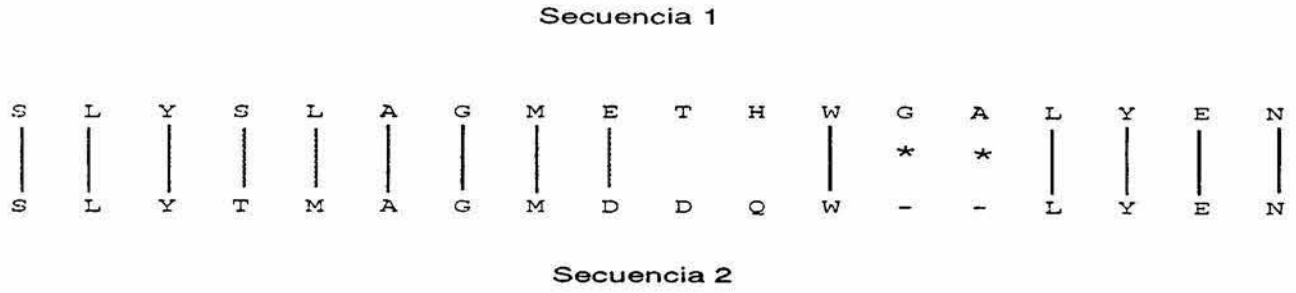


Figura 7. Alineamiento de dos secuencias de aminoácidos. Las líneas punteadas indican sustituciones en las secuencias, las líneas sólidas indican aminoácidos conservados y los asteriscos indican inserciones o deleciones.

1.4.5. Los árboles filogenéticos

Una vez obtenido el alineamiento múltiple, el paso siguiente es la construcción de un árbol filogenético. Un árbol filogenético es una representación gráfica de las relaciones evolutivas entre las unidades taxonómicas operativas (OTUs), las cuales pueden ser especies, poblaciones, individuos o moléculas. Esta gráfica está compuesta por nodos y ramas (Fig-8). Los nodos externos representan las OTUs actuales, y los nodos internos representan las unidades ancestrales de las OTUs (Figura 8-I). Las ramas muestran las relaciones entre los nodos en términos de ancestro y descendiente. Al patrón que presentan las ramas se le denomina topología.

Un árbol filogenético carece de escala evolutiva cuando el largo de las ramas no es proporcional al número de cambios evolutivos que se indican en ellas (Figura 8-I y 8-II), y tiene escala cuando los cambios son proporcionales al largo de las ramas. Por lo tanto, un árbol es aditivo si la distancia entre cualquier par de unidades taxonómicas es igual a la suma del largo de todas las ramas que los conectan.

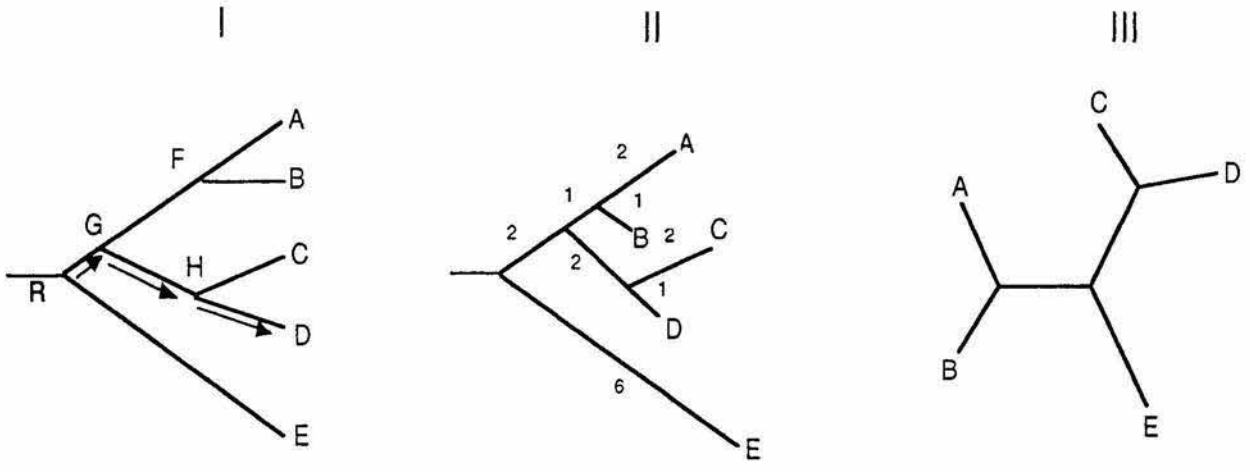


Figura 8. Características de los distintos árboles filogenéticos. En el árbol filogenético I, podemos observar que sus ramas no tienen escala, los nodos F, G y H son los nodos internos y los nodos A, B, C, D y E, son los nodos externos, la R significa que es un árbol con raíz y las flechas indican hacia donde se puede dirigir las ramas a partir de la raíz. En el árbol filogenético II, se observa que las ramas si tienen escala por lo que el largo de las ramas son proporcionales al número de cambios moleculares. El árbol III, es un árbol sin raíz y sin escala.

Los árboles filogenéticos pueden ser árboles con o sin raíz (Figuras 8-I y 8-III). La raíz de un árbol nos indica el ancestro común de todos los OTUs de ese estudio. La distancia que existe entre la raíz y cualquiera de los OTUs corresponde a un tiempo evolutivo. La raíz de un árbol no es otra cosa que una OTU que se encuentra evolutivamente más alejado de las OTUs que se esten analizando. En un árbol sin raíz solo se especifican las relaciones entre los OTUs y no se define un origen para estos. Si el análisis de un grupo de secuencias se realiza con raíz, aumenta la cantidad de árboles alternativos, debido a que las OTUs se pueden acomodar de distintas formas a partir de la raíz. Al tener una gran cantidad de árboles posibles, es necesario escoger aquel que represente mejor los datos obtenidos. Los métodos de reconstrucción filogenética están diseñados para resolver este problema (Li *et al*, 1990; Olsen, 1994).

1.4.6. Métodos de reconstrucción filogenética

Existen tres métodos principales para inferir filogenias, estos son los métodos de máxima parsimonia, los métodos de matriz de distancia y los métodos de máxima probabilidad. Otros métodos que se utilizan derivan de alguno de estos tres grupos (Nei 1987; Li 1991; Felsenstein 1988).

a) Método de Máxima Parsimonia

Este método se desarrolló originalmente para comparar secuencias de aminoácidos (Eck y Dayhoff 1966) y se modificó posteriormente para analizar secuencias de nucleótidos (Fitch 1967). El principio de la máxima parsimonia involucra la construcción de un árbol, en el cual los puntos de unión de dos secuencias representan a un ancestro común, obtenido con el mínimo número de cambios. Este método escoge las secuencias más parecidas entre sí, e identifica después los sitios informativos de las mismas para poder decidir la topología del árbol. Un sitio es filogenéticamente informativo, sólo si favorece algunos árboles sobre otros.

Por ejemplo, si tenemos cuatro secuencias, sólo se pueden obtener tres diferentes árboles sin raíz. Se acomoda cada sitio (entendiendo por sitio a cada nucleótido o aminoácido), de acuerdo con la posición que tiene la respectiva secuencia en cada uno de los árboles posibles. Posteriormente, se determina en cuantas ramas es necesario hacer un cambio para pasar de un sitio a otro y si en un árbol hay un menor número de cambios en comparación con los otros árboles posibles, el sitio es informativo. Por lo tanto, un sitio sólo puede ser informativo cuando hay dos o más tipos de nucleótidos o aminoácidos y además está representado en al menos dos secuencias. Como puede verse estos sitios permiten la elección del árbol de máxima parsimonia (Fig-9).

Este método ha sido muy utilizado, ya que permite reconstruir la historia evolutiva de los organismos, con un buen intervalo de confianza y de manera congruente con otros árboles, aunque utiliza estrategias diferentes como se pudo observar en la explicación.

Secuencia	Sitio 1	2	3	4	5	6	7	8	9
1	A	A	G	A	C	T	G	C	A
2	A	G	C	C	C	T	G	C	G
3	A	G	A	T	T	T	C	C	A
4	A	G	A	G	T *	T	C *	C	G *

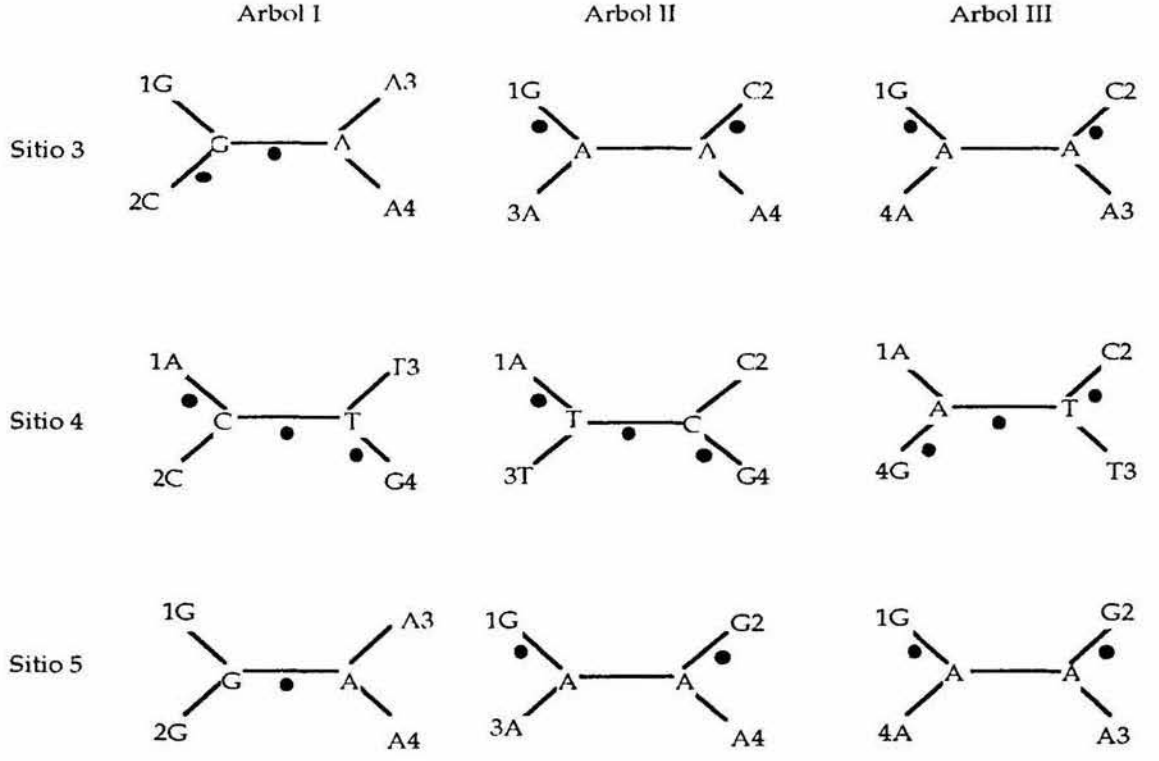


Figura 9. Ejemplo de los sitios informativos para el método de Parsimonia, los puntos representan los cambios en las secuencias (Tomado del Li pp 111-112).

b) Métodos de Matriz de Distancia

Estos métodos hacen una evaluación de la distancia genética entre cada par de secuencias, basada en el número de sustituciones que se observan entre ellas, produciendo una matriz con los datos resultantes. A partir de estos datos, se construye una representación gráfica, agrupando las secuencias en orden de mayor a menor similitud. Es posible obtener más de un árbol, por lo que es necesario que se elija aquel cuya topología refleje mejor las distancias obtenidas en la matriz de comparación de secuencias.

Los métodos más utilizados son el de UPGMA (método de promedios de distancias) y el de Fitch y Margoliash (1967), que fueron desarrolladas originalmente para construir fenogramas (diagramas que sólo expresan las relaciones fenotípicas en un grupo), pero pueden usarse para construir árboles filogenéticos, particularmente cuando las medidas de distancia del valor esperado son proporcionales al tiempo de divergencia entre las secuencias del estudio (Olsen, 1994; Li *et al,* 1990).

c) Método de Máxima Probabilidad

Este método utiliza tablas de frecuencias de sustitución para determinar las relaciones entre las secuencias en estudio. Estas tablas pueden ser predefinidas u obtenerse a partir de los datos (secuencias) que se introducen.

Para la construcción de los árboles filogenéticos es necesario saber que el par de OTUs que se va a unir primero es aquel que tenga la mayor probabilidad de estar más cercanas evolutivamente. Por otro lado, también es necesario saber que el largo de las ramas de las primeras dos secuencias se calcula sólo con respecto a una tercera, y de ésta con respecto a la cuarta y así sucesivamente (Fitch *et al,* 1967).

1.4.7. Consistencia de los métodos de reconstrucción filogenética

Todos los métodos mencionados son estadísticamente consistentes, en el sentido de que si hay una gran cantidad de datos aumenta la probabilidad de obtener un árbol que represente la historia evolutiva real. Sin embargo, cuando existen datos cuyo valor de divergencia es muy alto, o cuando la velocidad de evolución entre las OTUs es diferente, se pueden obtener árboles erróneos.

Para resolver este problema hay que llevar a cabo análisis estadísticos que permitan validar los resultados. Los programas como el *Bootstrap y Jacknife* no asumen ningún camino evolutivo previo y son completamente empíricos. Lo que hacen es modificar los datos, ya sea reduciendo el tamaño de alguna secuencia, eliminando secuencias enteras, etc, para reconstruir los árboles una y otra vez. Esto provee de un límite de confianza del 95% para cada una de las ramas en el árbol, aunque por ser análisis muy exhaustivos requieren de un gran tiempo de computo (Olsen 1994).

Como hemos visto, la metodología en evolución molecular ha permitido reconstruir de manera teórica algunas historias evolutivas, abriendo así diversas posibilidade para el mejor entendimiento de la evolución.

2. ANTECEDENTES

La mayoría de los trabajos de evolución molecular que se han realizado hasta el momento utilizan a las moléculas para determinar las relaciones filogenéticas de los organismos. Sin embargo, también se han desarrollado algunos trabajos donde se determina la historia evolutiva de las proteínas, tomando en cuenta la función que estas tienen dentro de los organismos. Este concepto de evolución relacionado con la función de las moléculas todavía no es muy familiar (Kleckner 1981).

Muchos trabajos nos indican que proteínas actuales estan compuestas de múltiples dominios funcionales y estructurales independientes los cuales provienen de distíntos orígenes.

Este modelo esta apoyado por el trabajo de E. Morett y L. Segovia (1992), en donde se hizo un estudio con las proteínas reguladoras que interaccionan con el factor σ^{54}. De los alineamientos que se realizaron con todo este grupo de proteínas, se identificaron tres dominios claramente separables; los cuales son; el dominio amino terminal, el central y el carboxilo terminal. En esta familia se observó que sólo el dominio central es homólogo en todas las proteínas y es el que define a la familia. En el dominio carboxilo terminal sólo se encuentran algunas regiones conservadas, mientras que el amino terminal, además de tener un tamaño variable, no contiene similitud obvia.

El análisis filogenético que se obtuvo con estas proteínas fué realizado con la secuencia de las proteínas completas y también con cada uno de los dominios independientes que antes se describieron. De esto se obtuvo que cada uno de los dominios habían evolucionado de manera independiente y que no ha habido mucho intercambio entre ellos, aunque algunos hayan coevolucionado.

En conclusión este trabajo ayuda a entender la evolución de los distintos dominios funcionales de las proteínas EBP's (del inglés, **E**nhancer **B**inding **P**roteins). El origen del sistema de σ^{54} es muy antiguo ya que este sistema ha sido encontrado en organismos tan alejados como firmicutes y proteobacterias.

Hasta el momento han sido descritas muchas proteínas que forman sistemas de dos componentes. Uno de los trabajos que se ha realizado con estas familias de proteínas es el de J.S. Parkinson y E.C. Kofoid (1992), en el cual se analizaron aproximadamente 53 proteínas HPK (proteínas histidin cinasas) y 91 proteínas RR (reguladoras de la respuesta).

Este sistema se encuentra principalmente en bacterias aunque ya han sido descubiertos sistemas homólogos en eucariotes. También se observó que las proteínas de las dos familias, tanto las sensoras como las reguladoras exhiben una gran variedad de arreglos con sus módulos. Los cuales pueden ser módulos solos, en tandem o ser la combinación de diferentes dominios. En ese análisis dendográfico se obtuvieron tres grupos principales, el primero al parecer esta relacionado por los módulos que se asocian a la utilización de nitrógeno, ya que muchas proteínas que estan involucradas con esta función forman un grupo tanto en las HPK como en las RR. El segundo grupo esta relacionado porque la disposición de sus módulos es muy similar, esta agrupación se puede hacer tanto en las proteínas HPK como en las RR. En el tercer grupo las proteínas RR no se encuentran relacionadas ni comparten alguna característica ni de función ni de arreglo en sus dominios, sin embargo, las proteínas HPK si se encuentran agrupadas.

Basados en que proteínas equivalentes en *E. coli* y en *B. subtilis*, organismos que divirgieron hace más de mil millones de años, contienen módulos muy similares Parkinson y Kofoid (1992) proponen que la comunicación entre los módulos se ha diseminado por transferencia horizontal relativamente reciente.

Por otro lado la gran variabilidad de módulos en una misma especie parecería indicar que la similitud de las secuencias entre las especies podría ser el resultado de evolución ya sea convergente o divergente. Por último, las secuencias más relacionadas no estan altamente correlacionadas con la especificidad en la comunicación, como se puede observar con las proteínas HPK ComP y DegS las cuales son muy similares en secuencia pero se comunican con receptores muy diferentes (ComA y DegU), por lo que es muy difícil determinar solo por comparación de las secuencias, cuales son los determinantes responsables de la comunicación específica.

La conservación de los dominios de las proteínas, en los pares de HPK y RR, sugieren que las interacciones entre cada par involucra un mecanismo consertado de evolución. Esto quiere decir que si las proteínas han estado relacionadas para responder a una función determinada, estas pudieron haber evolucionado juntas hasta formar un solo locus.

En este trabajo estamos interesados en reconstruir la historia evolutiva de las proteínas reguladoras del sistema de dos componentes, ya que este sistema es un buen modelo para determinar si existe evolución concertada entre proteínas, y además nos permite saber si esta evolución depende principalmente de la interacción que existe entre los módulos que conforman a estas proteínas. El hecho de que los dominios de comunicación entre estas proteínas se encuentren muy conservados nos lleva a considerar que su evolución no ha sido como proteínas independientes, sino como partes de un sistema coordinado que se encarga de regular de manera específica la respuesta adecuada a una señal recibida.

3. OBJETIVOS

3.1. OBJETIVO GENERAL

-El objetivo general de este proyecto es reconstruir la historia evolutiva de las dos familias de proteínas de los sistemas de dos componentes y evaluar si cada sistema de proteínas ha evolucionado de manera concertada.

3.2. Objetivos particulares

-Realizar una búsqueda en los bancos de secuencias de las proteínas relacionadas a las conocidas del sistema de dos componentes

-Realizar alineamientos de cada una de las familias de proteínas para comparar y obtener el grado de similitud que existe entre ellas.

-Realizar un análisis filogenético que nos sugiera el patrón evolutivo que han seguido estas dos familias de proteínas a lo largo del tiempo, utilizando distintos métodos como el de FITCH y el de máxima parsimonia

-Analizar la topología de los árboles obtenidos y determinar si estos son congruentes entre sí y con la de la otra familia, como para determinar si estas proteínas han mantenido una evolución concertada.

4. METODOLOGIA

4.1. Búsqueda

Para cumplir con los objetivos planteados en cuanto a la busqueda de todas las proteínas semejantes a las de los sistemas de dos componentes, fué necesario tener acceso a bancos de secuencias, como GenBank y EMBL de DNA y SwissProt de proteínas. Para esto se utilizaron programas que se encuentran en el paquete de GCG (Genetics Computer Group) versión 7.2 (25), de los cuales se hace una breve descripción a continuación.

FASTA: Este programa utiliza el método desarrollado por Pearson y Lipman (59), el cual busca secuencias en los bancos de datos similares a la secuencia en cuestión, sea DNA o proteína. En el primer paso de la búsqueda se compara a la secuencia en cuestión con las demás secuencias del banco y se identifican aquellas secuencias que tengan el mayor número de pequeños apareamientos en cada comparación. Este programa también utiliza la modificación de Wilbur y Lipman (77), por medio de la cual se seleccionan las 10 mejores regiones de similitud entre las secuencias. En el segundo paso, estas regiones son reevaluadas utilizando una matriz que en caso de ser proteínas se utiliza la matriz PAM250, la cual toma en cuenta reemplazos conservativos y símbolos ambigüos, entre otras cosas (Apéndice-2). Y por último en el tercer paso utilizando el procedimiento para alinear de Smith y Waterman se alinean con la secuencia problema, aquellas secuencias con el puntaje más alto de similitud.

TFASTA: Utiliza el método desarrollado por Pearson y Lipman (59), para buscar similitudes entre la secuencia del péptido en cuestión y cualquier grupo de secuencias nucleotídicas. El TFASTA traduce las secuencias nucleotídicas en los 6 marcos de lectura. Cada una de las 6 traducciones es tratada como una secuencia de aminoácidos diferente y se somete al procedimiento de comparación de la misma forma que el FASTA.

STRINGSEARCH: Si se utiliza el paquete de GCG, cada una de las secuencias de los bancos de datos tienen dos descripciones, la primera es la definición, la cual contiene el nombre del organismo, el nombre del gene, el largo de la secuencia y la fecha. La segunda es la documentación, en donde se presentan algunas características sobresalientes de la secuencia. Este programa busca tanto en la definición como en la documentación, aquella palabra que uno haya especificado para la búsqueda.

PROFILEMAKE: Utiliza el método de Gribskov para crear el perfil de un grupo de secuencias alineadas. Este perfil es una tabla que contiene la información cuantitativa (valores posición-específicos) obtenida de la comparación de un grupo de secuencias alineadas. Este perfil puede entonces ser usado por los programas de PROFILESEARCH, para la búsqueda o por PROFILEGAP para alineamientos de secuencias.

PROFILESEARCH: Este método usa el perfil creado por el PROFILEMAKE, que son un grupo de secuencias alineadas, como templado para buscar nuevas secuencias en el banco de datos. Utiliza el método desarrollado por Gribskov.

FETCH: Este programa copia las secuencias de la librería (bases de secuencias) a un archivo con formato GCG y lo muestra en la pantalla.

4.2. Alineamientos

Para obtener los alineamientos, que son el apareamiento de dos secuencias similares, con el fin de identificar el grado de conservación y los sitios de inserciones o deleciones en las proteínas, se utilizaron también programas del paquete GCG versión 7.2 de los cuales se hará una breve descripción:

LINEUP: Este es un programa que permite editar hasta 30 secuencias simultaneamente, acomodar y homogenizar su formato para que con el programa PILEUP estas puedan ser alineadas correctamente.

PILEUP: Este programa realiza alineamientos múltiples, los cuales pueden contener hasta con 300 secuencias de un tamaño máximo de 5000 caracteres. Utiliza una simplificación del alineamiento progresivo del método de Feng y Doolittle, que es parecido al descrito por Higgins y Sharp. Los alineamientos se realizan por pares de secuencias y son progresivos, las secuencias son agrupadas por similitud para producir un esquema que muestra las relaciones que existen entre las secuencias (dendograma).

PRETTY: Este programa muestra los alineamientos de secuencias múltiples y calcula la secuencia consenso. No crea el alineamiento, solo lo muestra.

4.3. Análisis filogenético

Para conocer la distancia genética que existe entre las distintas proteínas es necesario hacer pruebas estadísticas. El paquete Phylip (59), contiene programas que nos permiten realizar estos cálculos.

Muchas de las generalidades y diferencias de este tipo de métodos se muestran en la introducción, por lo que aquí se describen brevemente aquellos programas que se utilizaron.

PROTDIST: Con este programa se calculan las distancias de secuencias de proteínas, usando el estimdo de máxima probabilidad basado en la matriz de Dayhoff PAM250 (Apéndice-2) o en un modelo basado en el codigo genético. Este programa utiliza secuencias de proteínas para calcular la matriz de distancia bajo tres diferentes modelos de sustitución de aminoácidos. La matriz obtenida puede ser usada en los programas de matriz de distancia como el Fitch y el Neighbor. La distancia entre cada par de OTU's estima el largo total de las ramas entre los dos.

NEIGHBOR: Fué implementado por Mary Kuhner y John Yamato a partir del método de *Neighbor Joining*, original de Saitou y Nei que utiliza una matriz de distancia producto del programa PROTDIST. Este produce un árbol sin raíz y sin la suposición de un reloj molecular. El largo de las ramas no está optimizado, pero el método es muy rápido.

FITCH: Estima las filogenias de los datos obtenidos de una matriz de distancia (PROTDIST) bajo el modelo de árbol aditivo, UPGMA.

PROTPARS: Estima la filogenia de secuencias de proteínas utilizando el método de máxima parsimonia, donde solo los cambios de nucleótidos que provocan un cambio en aminoácido, son tomados en cuenta, ya que se asume que los cambios silenciosos se realizan con mayor frecuencia.

DRAWTREE: Dibuja los árboles filogenéticos de las secuencias proporcionadas ya sean de DNA o de proteínas.

4.4. Equipo

Las computadoras que se utilizaron fueron:

- MicroVAX, Mod. 3300 (de Digital Sistems), con un sistema operativo VMS versión 5.3 (Centro de Investigación sobre Fijación de Nitrógeno U.N.A.M.).

- PC 386 DX (Centro de Investigación sobre Fijación de Nitrógeno U.N.A.M.)

  -MsDOS versión 5

  -Windows versión 3.1

  -programa comunicación Telnet PC/TCP para DOS, versión

  2.05 p14.

  -Protocolo de comunicación FTP/IP (File Transfer Protocol)

- Silicon Graphics, Mod. 4D/35, con un sistema operativo irix 4.0.5 (Instituto de Biotecnología U.N.A.M.)

5. RESULTADOS Y DISCUSION

5.1. Búsqueda

Para cumplir con los objetivos planteados y obtener todas aquellas secuencias de proteínas pertenecientes a la familia de dos componentes, se realizó una búsqueda en los bancos de datos de GenBank, EMBL y SwissProt. Para esto fué necesario utilizar algunos programas como el FASTA, el TFASTA, el STRINGSEARCH y el PROFILISEARCH, los cuales se encuentran en el paquete GCG versión 7.2 (25) y fueron descritos en la metodología.

De las proteínas obtenidas se seleccionaron aquellas que guardaban un grado de identidad de por lo menos el 25% en cada 100 aminoácidos al compararlas con las proteínas más conocidas de los sistemas de dos componentes, como NtrB y EnvZ, las cuales son proteínas sensoras y NtrC y ArcA que son proteínas reguladoras. Además de esto, las proteínas seleccionadas tenían que presentar todas o la mayor parte de las regiones conservadas que han sido identificadas para la familia de proteínas de dos componentes. De esta manera se aseguraba la homología de las proteínas para realizar el análisis.

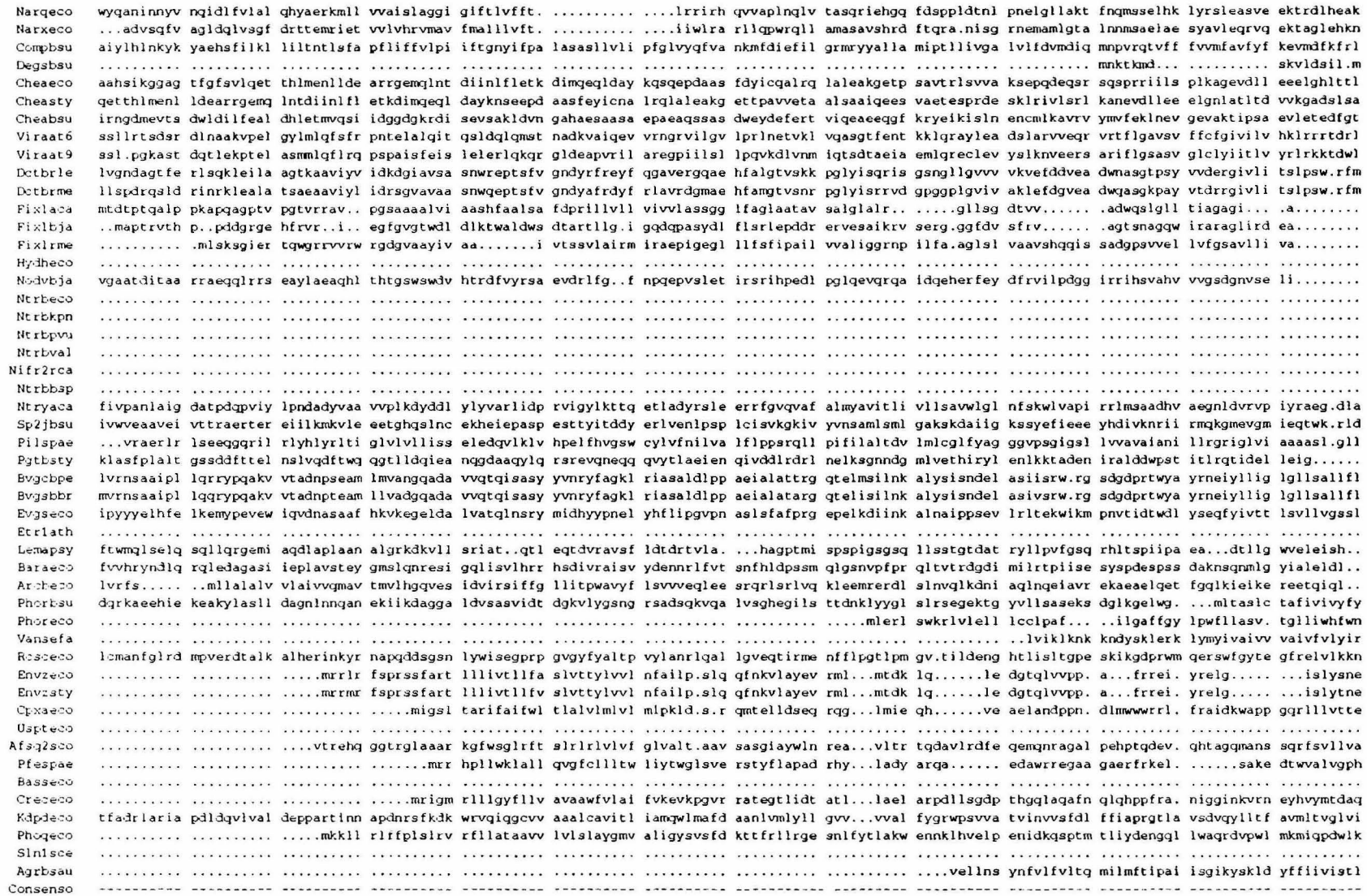
En total se obtuvieron 52 proteínas sensoras (HPK) y 83 proteínas reguladoras (RR). De estas proteínas se seleccionaron todos los pares que formaran sistemas de dos componentes, por lo que el análisis esta basado sólo en 49 proteínas HPK y 56 proteínas RR (Apéndice-1).

5.2. Alineamientos

Para identificar las regiones más conservadas entre los miembros de estas familias, así como los dominios ya sea funcionales o estructurales de los que estan compuestas las proteínas, se realizaron alineamientos múltiples mediante el programa PILEUP, los cuales fueron optimizados manualmente. Para identificar los aminoácidos presentes en más del 75% de las proteínas se utilizó el programa PRETTY, este programa nos presenta claramente las regiones consenso para cada una de las familias de proteínas. Inicialmente, se realizaron los alineamientos de las secuencias completas con las proteínas HPK y RR seleccionadas, las cuales formaron 33 sistemas de dos componentes (Figs. 10 y 11).

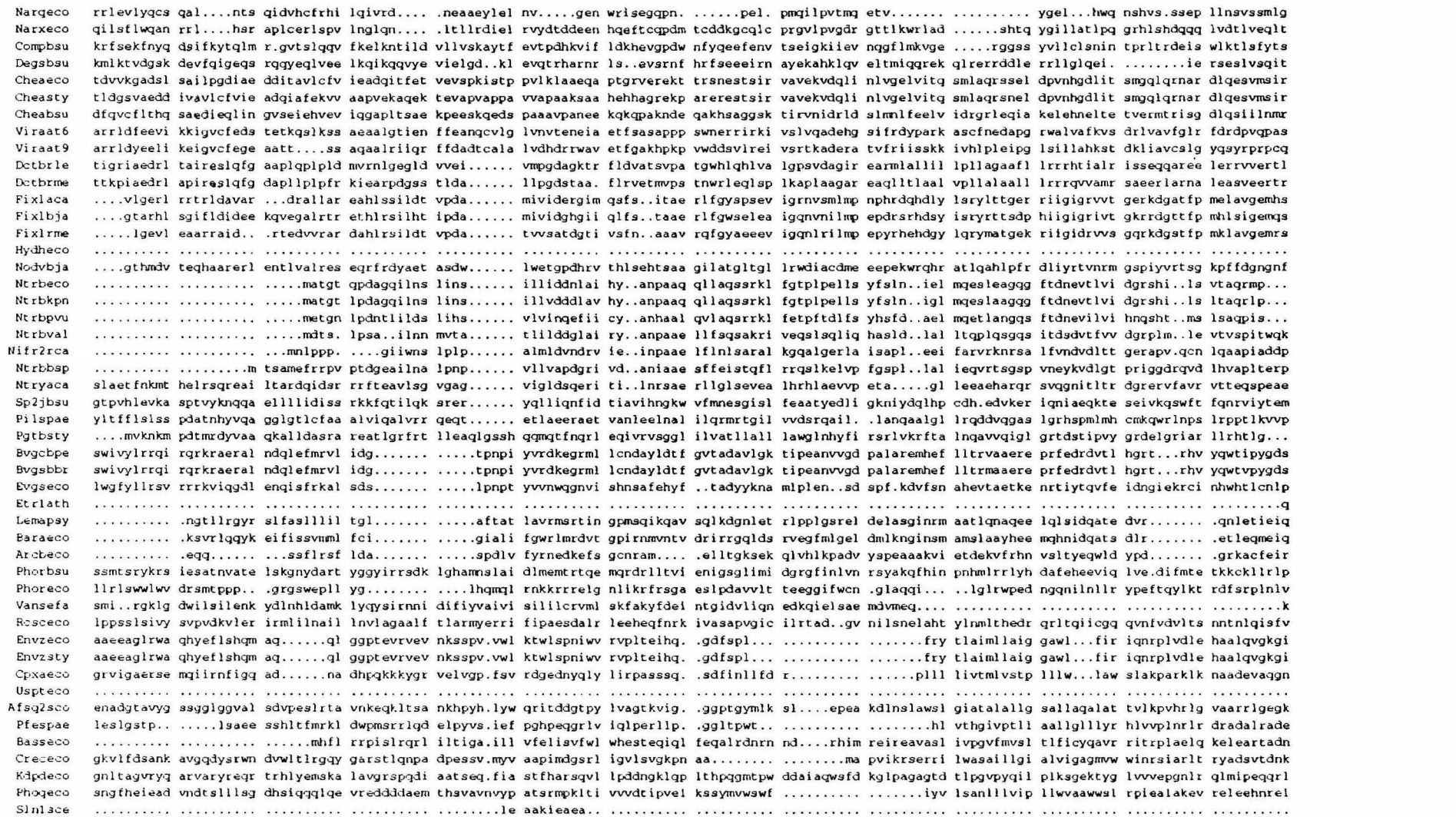
Del alineamiento con las proteínas HPK completas se pudieron identificar claramente los límites del módulo de transmisión de la señal, el cual está localizado en la región carboxilo terminal de todas las proteínas HPK. En la región amino terminal de estas proteínas, se identificaron por lo menos dos regiones hidrofóbicas para cada una de ellas.

Figura 10 (a). Alineamientos de las proteínas HPK completas realizados por el programa PILEUP, en donde se muestran los dominios de los que estan compuestas estas proteínas. La caja negra nos indica el módulo de transmisión conservado de estas proteínas. Las rayas paralelas encontradas en la secuencia de la proteína Sln1 representa un "indel" que no fué considerado para el análisis.



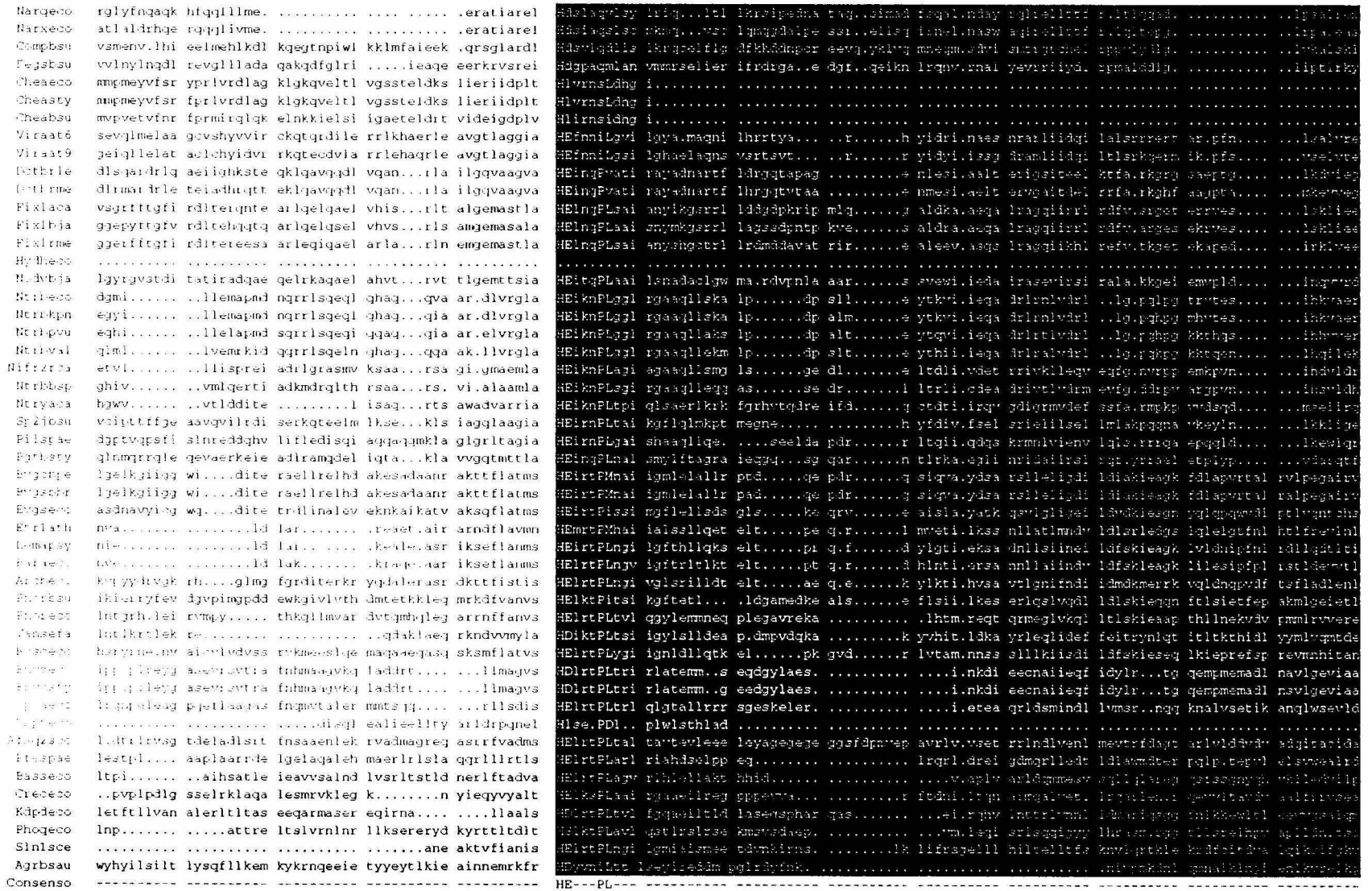
Región N-terminal (donde se encuentran las regiones hidrofóbicas) **Dominio de entrada**

Figura 10 (b)



Región N-terminal (Regiones hidrofóbicas) **Dominio de entrada**

Figura 10 (c)



Región C-terminal **Módulo de transmisión**

Figura 10 (d)

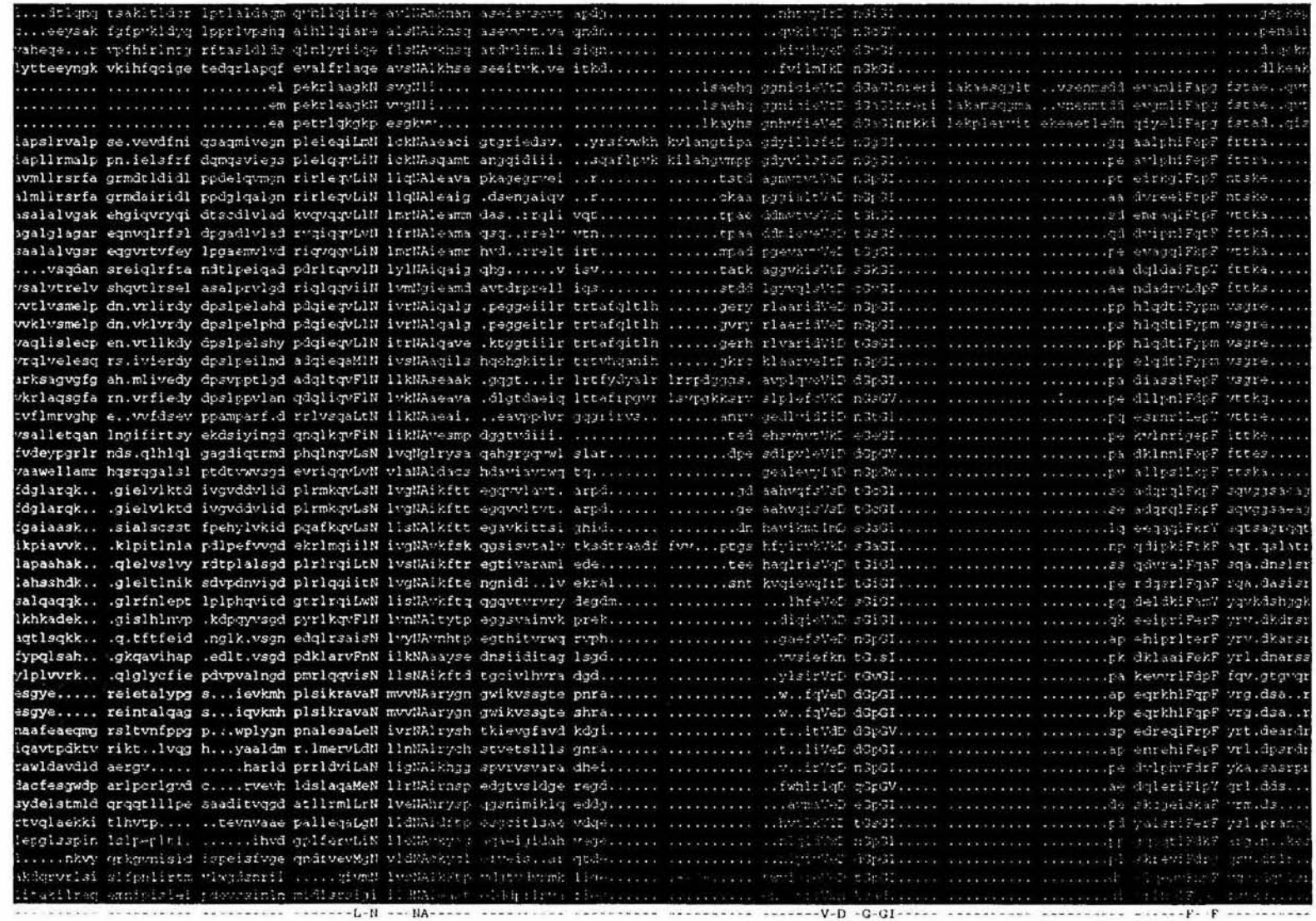


Figura 10 (e)

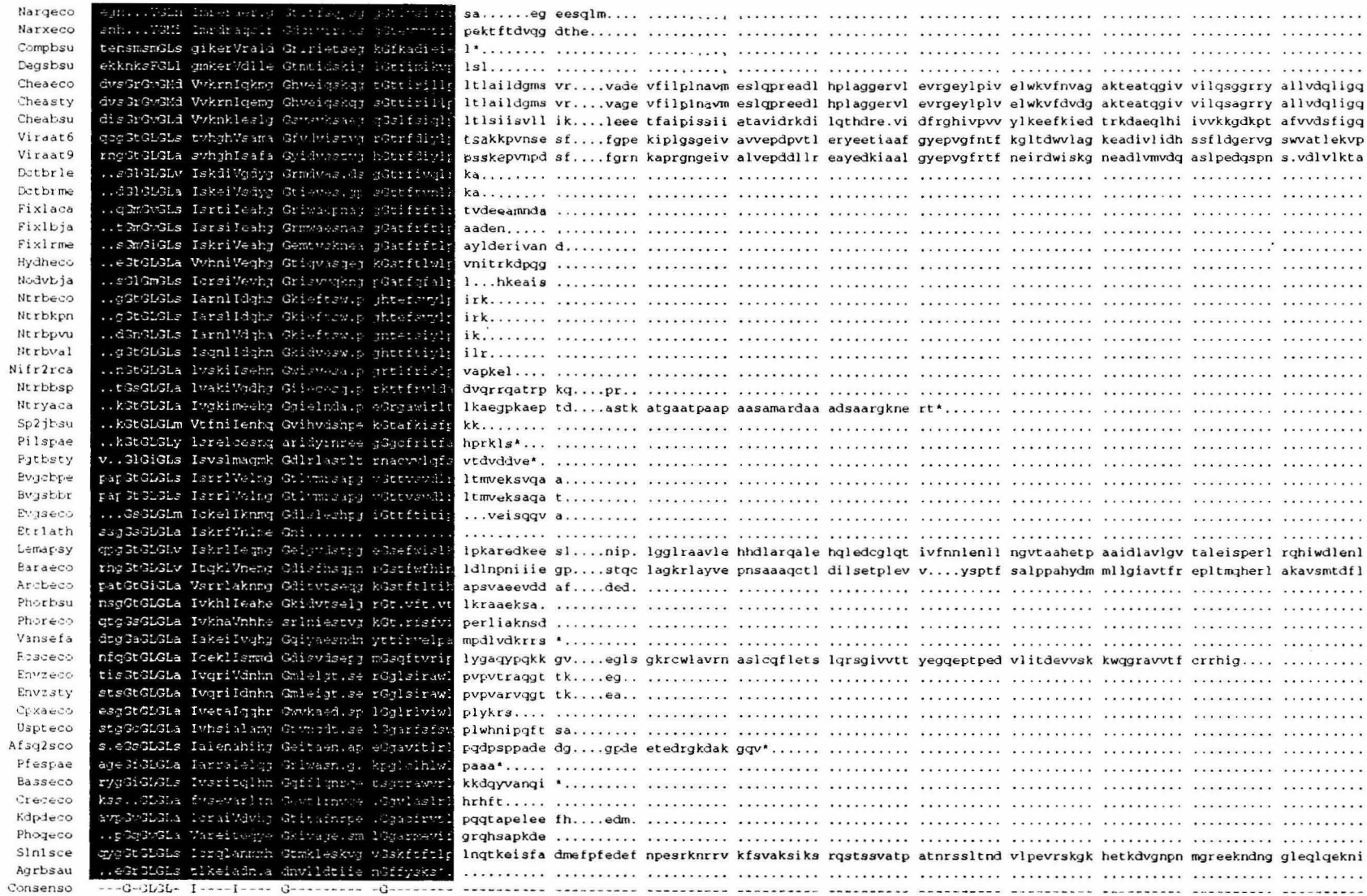
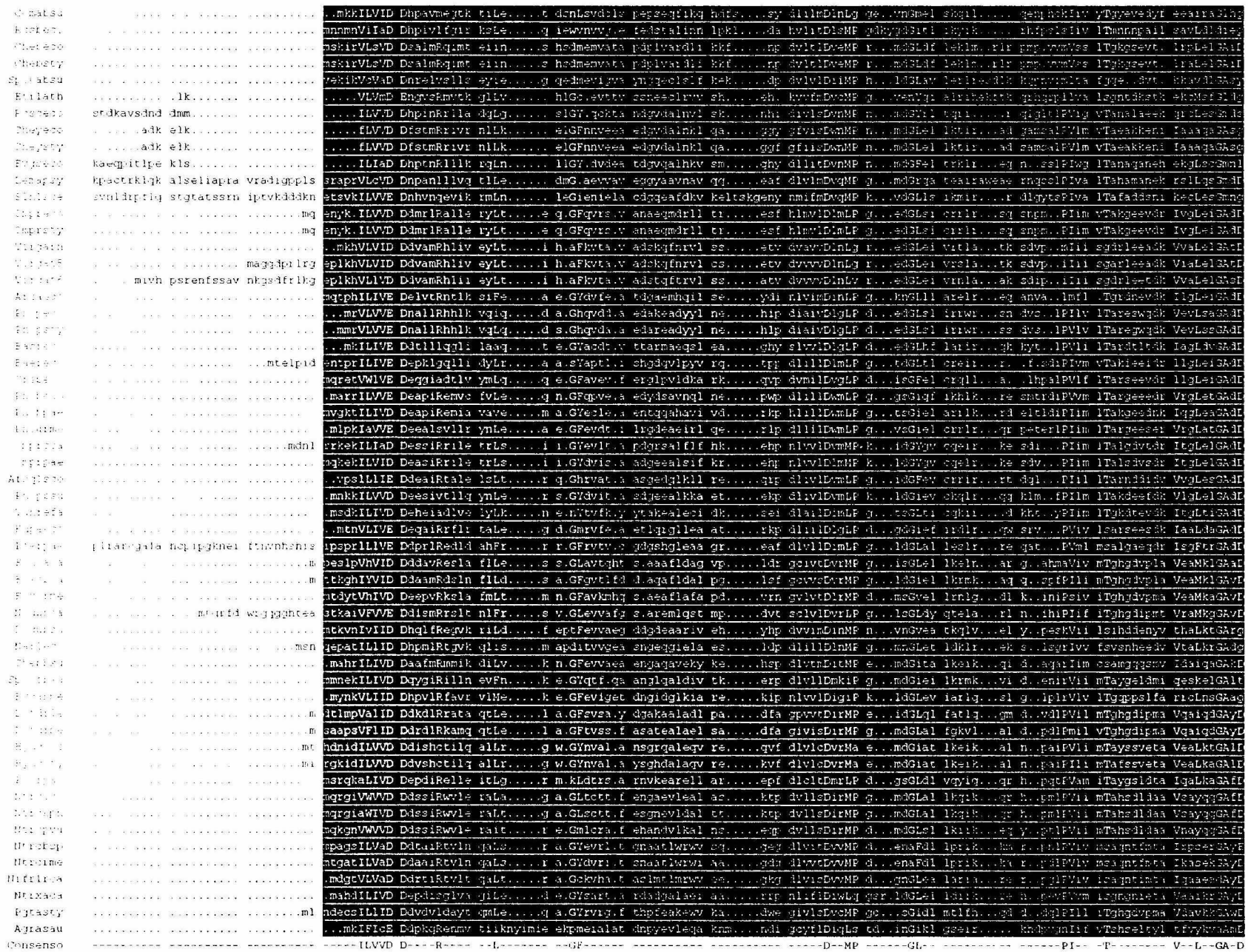
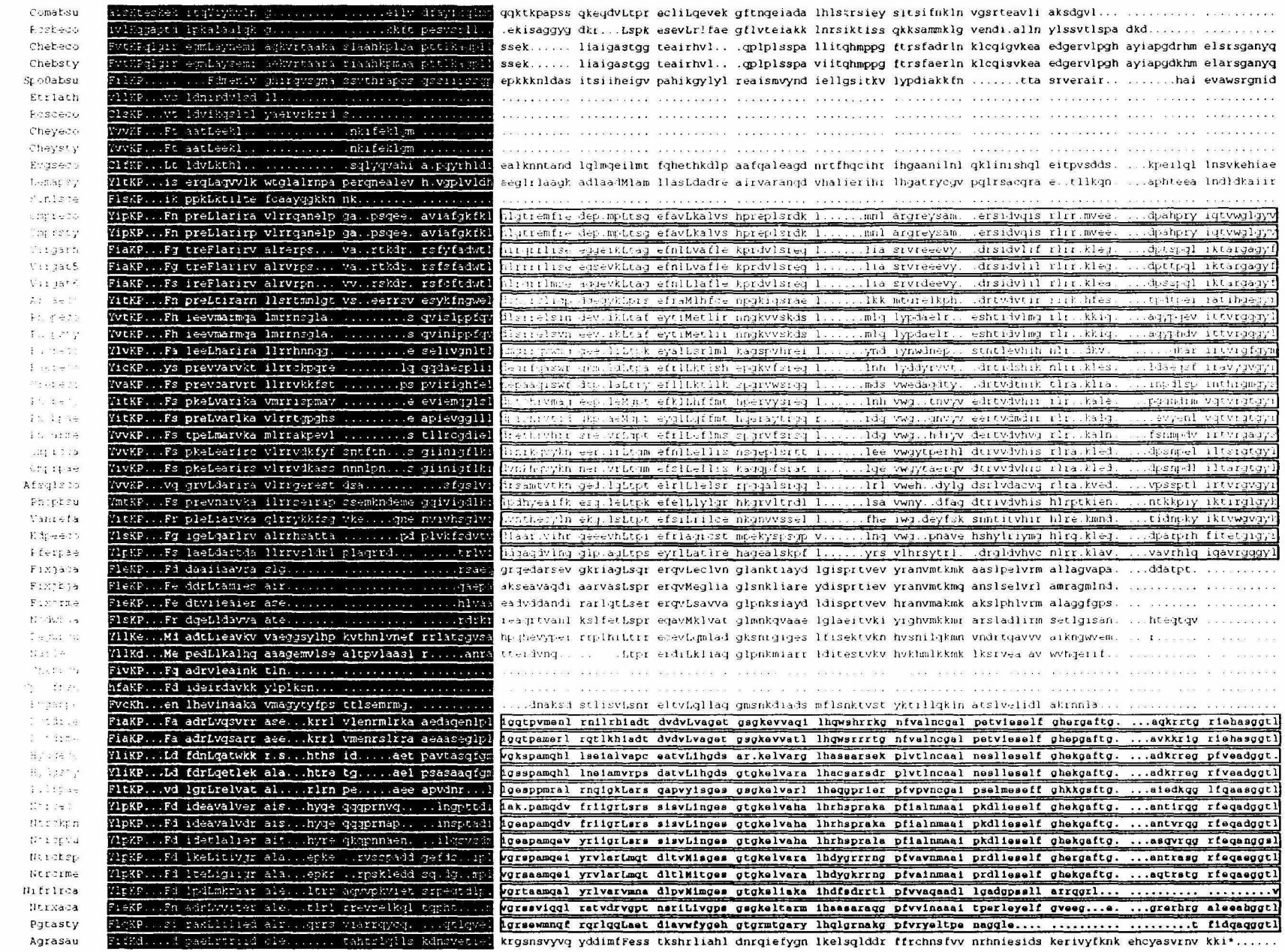


Figura II (a). Alineamientos de las proteínas RR completas realizados por el programa PILEUP, en donde se muestran los dominios de los que estan compuestas estas proteínas. La caja negra indica el módulo receptor que es el más conservado. La caja transparente nos muestra el dominio de las proteínas activadoras de la familia de σ^{54}, por último la caja transparente con las letras resaltadas es un dominio identificado como de unión a DNA.



Región N-terminal **Módulo receptor**

Figura 11 (b)

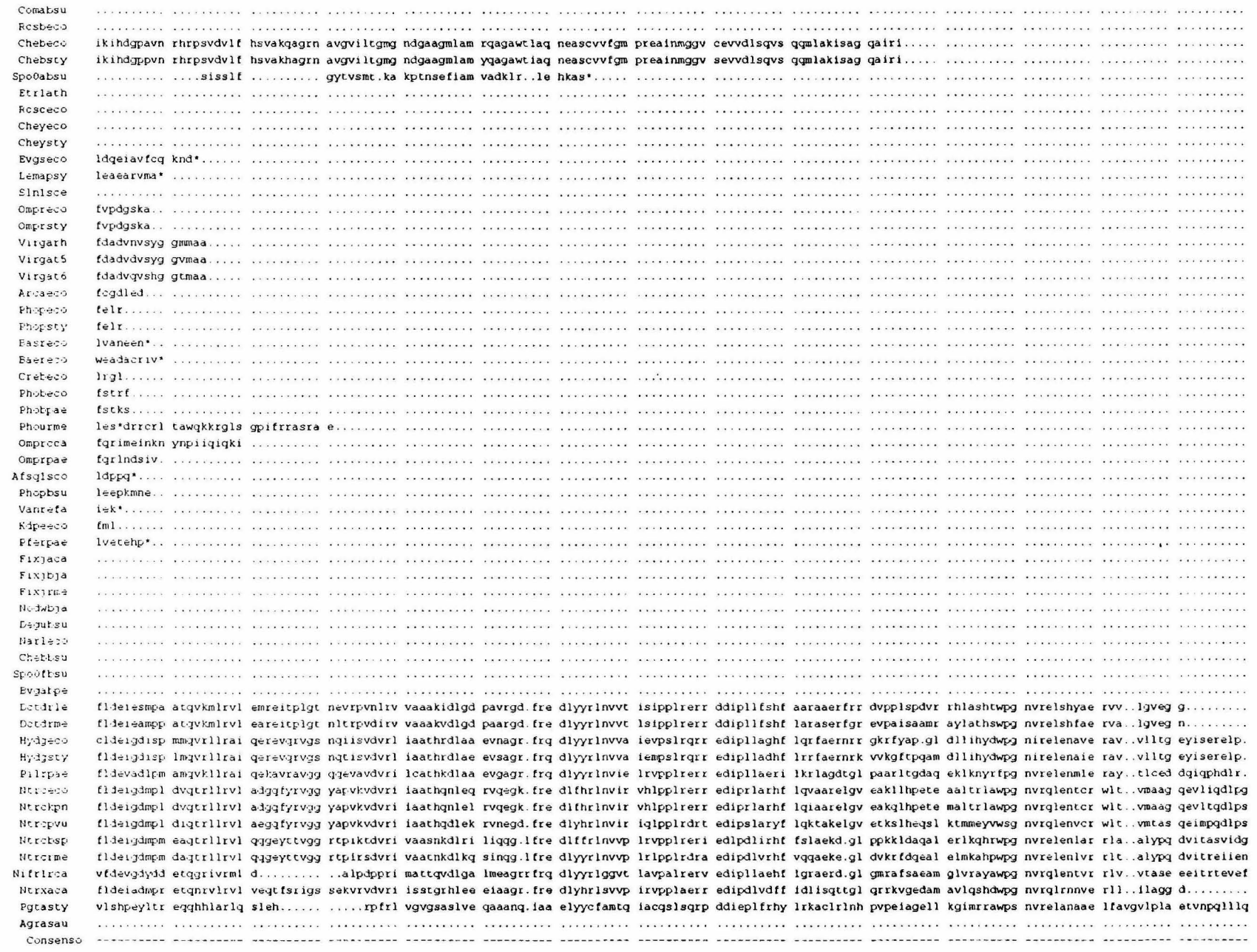


Módulo receptor

Región central para algunas y C-terminal para otras

Dominio de salida

Figura 11 (c)



Región C-terminal **Dominio de salida**

Estas regiones no presentan similitud en la misma proteína, ni entre las regiones de las demás proteínas. Se ha determinado que estas regiones permiten que las proteínas se anclen a la membrana para favorecer la recepción de las señales ambientales (63). Algunas de las proteínas de esta familia, como NtrB, DegS, Sp2J y CheA, reciben las señales en el citoplasma y se comprobó que carecen de regiones hidrofóbicas (Figura 12-A).

Por otro lado, en el alineamiento realizado con las proteínas RR completas, se observó que todas las proteínas contienen el módulo receptor de la región amino terminal conservado. Algunas de estas proteínas contienen otros dominios, los cuales permiten formar 4 subgrupos principales (Figura 12-B), los cuales se describen a continuación:

El primer subgrupo contiene a las proteínas NtrC, DctD, HydG, PilR, NtrX y PgtA, las cuales constan de 3 dominios; el dominio amino terminal o módulo receptor; el dominio central y el carboxilo terminal. El dominio central de estas proteínas es una región muy conservada, la cual pertenece a la familia de proteínas que se unen a regiones de DNA en procariontes parecidas a los *enhancers* eucariontes. Al unirse al DNA activan genes que se transcriben por la familia de factores σ^{54}. Por último, la región carboxilo terminal en este grupo de proteínas no está conservada.

El segundo subgrupo contiene a las proteínas, OmpR, VirG, ArcA, PhoP, PhoB, BasR, BaeR, CreB, AfsQ1, VanR, KdpE, y PfeR. Estas proteínas están formadas por dos dominios muy conservados, el primero es el módulo receptor, localizado en la región amino terminal, y el segundo en la región carboxilo terminal. Este último agrupa a las proteínas en una familia de activadores que se unen a DNA. Mediante este dominio las proteínas son capaces de reconocer regiones específicas del promotor favoreciendo la regulación de los genes involucrados. Para determinar si existen más proteínas con este dominio, se realizó una búsqueda en los bancos de datos mediante el programa FASTA utilizando el dominio carboxilo terminal como referencia y no se encontraron más proteínas homólogas a esta familia.

En el tercer subgrupo se encuentran las proteínas FixJ, NodW, DegU, NarL, ComA, RcsB, CheB, Spo0A, AgrA, BvgA, EvgS y LemA, que contienen dos dominios, el módulo receptor muy conservado y un dominio de salida, no conservado.

El último subgrupo contiene a las proteínas Spo0F, CheY y a la proteína CheB de *B. subtilis*, que estan compuestas solo por el módulo receptor, y carecen de dominio de salida.

Por otro lado las proteínas Etr1, Sln1, LemA, EvgS, ArcB, BarA y RcsC, son un solo polipéptido, formado por los dos dominios de comunicación, el de las HPK y el de las RR (Figura 12-C). El arreglo que presentan los dominios en las proteínas nos muestra la gran variedad de combinatorias entre estos.

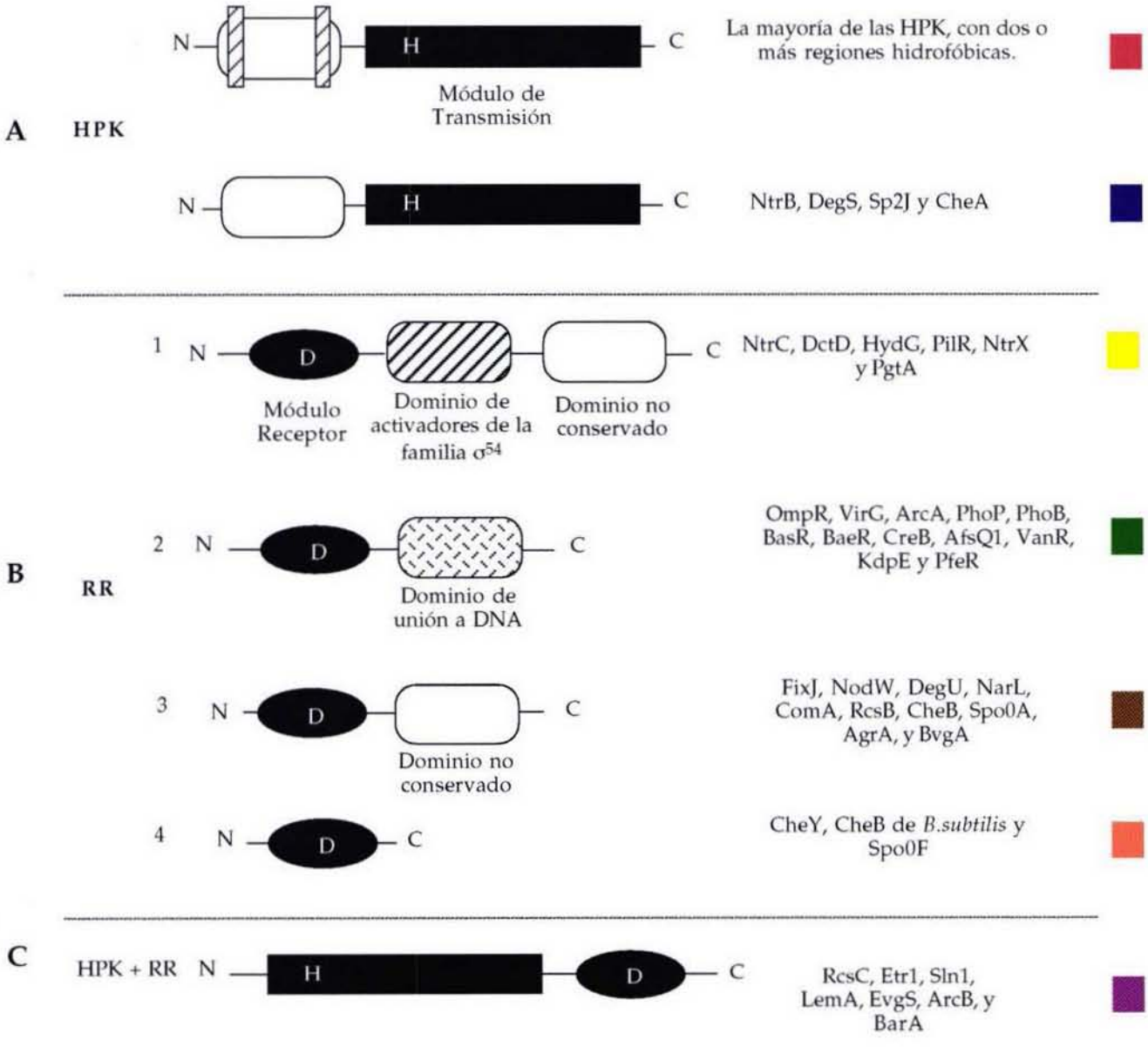


Figura 12. Dominios de las proteínas HPK (**A**) y RR (**B**), determinados mediante los alineamientos con las proteínas completas. El tercer agrupamiento incluye a las proteínas que contienen los dos dominios conservados, tanto de las HPK como el de las RR (**C**). Los colores ilustran el grupo estructural que estan formando

Como se mencionó con anterioridad, el principal objetivo de este proyecto es analizar la historia evolutiva en los dominios de comunicación de la familia de proteínas de dos componentes; ya que es interesante saber si la interacción de los módulos entre las proteínas ha influido en su evolución.

Se realizaron alineamientos únicamente con los módulos de comunicación ya que estos son los módulos de interés además de que las regiones no conservadas podrían desviar los resultados. En estos alineamientos se observó claramente en las dos familias de proteínas que los aminoácidos importantes para la comunicación, histidina (H) para las proteínas HPK y ácido aspártico (D) o glutámico (E) para las proteínas RR se mantienen muy conservados. En general, los residuos de aminoácidos que forman los bloques más conocidos de las dos familias de proteínas (HPK y RR) se encuentran conservados en más del 70% de estas.

5.3. Análisis filogenético

Los alineamientos realizados con los módulos de comunicación se utilizaron para realizar la reconstrucción filogenética de ambas familias de proteínas, mediante dos programas que utilizan estrategias de reconstrucción filogenética distintas: uno de los programas es el de matriz de distancia (FITCH*) (Figs. 13 y 14) y el otro es el de máxima parsimonia (PROTPARS) (Figs. 15 y 16). En los árboles obtenidos con cada uno de los programas, se mantiene de manera general, la misma relación evolutiva entre las proteínas, tanto en las HPK como en las RR.

Como ya se ha mencionado, los árboles filogenéticos son los que mejor representan las relaciones evolutivas de los OTU´s. Cabe mencionar que aunque la mayoría de las proteínas de este estudio se encuentran más relacionadas con proteínas de su grupo, algunas de ellas mantienen una menor distancia con otras proteínas. Esto se observó utilizando las distancias entre las proteínas dadas por el programa PROTDIST (Apéndice-4), lo que permitió mejorar la interpretación de las relaciones evolutivas entre estas.

De los árboles realizados con el programa FITCH se determinaron aquellas proteínas más alejadas evolutivamente o con una mayor divergencia para poder enraizar los árboles y determinar las relaciones filogenéticas de las dos familias de proteínas. La proteína que permitió enraizar el árbol de las HPK, fué CheA de *B. subtilis* y para enraizar el árbol de las RR se utilizó la proteína ComA de *B. subtilis* (Figs. 13 y 14). Estas raices no son verdaderas, debido a que no conocemos el ancestro en común de cada una de estas familias de proteínas, sin embargo, nos permiten agrupar y observar las relaciones filogenéticas que existen entre ellas.

*Para obtener la matriz de las distancias entre las proteínas, se utilizó el programa PROTDIST. A partir de la matriz obtenida se puede reconstruir el árbol filogenético con el programa FITCH.

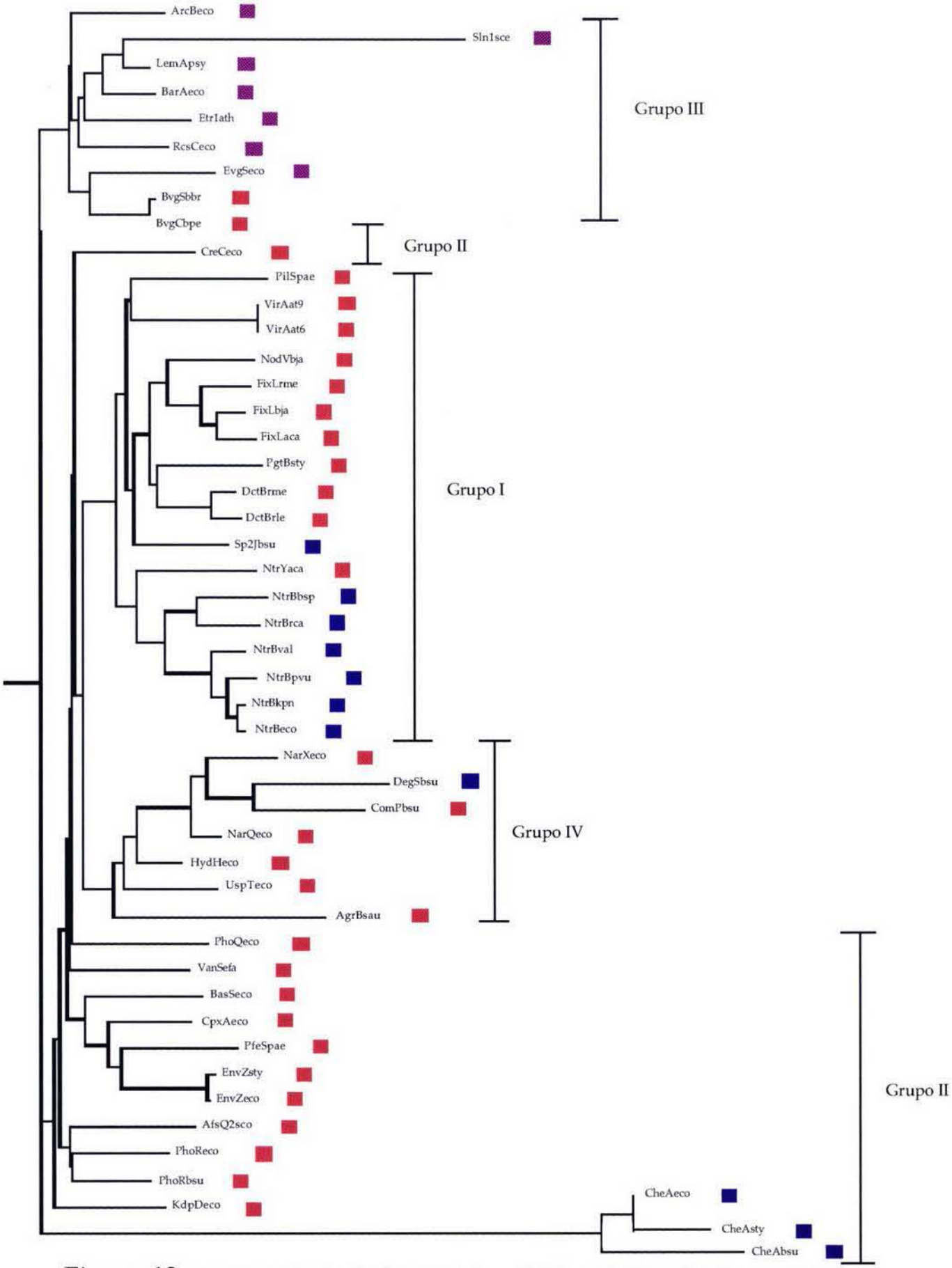


Figura 13. Agrupamiento de las proteínas HPK. Arbol realizado con el programa FITCH. Los colores son equivalentes a los de la Fig-12.

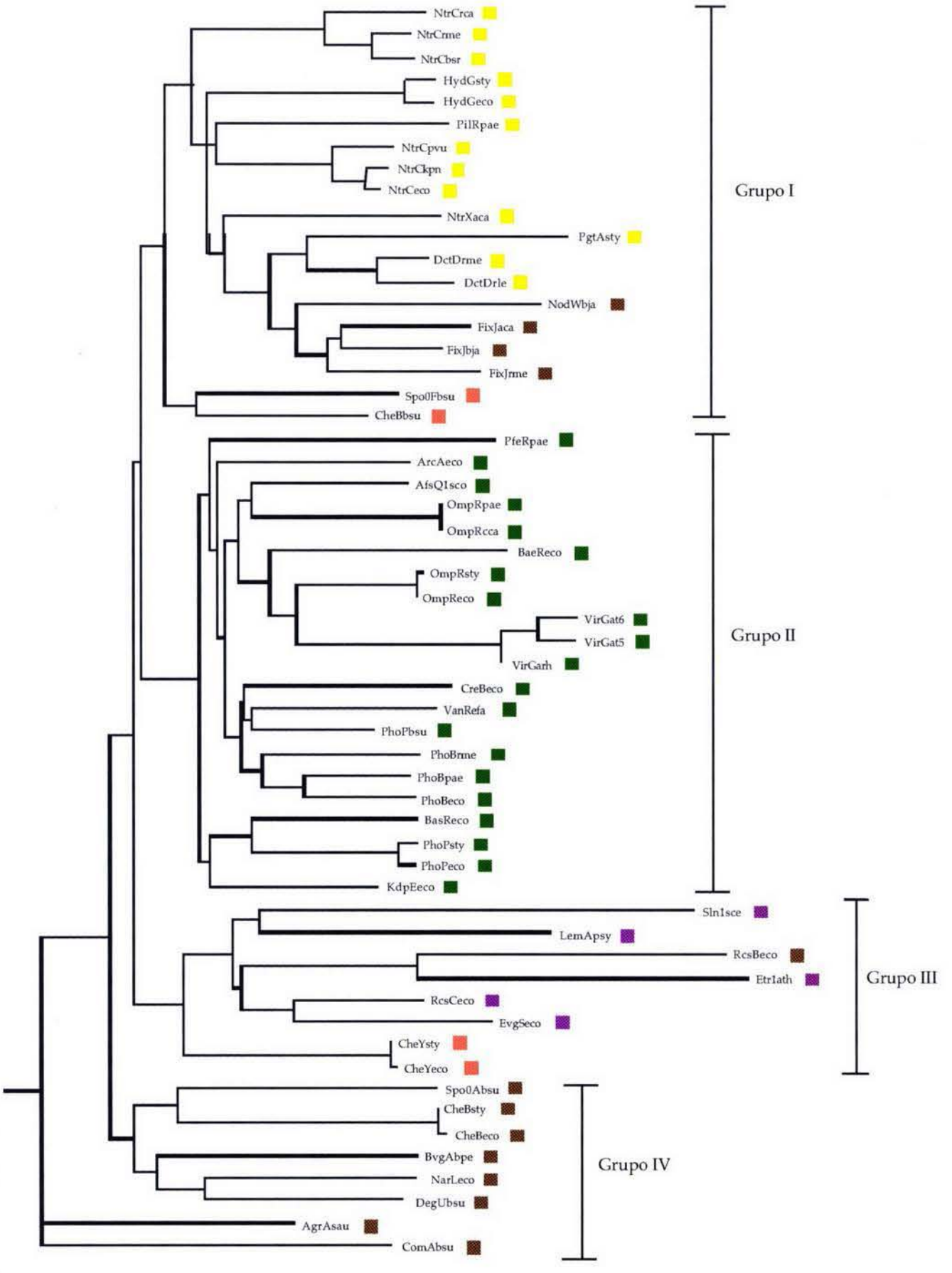


Figura 14. Agrupamiento de las proteínas RR. Arbol realizado con el programa FITCH. Los colores son equivalentes a los de la Fig-12.

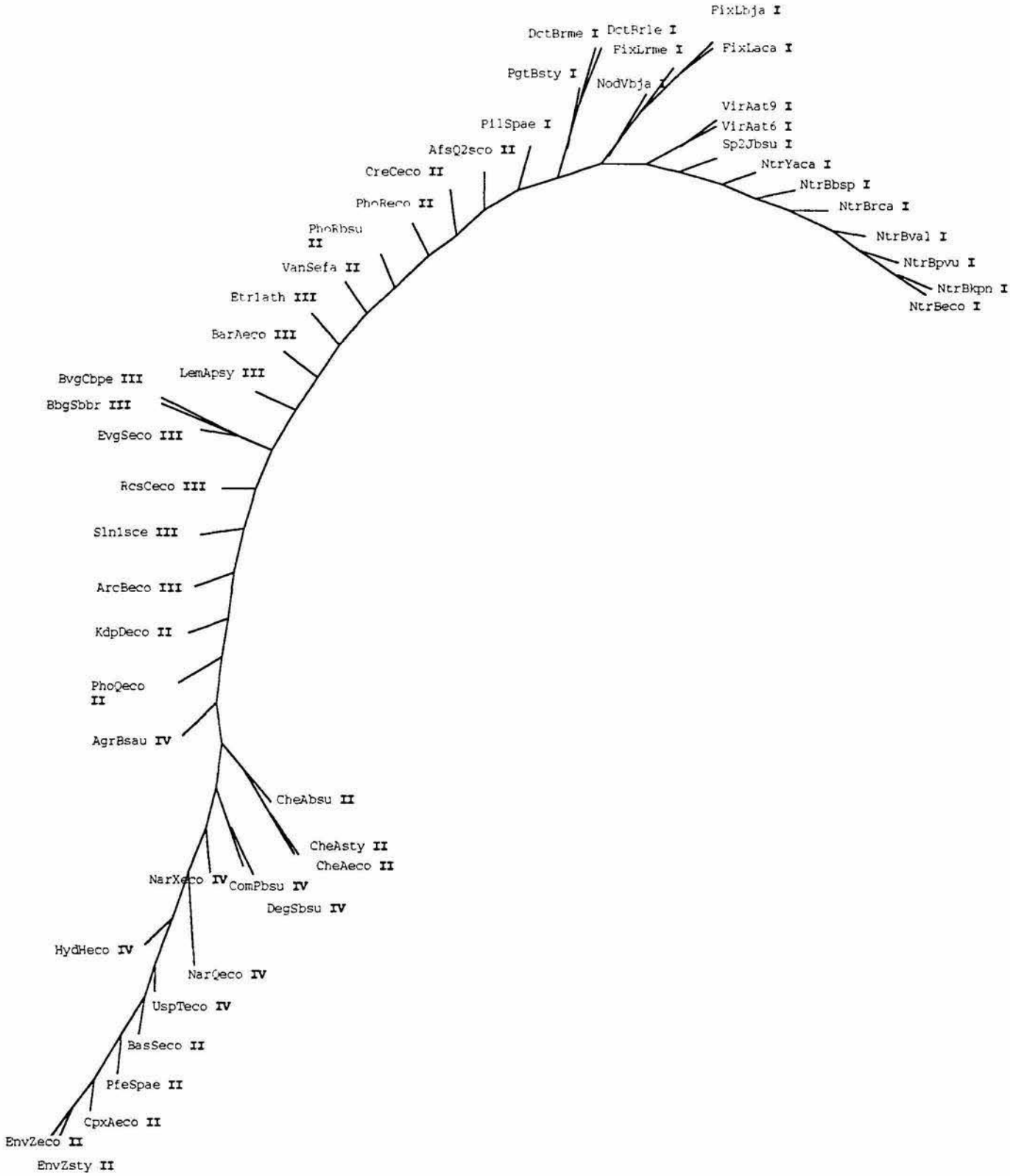


Figura 15. Proteínas HPK. Arbol realizado mediante el programa PROTPARS, los números romanos indican el grupo al que pertenecen las proteínas en el árbol realizado con el programa de FITCH

Figura 16. Proteínas RR. Arbol realizado mediante el programa PROTPARS, los números romanos indican el grupo al que pertenecen las proteínas en el árbol realizado con el programa de FITCH

Los árboles que se analizaron con más detalle fueron los que se obtuvieron por el programa de FITCH, y los que se realizaron mediante el programa PROTPARS ayudaron a corroborar si mediante los dos métodos las proteínas mantienen la misma relación.

En el árbol de FITCH de las proteínas HPK y utilizando las distancias de sustitución que se muestran en el Apéndice-4, se pudo identificar un grupo bien definido (Fig. 13-I), el cual contiene a las proteínas PilS de *P. aeruginosa*; VirA de *A. tumefaciens*-6, *A. tumefaciens*-9; NodV de *Bradyrhizobium japonicum*; FixL de *R. meliloti*, *B. japonicum* y *Azorhizobium caulinodans*; PgtB de *A. caulinodans*; DctB de *R. leguminosarum* y *R. meliloti*; Sp2J de *S. typhimurium*; NtrY de *A. caulinodans*; NtrB de *Rhodobacter capsulatus*, *Vibrio alginolyticus*, *Proteus vulgaris*, *K. pneumoniae* y *E. coli*. El hecho de que todos los árboles mantengan agrupadas a estas proteínas y que incluso conserven la misma relación entre ellas, nos habla de que este agrupamiento es verdadero. Una de las principales características de este grupo es que contiene a todas las proteínas involucradas en el proceso simbiótico de fijación de nitrógeno, como las FiXL, NodV, NtrB, NtrY y DctB.

De este grupo, las proteínas NtrB se encuentran más cercanas entre sí que con cualquier otra proteína. En los árboles realizados tanto con el programa FITCH (Fig-13) como con el de PROTPARS (Fig-15), estas proteínas siempre estan juntas.

Por otro lado, el agrupamiento de estas proteínas es congruente con los árboles realizados con 16S ribosomales (Apéndice-3), ya que las proteínas NtrB de proteobacterias de la subdivisión α se encuentran separadas de las proteobacterias de la subdivisión γ, siendo esto congruente con la separación evolutiva de los organismos. Las proteínas NtrB son ortólogas, es decir, que se originaron antes de la especiación, ya que se mantiene la función en organismos diferentes.

La proteína NtrY de *A. caulinodans*, se encuentra más cercana a las NtrB que a cualquier otra proteína. En todos los árboles realizados siempre se agrupa con NtrB, por lo que podemos decir que estas proteínas comparten un ancestro, siendo esto altamente probable ya que incluso las dos participan en la regulación de la asimilación de nitrógeno.

También dentro de este grupo están las proteínas FixL, la relación que guardan estas proteínas es congruente con los árboles obtenidos por 16S ribosomales (Apéndice-3) estas proteínas se han identificado en tres organismos pertenecientes a la subdivisión α de las proteobacterias, estos son *B. japonicum*, *A. caulinodans*, *R. meliloti* y guardan la misma relación en los árboles. Estas proteínas son ortólogas, ya que las encontramos en distintos organismos realizando la misma función.

Se observó que FixL y NodV siempre aparecen relacionadas. La proteína NodV de *B. japonicum* se encuentra más cercana evolutivamente de las proteínas FixL que de cualquier otra. Se puede decir que incluso NodV de *B. japonicum* puede ser una proteína paráloga a FixL de *B. japonicum*, ya que las dos se encuentran en un solo organismo y realizan funciones diferentes. La explicación a esto es que antes de la especiación la proteína ancestral se duplicó dando origen a NodV y a FixL, posteriormente las especies divirgieron conservando las dos proteínas con funciones diferentes. No sabemos si las proteínas NodV de *A. caulinodans* y *R. meliloti* no han sido encontradas o secuenciadas en estos organismos, o estos sufrieron una deleción de NodV en algún momento de su historia.

Las proteínas DctB y PgtB se encuentran agrupadas en todos los árboles. Las proteínas DctB en *R. meliloti* y *R. leguminosarum*, estan involucradas en el transporte de C4-dicarboxílicos y la proteína PgtB de *S. typhimurium* regula el transporte de fosfoglicerato. Es muy interesante que se agrupen proteínas que comparten una función relacionada, en este caso el transporte. Estas proteínas comparten un ancestro, el cual quizá comenzó siendo un transportador más laxo y a partir de este, mediante mutaciones y la fijación de estas, se fueron especializando hasta llegar a las proteínas que ahora conocemos, aunque también una de las protéinas pudo dar origen a la otra.

Las distancias que existen entre Sp2J y las proteínas de el grupo I de las HPK, son menores que las de Sp2J con otras proteínas (Apéndice-4). Cabe mencionar que en los árboles realizados por el programa FITCH, esta proteína aparece más relacionada con las proteínas DctB, PgtB, FixL y NodV que con las demás, en tanto que en los árboles realizados por el programa PROTPARS, Sp2J tiene una relación más cercana con VirA y NtrB que con las demás del grupo. Sin embargo, a pesar de esta diferencia esta proteína pertenece a este primer grupo independientemente de con cual proteína se relacione más.

Las proteínas VirA de *A. tumefaciens*-6 y 9 se encuentran muy relacionadas con las proteínas de el grupo I de las HPK. Es muy probable que la relación se deba a que estas proteínas están involucradas en la inducción de tumores en plantas, lo cual puede ser un proceso parecido al de simbiosis para la fijación de nitrógeno, siendo ambos eventos una interacción entre bacterias y plantas. Aunque hay que tomar en cuenta que las distancias de las proteínas VirA con PgtB es mayor que con otras proteínas que no son del grupo (Apéndice-4), como CreC, VanS, BvgC, BvgS, BarA y PhoR. Una de las explicaciones de que la proteína PgtB guarde una evidente relación con las proteínas de este grupo y que sin embargo este muy alejada evolutivamente, es que pudo haber divergido con mayor rapidez que las demás.

Por último, la proteína PilS de *P. aeruginosa*, la cual tiene que ver con la presencia de fimbrias en el organismo, mantiene valores de distancia menores con este grupo de proteínas que con cualquier otra proteína y en todos los árboles observados forma parte de este grupo, aunque sea la proteína más alejada del mismo (Apéndice-4).

Debido al arreglo que tienen las proteínas del grupo I, quizá indique muy probablemente, que el proceso simbiótico de fijación de nitrógeno como tal, se originó una sola vez y que a partir de las proteínas ancestrales el sistema se ha ido especializando y divirgiendo, de manera que ahora es un proceso muy complejo que requiere ser regulado por distintas señales para mantener un buen funcionamiento. Cuando los organismos entran en simbiosis necesitan coordinar todos aquellos genes que participan en el evento, es por esto que no es de extrañarse que estas proteínas se mantengan evolutivamente muy relacionadas.

En el segundo grupo (Fig. 13-II), las proteínas EnvZ de *E. coli* y *S. typhimurium*; PfeS de *P. aeruginosa* y CpxA de *E. coli* se encuentran claramente relacionadas en todos los árboles realizados. Al parecer, estas tres proteínas comparten un ancestro común, siendo la proteína PfeS la que más ha divergido de las tres. En cuanto a función estas proteínas no se encuentran relacionadas. Cabe destacar que la proteína CpxA transduce la señal a la proteína ArcA de *E. coli*, al igual que la proteína ArcB. Sin embargo estas dos proteínas HPK no son muy parecidas entre sí. En general, las demás proteínas de este grupo se encuentran muy relacionadas entre ellas. Las distancias sin embargo, no son tan diferentes entre estas proteínas y las del segundo grupo (Apéndice-4). Inclusive la proteína CreC se encuentra tan relacionada con proteínas del segundo grupo como con proteínas del cuarto grupo.

Finalmente se encuentran las proteínas CheA, las cuales son las proteínas más alejadas. Estas son proteínas ortólogas y su arreglo es congruente con los datos obtenidos por 16S (Apéndice-3). La proteína CheA transduce la señal a dos proteínas RR diferentes en secuencia, que son CheB y CheY.

El tercer agrupamiento de las HPK, contiene a las proteínas ArcB de *E. coli*, Sln1 de *Sacaromyces cereviciae*, LemA de *Pseudomona syringae*, BarA de *E.coli*, Etr1 de *Arabidopsis thaliana*, RcsC de *E. coli*, EvgS de *E. coli*, BvgS de *Bordetella bronchiseptica* y BvgC de *B. pertusis* (Fig. 13-III). Este agrupamiento se mantiene en cualquiera de los árboles realizados, sin embargo, al corroborar esto con las distancias (Apéndice-4), se observó que no siempre estas son menores entre las proteínas de este grupo que con las demás, aunque generalmente ocurre así. Una de las proteínas más llamativas en este grupo es Sln1 de *S. cereviciae*, esta proteína aunque está muy alejada, todavía conserva las regiones que le permiten relacionarse con esta familia y en particular con este grupo, ya que en todos los árboles se encuentra agrupada de la misma forma, inclusive la distancia evolutiva es menor con estas que con las demás (Apéndice-4).

Por otro lado, la mayoría de las proteínas de este grupo estan más cercanas a otras que a Sln1, esto nos habla de que muy probablemente esta proteína ha divergido más rápidamente que las demás de su grupo aunque sigan compartiendo un ancestro. Con la proteína EvgS de *E. coli* pasa algo similar ya que LemA de *P. syringae*, está más cercana a PhoR que a EvgS que sí es de su grupo (Figura 13-III).

Es importante resaltar que en el árbol realizado con esta familia de proteínas, se observa una clara división entre las proteínas del grupo III (Fig-13) y los demás grupos. Una de las características más sobresalientes del grupo II es que casi todas las proteínas contienen fusionados tanto el dominio conservado de las HPK como el de las RR (Figura 12-C), tal es el caso de las proteínas de eucariotes Sln1 y Etr1 y las mayoría de las otras proteínas de procariotes que se encuentran en este grupo. Una de las explicaciones a esto puede ser que debido a que en la mayoría de los sistemas de dos componentes, las proteínas HPK y RR se encuentran codificadas en el mismo operon, al darse una duplicación en alguna de las dos proteínas, ocurra la fusión entre una proteína HPK y una RR, este sistema ya fusionado tuvo que haberse formado antes de la divergencia entre procariotes y eucariotes, por lo que las proteínas de eucariotes fusionadas que se encuentran en este grupo (Figura 13-III), tuvieron que haber sufrido la inserción de los intrones posterior a la divergencia ya que ninguno de los organismos procariotes contienen intrones. La teoría de los intrones tardíos está más aceptada que la de los intrones tempranos, no porque alguna de ellas se pueda descartar, si no porque existen más evidencias a favor de la primera que de la segunda.

El cuarto grupo (Fig. 13-IV) contiene a las proteínas NarX de *E. coli*, DegS de *B. subtilis*, ComP de *B. subtilis*, NarQ de *E. coli*, HydH de *E. coli* UspT de *E. coli* y AgrB de*S. aureus*. Como en el grupo anterior, las distancias que existen entre ellas son en general menores que con las demás proteínas a excepción de AgrB, la cual se encuentra más cercana a proteínas fuera del grupo que a algunas de su grupo (Apéndice-4). Las proteínas DegS y ComP de *B. subtilis*, pudieron haberse generado a partir de una duplicación, siendo así proteínas parálogas.

Las proteínas NarQ y NarX son muy parecidas entre sí, es muy probable que estas hayan surgido por un evento de duplicación reciente, ya que incluso las dos proteínas son capaces de transducir la señal a la proteína NarL. Que estas dos proteínas HPK sean funcionales para una misma proteína RR, quizá nos pueda dar evidencia de como han ido surgiendo los sistemas de dos componentes (Fig. 13-IV) (ver más adelante). Finalmente, las proteínas UspT y HydH de *E. coli*, son más parecidas a este grupo de proteínas, sin embargo, las distancias (Apéndice-4) con las demás proteínas de la familia HPK no varían demasiado, es decir que estas proteínas no han divergido tanto como las demás de su grupo o que acaban de surgir de algún evento de duplicación reciente.

Lo que podemos decir en general del dominio de las proteínas HPK es que se encuentra muy conservado ya que los intervalos de divergencia que existe entre ellas no es muy grande, las proteínas más alejadas incluso comparten más de un 40% de similitud, evidentemente estas proteínas forman una gran familia de HPK.

Al igual que con las proteínas HPK, el análisis con las RR se corroboró con las distancias obtenidas por el programa PROTDIST (Apéndice-4).

En esta familia de proteínas fué identificado un grupo bien definido, el cual contiene a las proteínas NtrC de *R. capsulatus, R. leguminosarum, Bradyrhizobium sp, P. vulgaris, R. meliloti* y *E. coli*; HydG de *K. pneumoniae* y *S. typhimurium*; PilR de *P. aeruginosa*; NtrX de *A. caulinodans*; PgtA de *S. typhimurium*; DctD de *R. meliloti* y *R. leguminosarum*; NodW de *B. japonicum*; FixJ de *A. caulinodans, B. japonicum* y *R. meliloti*; Spo0F de *B. subtilis* y CheB de *B. subtilis* (Fig. 14-I).

Las proteínas FixJ son proteínas ortólogas; las relaciones evolutivas que estas guardan entre sí son congruentes con los árboles obtenidos para los genes ribosomales 16S (Apéndice-3). En los árboles realizados por el programa PROTPARS (Fig-16), se observó que las proteínas FixJ están más relacionadas con la proteína NodW, sin embargo, mediante el método de FITCH, aunque las proteínas se encuentran aparentemente más relacionadas con NodW, las distancias que existen son menores con las proteínas DctD que con NodW (Apéndice-4), muy probablemente esto quiera decir que la proteína NodW ha divergido más rápidamente que las demás proteínas. Las proteínas DctD, en los árboles realizados con el programa PROTPARS muestran que están más relacionadas a NtrX que a PgtA, al obtener las distancias entre estas proteínas, efectivamente DctD se encuentra más cerca a NtrX que a PgtA. Por otro lado, es posible pensar que las proteínas DctD y PgtA comparten un ancestro común ya que estas proteínas estan encargadas de regular sistemas de transporte. NtrX es una proteína más parecida a DctD que a NtrC, aunque con estas últimas compartan sistemas de regulación parecidos, ya que los dos estan involucrados en la asimilación de nitrógeno.

En este árbol se observó que las proteínas NtrC se encuentran claramente separadas por las proteínas HydG y PilR. Sin embargo, existe una menor distancia entre las proteínas NtrC que con cualquier otra proteína (Apéndice-4). Por otro lado, las relaciones evolutivas que guardan estas proteínas son congruentes con los árboles obtenidos con los genes ribosomales 16S (Apéndice-3), ya que *R. capsulatus, R. melioloti* y *Bradyrhizobium sp*; son proteobacterias que pertenecen a la subdivisión α y *P.vulgaris, K.pneumoniae* y *E. coli* son de la subdivisión γ. Sin embargo, las proteínas HydG al igual que PilR se encuentren más cercanas de las NtrC que de cualquier otra proteína y es por esto que se agruparon ahí.

Las proteínas de este primer grupo estan relacionadas con las HydG que se encargan de la regulación de las hidrogenasas. Estas proteínas mantienen una alta similitud con NtrC y un alto grado de conservación en algunos aminoácidos al igual que con otras proteínas. De la proteína PilR se sabe que participa en la regulación de fimbrias. Esta proteína, al igual que en las sensoras, se encuentran relacionadas con este primer grupo sin compartir una función muy parecida. Estas proteínas además de estar muy relacionadas en todos los árboles realizados, pertenecen a la familia de activadores del factor σ^{54}.

La proteína Spo0F de *B. subtilis*, se encuentra más cercana de las proteínas de este grupo que de cualquier otra. Spo0F es mejor receptor que Spo0A y que Spo0F que es el sustrato real de Sp2J. Spo0F es menos parecida a Spo0A que a otras proteínas.

La proteína CheB de *B. subtilis*, es más parecida a Spo0F de *B. subtilis*, luego a las proteínas CheY y por último a las otras proteínas CheB. La primera explicación es que esta proteína, junto con Spo0f de *B. subtillis* pueden provenir de un ancestro común, y que divigieron en *B. subtillis* después por un evento de duplicación surgieron las dos proteínas, las cuales divirgieron hasta responder a distintos estímulos. El que CheB realice la misma función en *B. subtillis* que las demás CheB de otros organismos es quizá un evento de convergencia funcional o transferencia horizontal.

En un segundo grupo (Fig. 14-II), están las proteínas PfeR de *P. aeruginosa*; ArcA de *E. coli*; Afsq1 de *S. coelicolor*; OmpR de *P. aerugineum, Cyanidium caldarium, S. typhimurium* y *E. coli*; VirG de *A. tumefaciens*-5, *A. tumefaciens*-6 y *A. rhizogenes*; CreB de *E. coli*; VanR de *Enterococcus faecium*; BaeR de *E. coli*; PhoP *B. subtilis, S. typhimurium* y *E. coli*; PhoB de *R, meliloti, P. aeruginosa*, y *E. coli*; BasR de *E. coli*, y KdpE de *E. coli*. Este agrupamiento coincide con el agrupamiento por estructura, ya que todas estas son proteínas de que contienen un dominio particular de unión a DNA.

Las proteínas VirG se encuentran más cercanas entre ellas que con cualquier otra proteína, y todas ellas se parecen más a las proteínas de este grupo que a las demás. Las proteínas PhoB se encuentran más cercanas entre ellas que con cualquier proteína, su distribución en el árbol es congruente con los árboles obtenidos con RNA ribosomal 16S (Apéndice-3), ya que PhoB de *E. coli* y *P. aeruginosa* se encuentran más cercanas y los dos organismos pertenecen al grupo de proteobacterias de la subdivisión γ. La de *R.meliloti* está un poco más alejada, ya que pertenece a la subdivisión α.

Por otro lado las proteínas PhoP de *S. typhimurium* y *E. coli* estan muy cercanas, ya que pertenecen a la misma subdivisión γ de las proteobacterias.

Sin embargo, no todas las proteínas PhoP forman un grupo, ya que PhoP de *B. subtilis* se encuentra más alejada de las demás proteínas Pho que de otras proteínas. La proteína BasR de *E. coli* está muy relacionada con las proteínas PhoP, estas se encuentran distribuidas de la misma forma en todos los árboles realizados y comparten un ancestro.

La proteína KdpE de *E. coli* funciona como regulador del transporte de potasio, y se encuentra muy relacionada con las proteínas PhoB, PhoP y con CreB. La proteína CreB se encuentra muy relacionada con las proteínas Pho y comparten un ancestro. Incluso realizan funciones involucradas en la regulación de fosfato. La proteína VanR de *E. faecium* regula la expresión de un gene que da resistencia a antibióticos y la producción de D,D-carboxipeptidasa. Está muy relacionada con este grupo de proteínas y en particular con las proteínas Pho.

Otras proteínas de este grupo son las OmpR, las de *E. coli* y*S. typhimurium* se encuentran muy cercanas debido a que pertenecen a la misma subdivisión γ. Las otras OmpR son de plantas, por lo que se encuentran más alejadas y coinciden con los árboles obtenidos mediante los 16S ribosomales (Apéndice-3). Estas se encuentran separadas por la proteína BaeR de *E. coli*, de la cual no se conoce su función. En estas proteínas es interesante saber como se pueden presentar proteínas OmpR tanto en bacterias como en plantas. Uno de los caminos que esto pudo haber seguido es que debido a que las proteínas OmpR de bacterias son muy parecidas a las proteínas VirG, es muy probable que estas compartan una proteína ancestral, la cual se fué especializando ya sea hacia VirG o a OmpR. Posteriormente por un evento de transferencia horizontal la proteína VirG de bacterias pasó hacia las plantas y una vez ahí se diferenció nuevamente hacia OmpR, por un evento de reversión en la función. Otra de las explicaciones es que se pudo haber dado un evento de convergencia funcional entre las OmpR de bacterias y las de plantas.

Las proteínas de este grupo, guardan en general la misma relación entre ellas y comparten un ancestro en todos los árboles que se realizaron.

En el tercer grupo (Fig. 14-III), encontramos a las proteínas Sln1 de *S. cerevisiae;* LemA de *P. syringae;* RcsB de *E. coli;* Etr1 de *A. thaliana;* RcsC de *E. coli;* EvgS de *E. coli* y a las CheY de *S. typhimurium* y *E. coli.* Estas proteínas se encuentran agrupadas de la misma forma en todos los árboles realizados, aunque en algunos de ellos la relación entre las proteínas es un poco promiscua, ya que en algunas ocasiones estas proteínas se encuentran en distancia más cercanas a otras que con las del grupo (Apéndice-4), sobre todo con proteínas del grupo IV como CheB.

Por otro lado tenemos a la proteína Sln1 de *S.cerevisiae,* esta proteína no esta bien caracterizada, solo se sabe que mutantes en el gene son letales.

El último grupo (Fig. 14-IV) no está muy bien definido. En este se encuentran las proteínas Spo0A de *B. subtilis*; CheB de *S. typhimurium* y *E. coli*; BvgA de *B. pertusis*; NarL de *E. coli*; DegU de *B. subtilis*; estas proteínas estan muy relacionadas con las del grupo tres, incluso en los árboles realizados con el programa PROTPARS, el grupo tres y cuatro se encuentran mezclados. Las proteínas más alejadas de toda la familia son las proteínas ComA de *B. subtilis* y AgrA de *S. aureus*.

Como se puede observar, en todos los árboles realizados se encontró que estas proteínas mantienen de manera general las mismas relaciones, sobre todo y más evidentemente en el grupo uno, en donde se encuentran todas aquellas proteínas que participan en el proceso simbiótico de fijación de nitrógeno. Las proteínas más alejadas de esta familia mantienen un porcentaje de similitud del 50%. En estos árboles, el intervalo de variabilidad es mayor entre las proteínas RR que entre las HPK, es por esto que el tamaño de las ramas son menores en este último.

Es muy interesante que al comparar los árboles de las proteínas reguladoras con las sensoras se observa el mismo agrupamiento entre los pares de proteínas que forman los sistemas de dos componentes.

Las relaciones son más evidentes entre las proteínas de los grupos uno de las HPK y de las RR, en las dos familias de proteínas además de estar agrupadas de la misma forma, comparten historias evolutivas similares. Por ejemplo, proteínas como FixJ y NodW comparten un ancestro en común al igual que las proteínas FixL y NodV. También se guarda la misma relación entre PgtA y DctD que entre PgtB y DctB. Todas estas a su vez se encuentran relacionadas con las proteínas NtrB y NtrC respectivamente.

Como se pudo observar en este grupo se encuentran una gran parte de proteínas que participan en el proceso simbiótico de fijación de nitrógeno. Muy probablemente esto nos hable de que originalmente exitían proteínas encargadas de realizar estas funciones y mediante eventos de duplicación y mutación estas se fueron especializando para responder a cada uno de los estímulos que afectan la simbiosis, hasta formar distintos sistemas capaces de sensar y responder a cada uno de los estímulos, esto comprueba que las proteínas HPK y RR han compartido una historia evolutiva.

Los árboles realizados con todas las proteínas reguladoras y sensoras localizadas en los bancos de datos nos permitieron saber que las proteínas mantenían su relación, y que esta no se veía afectada por el número de proteínas, comparándolas con los árboles que solo contenían aquellas proteínas que forman sistemas de dos componentes.

Evidentemente no podemos decir que la relación evolutiva entre las proteínas sea exactamente como la presentamos, ya que creemos que todavía existen muchas proteínas del sistema de dos componentes las cuales no se han encontrado y que pudieran cambiar de alguna manera la topología de los árboles aunque no la relación general que hemos observado de estas dos familias de proteínas.

Ahora con todos estos datos, es necesario proponer como han surgido y como ha sido la historia evolutiva de las proteínas de dos componentes.

5.4. Propuesta de evolución del sistema de dos componentes

Una de las vias evolutivas que pudieron haber seguido estas proteínas se describe a continuación (Fig. 17). Inicialmente, existió un sistema de dos componentes HPK-RR capaz de sensar y responder a los cambios en el ambiente, transduciendo la señal de manera específica. Al surgir un evento de duplicación en la proteína HPK, este puede fijarse o no en la población, si se fijase, el sistema cambia un poco, ya que ahora dos proteínas HPK las cuales responden a la misma señal son capaces de transducirla a una sola proteína RR, como por ejemplo, en el sistema NarQ y NarX que transducen la señal a NarL o en el de BvgC y BvgS que la transducen a BvgA. Algún tiempo después una de las dos proteínas sensoras comienza a divergir, pero siguen manteniendo la comunicación con la misma proteína RR, como por ejemplo las proteínas CpxA y ArcB que son proteínas muy diferentes las cuales transducen la señal a la misma proteína que en este caso es la proteína ArcA, y lo que se observa en este caso es que ahora cada una de las proteínas HPK provoca que la proteína RR regule la respuesta de manera diferente, ya sea positiva o negativamente. Luego al darse una duplicación de la proteína RR solo una de las proteínas HPK es capaz de interaccionar con las dos proteínas RR. Al darse la transducción de la señal la proteína HPK interacciona preferencialmente con alguna de las dos proteínas RR, como en el caso de Sp2J que envía la señal a las proteínas Spo0A y a Spo0F. A partir de este paso se pueden seguir dos caminos, el primero es que la proteína HPK se especialice por una sola de las proteínas RR y que de esta manera se forme el nuevo sistema, como en la mayoría de los casos que se encontraron en el estudio. En el segundo camino, puede ser que la proteína HPK establezca una interacción eficaz con las dos proteínas RR, siendo necesaria una mayor divergencia de la proteína HPK para poder seguir interaccionando con las dos, tal es el caso de CheA que es una proteína muy alejada e interacciona tanto con CheY como con CheB, que son dos proteínas diferentes.

Esto corrobora que aunque la comunicación cruzada se pueda dar en condiciones forzadas, esto no se ha dado a lo largo del tiempo ya que las proteínas tienden a especializarse por pares, debido a la función compartida que estas guardan.

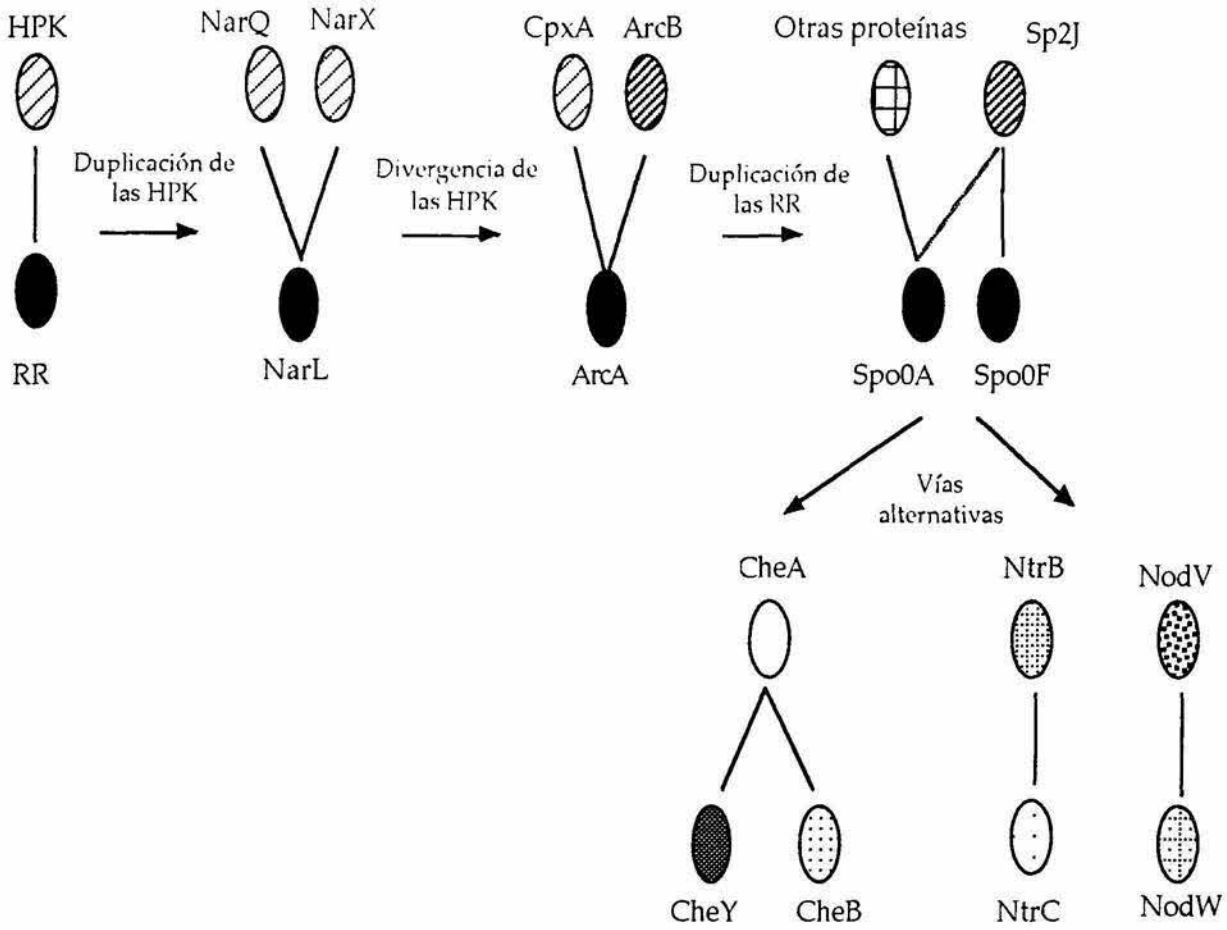


Figura 17. Representación esquemática del surgimiento de los sistemas de dos componentes, con algunos ejemplos de proteínas actuales para cada uno de los cambios.

Según este modelo, la duplicación de un miembro de una de las familias de proteínas podría fijar la duplicación de su contraparte, esto está favorecido por la interacción que llevan a cabo las proteínas, principalmente en los módulos de comunicación. Debido a que hemos observado que los módulos de comunicación comparten una historia evolutiva en común, podemos decir que estos han evolucionado de manera concertada.

Mediante este análisis se pudo observar también que este sistema de transmisión de señales esta ampliamente distribuido principalmente en las bacterias aunque también se ha encontrado en organismos eucariontes. Y que además es un sistema muy antiguo ya que éste, era funcional antes de que los organismos divirgieran (inclusive antes de la divergencia de *B. subtillis*).

6. CONCLUSIONES

-Como se pudo observar en todos los árboles realizados, tanto por el programa FITCH como por el de PROTPARS, las proteínas mantienen su relación tanto en las proteínas HPK como en las RR. La historia de las dos familias es común por lo que se trata de un evento de evolución concertada, el cual está evidentemente favorecido por la interacción de los módulos de comunicación que existe entre las dos familias.

-Los grupos I de las dos familias (HPK y RR), contienen una gran parte de proteínas que participan en el proceso simbiótico de fijación de nitrógeno, esto es muy interesante ya que en primera se eligió el sistema de dos componentes para regular este proceso, que aunque no es vital, si es muy importante; y en segundo lugar estas proteínas surgieron de un ancestro en común del cual se han especializado los distintos sistemas que ahora regulan el proceso simbiótico.

-Finalmente, estos sistemas son muy antiguos y se han mantenido a lo largo del tiempo, ya que incluso se encuentran en organismos que divirgieron hace mucho tiempo como *B. subtilis* que divirgió de *E. coli* desde hace aproximadamente mil millones de años. Por otro lado este sistema está ampliamente distribuido ya que se ha encontrado en organismos eucariotes como plantas y levaduras.

Falta página

N° 55

Etr1	Etr1	Arabidopsis thaliana		Participa en la regulación de la producción de etileno, teoricamente se cree que forman un sistema de dos componentes	10,36
EvgS	EvgS	Escherichia coli		No se conoce	D14008
FixL	FixJ	Bradyrhizobium japonicum Azorhizobium caulinodans Rhizobium meliloti	En el operon *fixLJ*	Participan en la capacidad de crecer de manera anaeróbica, estimuladas por las diferencias en la concentración de oxígeno	3
HydH	HydG	Escherichia coli Salmonella typhimurium	En el operon *hydGH*	Se encargan de regular la actividad de las hidrogenasas	67
KdpD	KdpE	Escherichia coli	En el operon *kdpDE*	Participan en la regulación del transporte de potasio	54,71
Sp2J	Spo0A, SpoF	Bacillus subtilis		Participan en el inicio de la esporulación	21,28,57,70
LemA	LemA	Pseudomona syringae	En el gene *lemA*	Encargada de la regulación de la infección en plantas, en la producción de proteasa y syringomicina	29
NarQ, NarX	NarL	Escherichia coli	En el operon *narRL*	Regulan la síntesis de enzimas envueltas en la respiración anaeróbica y la fermentación	11,18,59,62
NifR2	NifR1	Rhodobacter capsulatus	En el operon *nifR1,2*	Regula la expresión de genes requeridos para la asimilación de nitrógeno	32
NodV	NodW	Bradyrhizobium japonicum	En el operon *nodVW*	Participan en el inicio de la nodulación y en el número de nódulos, pero no tiene influencia en el número de bacteroides	25
NtrB	NtrC	Escherichia coli Rhizobium meliloti Rhodobacter capsulatus Klebsiella pneumoniae Bradyrhizobium sp. Proteus vulgaris Vibrio alginolyticus	En el operon *ntrBC*	Regula la expresión de genes requeridos para la asimilación de nitrógeno	49,50
NtrY	NtrX	Azorhizobium caulinodans	En el operon *ntrYX*	Participan en la regulación de la asimilación de nitrógeno	55
PfeS	PfeR	Pseudomonas aeruginosa		Participa en la expresión del receptor de enterobactina férrica	L07739

PgtB	PgtA	Salmonella typhimurium		Activador positivo del sistema de transporte inducible de fosfoglicerato. Regulación del transporte de fosfoglicerato.	M21279
PhoQ	PhoP	Escherichia coli Bacillus subtilis Salmonella typhimurium	En el operon *phoQP*	Participan en la regulación de la expresión de los genes de virulencia y sobrevivencia a macrófagos y en la activación de los genes de ATPasa	43
PhoR	PhoB	Escherichia coli Bacillus subtilis Pseudomonas aeruginosa Rhizobium meliloti	En el operon *phoRB*	Participan en la activación del regulon de fosfato	2,4,5,39, 40,61
PilS	PilR	Pseudomonas aeruginosa		Controla la expresión de fimbrias en Pseudomonas aeruginosa tipo 4.	Z12154
RcsC	RcsB, RcsC	Escherichia coli	En el operon *rcsBC*	Regulación de la síntesis de la cápsula	68
Sln1	Sln1	Sacaromyces cereviciae		Proteína hipotética	36,53
VanS	VanR	Enterococcus faecium		Regulan la expresión de *vanX*, necesario para la resistencia a antibióticos. También regulan la producción de D,D-carboxipeptidasa	9
VirA	VirG	Agrobacterium tumefaciens Agrobacterium rhizogenes		Escenciales para la inducción de tumores y para la transferencia del plásmido Ti	42,75
NifL		Azotobacter vinelandii		Participa en la inhibición de la actividad de NifA en respuesta al oxígeno	P30663
PleC		Caulobacter crescentus		Tiene que ver con la sintesis de aminopeptidasa, para la resistencia a fagos	M91449
VirL		Agrobacterium tumefaciens		Formación de tumores	P07167
	AlgB	Pseudomonas aeruginosa		Regulador positivo de la biosíntesis de Alginato *algD*	P23747
	AlgR	Pseudomonas aeruginosa		Regulador positivo de la biosíntesis de Alginato *algD*	P26275
	CopR	Pseudomonas syringae			L05176
	CutR	Streptomyces lividans			X58793

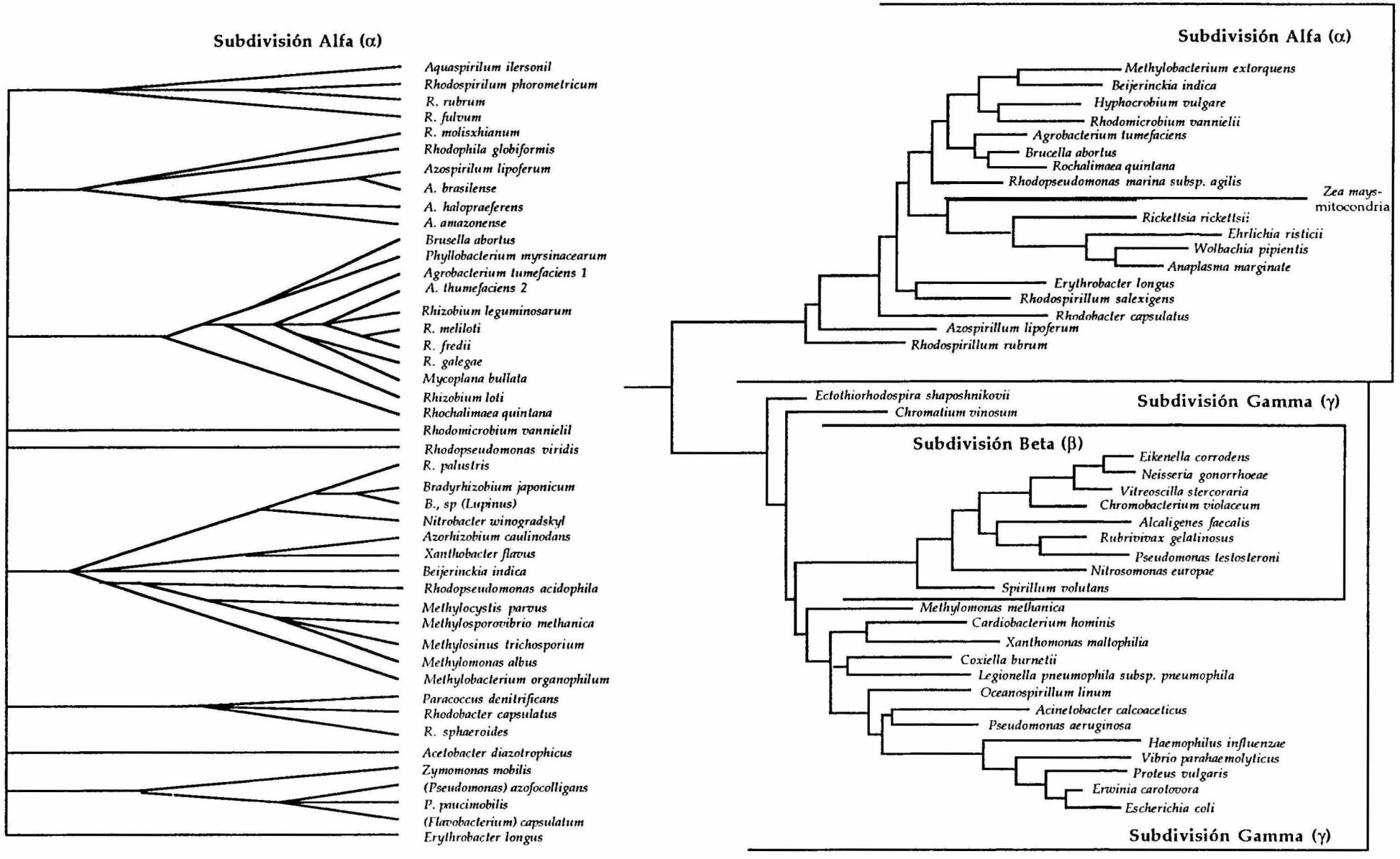
Apéndice 1 (d)

FimZ	Escherichia coli	En la sintesis de fimbrias	P21502
FliA	Pseudomonas aeruginosa	Requerido para la síntesis de flagelina	X61231
FrzE	M. xanthus	Proteína reguladora de la motilidad	M35192
GlpR	Pseudomonas aeruginosa	Proteína reguladora de glicerol	M60805
HoxA	Alcaligenes eutrophus	Regulación de la actividad de hidrogenasa	P29267
HupR	Rhodobacter capsulatus	Involucrado en la regulación de *nifE*	P26408
MrkE	Klebsiella pneumoniae	Regula la expresión de fimbrias	P21649
PatA	Anabaena sp	Formación del patrón de heterocisto	M87501
PcoR	xxx		
PetR	Rhodobacter capsulatus		Z12113
RcaC	Fremyella diplosiphon		M95680
RegA	Rhodobacter capsulatus		M64976
RteB	B. th	Produce formas como plásmidos	M81439
Skn7P	Sacaromyces cereviciae		
SpaR	Bacillus subtillis		L07785
TctD	Salmonella typhimurium	Activador transcripcional del operon de sistema de transporte de tricarboxilato	P22104
UhpA	Escherichia coli	Activador de UHPT, transporte de azúcares	P10940
UhpA	Salmonella typhimurium		P27667
VirL	Agrobacterium tumefaciens	Formación de tumores	P07167
Xcc1	Xanthomonas campestris		X54015
YecB	Escherichia coli		P07027
Ylb3	Leptospira interrogans		P24086

C Cys	12																					C-sulfidrilo
S Ser	0	2																				S-pequeños hidrofílicos
T Thr	-2	1	3																			
P Pro	-3	1	0	6																		
A Ala	-2	1	1	1	2																	
G Gly	-3	1	0	-1	1	5																
N Asn	-4	1	0	-1	0	0	2															N-ácido, ácido amida, hidrofílico
D Asp	-5	0	0	-1	0	1	2	4														
E Glu	-5	0	0	-1	0	0	1	3	4													
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4												
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6											H-básicos
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6										
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5									
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6								V-pequeños hidrofóbicos
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5							
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6						
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4					
F Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9				F-aromáticos
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10			
W Trp	8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17		
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W		
	Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp		

Apéndice 2. Matriz de Dayhoff PAM250 (Mutaciones de Aminoácidos por cada 100 residuos de aminoácidos). Los aminoácidos estan arreglados asumiendo los valores positivos que representan los reemplazos evolutivos conservados, los grupos estan formados en base a las propiedades fisicoquímicas de los aminoácidos.

Apéndice 3. Arboles realizados con los genes ribosomales 16S donde se muestran las subdivisiones α, β y γ (Young, J. P. 1992; Olsen, G. *et al* 1994).



Apéndice 4 (a). Distancias de las proteínas HPK.

Narqeco	1	000
Narxeco	2	035	000
Compbsu	3	081	082	000
Degsbsu	4	083	079	082	000
Cheaeco	5	163	231	319	340	000
Cheasty	6	320	410	492	446	019	000
Cheabsu	7	307	390	412	409	076	076	000
Ntrbeco	8	099	115	154	173	261	305	285	000
Ntrbkpn	9	099	118	151	171	257	311	286	004	000
Ntrbpvu	10	106	118	167	177	272	318	284	014	015	000
Ntrbval	11	103	116	154	166	271	309	289	022	024	025	000
Ntrbrca	12	106	119	166	177	266	335	342	059	058	061	060	000
Ntrbbsp	13	112	139	193	188	266	301	298	060	061	061	062	043	000
Ntryaca	14	114	140	147	162	312	349	291	075	077	078	077	080	079	000
Viraat6	15	119	137	158	179	235	325	293	084	083	089	088	099	094	091	000
Viraat9	16	120	137	159	180	236	325	294	084	083	089	088	100	094	091	000	000
Hydheco	17	037	070	110	112	158	294	252	081	082	085	081	084	084	097	089	089	000
Sp2jbsu	18	094	124	134	157	274	353	285	088	089	093	085	082	085	085	078	077	067	000
Dctbrle	19	098	107	145	148	233	300	291	088	088	091	090	080	087	079	091	090	080	065	000
Dctbrme	20	094	106	136	142	226	301	289	088	088	090	087	079	086	081	090	090	069	068	018	000
Fixlaca	21	101	114	153	164	273	382	317	088	089	093	090	089	095	085	084	084	076	067	065	069	000
Fixlbja	22	099	111	143	165	221	347	280	085	085	091	090	086	089	086	080	080	074	067	062	064	023	000
Fixlrme	23	104	112	148	183	264	368	295	086	086	091	091	083	090	085	086	085	083	069	059	062	032	034	000
Nodvbja	24	097	112	149	141	234	340	284	087	085	090	094	091	093	092	079	078	073	072	066	062	060	057	053	000
Pgtbsty	25	094	108	154	139	264	359	341	104	102	110	100	106	113	106	100	099	080	078	062	060	069	065	071	077	000
Pilspae	26	104	128	138	158	229	316	273	095	092	099	099	092	108	112	087	087	088	077	086	086	089	080	090	086	092	000
Phorbsu	27	100	102	132	151	234	300	282	089	089	095	090	097	105	093	092	092	085	080	090	093	099	090	098	102	102	101	000
Phoreco	28	089	106	127	128	246	299	253	101	101	103	096	109	096	106	105	105	094	096	105	103	116	109	110	101	104	109	057	000
Vansefa	29	100	120	134	136	241	303	298	095	094	095	093	098	102	101	098	098	084	091	097	093	110	100	107	109	102	107	060	075	000
Crececo	30	089	109	127	134	282	367	352	098	094	096	094	095	103	095	092	093	094	092	094	088	111	104	112	098	106	105	072	070	076	000
Afsq2sco	31	101	108	136	156	244	305	287	110	108	113	107	110	115	109	114	114	104	100	099	100	118	105	111	119	102	096	070	073	080	079	000
Bvgcbpe	32	099	119	146	169	220	278	278	104	102	105	100	114	114	122	099	099	095	099	106	103	098	093	096	102	104	106	076	084	093	092	087	000
Bvgsbbr	33	102	121	144	167	232	279	273	106	103	107	101	117	113	120	098	098	097	098	105	104	097	095	096	104	105	108	076	084	094	091	088	001	000
Evgseco	34	124	136	215	167	198	264	265	133	131	131	125	138	139	131	120	120	110	112	121	124	132	120	132	138	119	129	090	105	112	107	102	061	062	000
Baraeco	35	096	120	149	158	243	313	301	106	107	107	101	112	111	112	096	096	090	092	101	105	107	100	102	109	098	117	067	085	085	097	095	057	057	081	000
Lemapsy	36	100	121	159	151	235	304	258	104	101	107	103	104	115	108	104	104	088	095	096	101	105	097	101	103	099	115	066	081	085	102	096	053	053	070	033	000
Etrlath	37	116	151	169	154	298	344	295	112	111	115	109	126	116	123	114	114	104	103	115	116	128	123	125	121	117	130	076	091	095	097	101	074	072	090	058	060	000
Rcsceco	38	105	129	132	142	273	333	303	124	124	124	123	110	123	116	109	109	097	098	110	107	115	112	109	102	104	124	073	086	089	090	096	063	062	080	054	050	067	000
Arcbeco	39	095	119	146	163	257	335	285	112	112	114	109	116	114	109	116	116	095	102	120	121	120	109	106	114	118	119	074	079	095	088	096	062	061	080	060	056	067	069	000
Slnisce	40	328	431	432	406	286	336	319	182	189	196	194	212	207	255	213	213	218	206	221	224	247	229	219	218	226	228	148	169	192	218	206	151	150	149	129	123	142	135	136	000
Kdpdeco	41	102	119	132	141	250	315	317	097	097	099	097	102	108	100	106	107	102	102	096	091	119	102	110	111	110	106	077	084	084	088	090	078	078	097	082	082	088	081	085	177	000
Envzeco	42	101	119	143	136	280	327	293	112	110	115	102	115	110	100	125	126	089	111	100	099	121	121	115	121	116	112	081	076	097	098	092	097	096	129	099	102	115	090	103	206	083	000
Envzsty	43	103	121	146	135	287	322	299	110	111	116	101	116	114	099	126	127	091	114	105	105	121	124	120	127	121	117	081	078	099	100	097	098	098	129	100	102	115	091	104	212	084	004	000
Cpxaeco	44	107	119	140	147	236	273	243	100	099	106	100	116	115	095	113	114	092	096	100	095	110	110	115	113	108	115	073	074	083	087	087	092	093	107	090	083	094	089	092	177	084	062	063	000
Uhpteco	45	055	083	108	130	159	288	294	102	102	109	114	113	116	120	124	124	058	115	107	098	118	117	114	121	120	118	093	086	091	086	105	101	104	130	105	106	118	100	097	234	093	075	078	064	000
Pfespae	46	106	126	155	162	243	293	280	123	122	124	119	127	126	116	152	153	092	120	127	123	138	136	133	139	133	132	094	090	102	097	104	099	099	123	104	105	118	097	099	232	092	069	071	072	084	000
Basseco	47	091	107	118	140	265	308	319	101	100	109	099	107	116	116	124	124	090	095	104	097	113	109	109	121	112	110	086	081	092	090	095	095	096	117	094	095	110	095	096	197	091	080	082	077	076	085	000
Phoqeco	48	092	108	128	164	244	319	288	098	100	103	105	104	113	106	122	122	094	100	103	100	115	104	112	118	100	116	087	082	084	083	096	095	097	116	100	099	113	092	091	236	081	093	096	082	084	094	088	000
Agrasau	49	100	142	153	143	294	389	375	139	138	143	138	156	175	149	163	163	104	143	146	138	149	150	163	169	155	169	125	122	108	119	132	130	130	155	126	128	145	125	123	401	127	120	118	122	105	123	121	103	000
Proteína #		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49

Apéndice 4 (b). Distancias de las proteínas RR.

Proteínas #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56
Agrasau 1	000																																																							
Comsbau 2	115	000																																																							
Chebaco 3	128	154	000																																																				
Chebsty 4	128	152	001	000																																																			
Degubau 5	115	103	113	112	000																																																		
Marlaco 6	138	127	123	122	079	000																																																		
Bvgebpe 7	129	144	119	119	094	102	000																																																	
Sposbau 8	126	167	101	098	104	104	104	000																																																
Cheyaco 9	138	144	135	135	119	118	113	130	000																																														
Cheysty 10	136	144	132	132	119	114	115	129	001	000																																													
Bvgeaco 11	169	162	157	157	117	111	104	142	092	092	000																																												
Rcaceco 12	125	163	119	118	120	120	108	132	082	083	062	000																																											
Lamapey 13	165	176	157	157	163	153	151	177	111	110	109	096	000																																										
Blnlaco 14	173	159	174	172	148	151	177	188	135	135	122	132	139	000																																									
Virgarh 15	122	133	111	108	124	125	138	125	124	123	141	134	172	206	000																																								
Virgat5 16	156	174	139	135	155	140	169	157	150	149	177	166	184	208	011	000																																							
Virgat4 17	151	174	127	123	145	133	162	150	130	137	161	158	176	213	016	014	000																																							
Phopaco 18	118	147	114	114	132	123	095	121	099	099	112	095	132	164	088	113	109	000																																						
Phopsty 19	128	148	119	118	130	135	098	120	101	101	131	092	131	166	087	113	109	006	000																																						
Baeraco 20	114	153	132	130	136	125	124	133	103	105	123	101	140	170	099	123	112	062	063	000																																					
Phokaco 21	125	124	119	119	117	125	116	125	098	093	141	113	139	158	099	121	112	076	075	087	000																																			
Phokpse 22	119	126	116	116	116	116	114	117	107	105	129	109	139	156	100	118	116	080	087	087	041	000																																		
Phobrae 23	133	135	137	137	115	137	112	112	104	103	118	117	125	170	098	119	114	075	072	078	054	063	000																																	
Kdpaaco 24	113	121	121	118	107	103	109	119	105	100	115	111	127	166	079	101	093	073	070	072	068	069	069	000																															
Gmpraa 25	124	149	119	117	136	133	140	122	106	102	109	108	130	148	100	109	103	110	107	091	085	081	087	091	000	000																											
Gmppse 26	124	149	119	117	134	133	140	122	104	102	109	108	130	148	099	109	103	110	107	091	085	081	087	091	000	000																											
Afaplaco 27	122	101	131	129	119	107	103	121	118	113	124	106	130	151	084	105	101	083	082	085	075	077	067	078	066	066	000																										
Phopbau 28	120	130	107	107	112	113	098	109	097	094	109	101	116	141	087	107	103	082	083	077	060	058	053	071	060	060	056	000																										
Gmpraco 29	119	159	132	130	136	119	128	137	112	108	123	118	139	156	066	075	074	079	081	076	070	069	072	089	070	070	066	064	000																										
Gmpsty 30	120	161	133	131	137	119	131	116	112	108	123	118	139	157	066	076	075	081	083	077	071	070	078	090	070	070	067	065	000	000																									
Baeraco 31	126	158	152	151	166	137	152	158	126	131	144	128	130	162	110	112	106	110	112	103	082	082	080	102	088	088	087	072	068	068	000																									
Venrafa 32	108	126	131	131	130	123	126	123	120	119	148	124	137	160	101	131	122	085	091	072	077	065	071	089	077	077	082	058	071	072	078	000																									
Crebaco 33	117	174	124	122	130	141	133	131	121	120	135	130	136	173	109	133	127	089	087	092	079	078	067	075	086	086	072	062	088	088	080	080	000																									
Azcaaco 34	100	124	115	117	099	110	110	130	108	108	106	107	129	145	089	109	100	082	086	081	082	074	083	080	086	086	087	069	070	070	084	067	071	000																									
Pfarpae 35	135	185	148	148	172	118	149	158	126	125	134	132	122	167	126	126	116	105	102	098	097	108	111	105	092	092	090	092	089	090	093	095	109	101	000																								
Hydpaco 36	126	148	134	132	120	114	131	126	088	087	105	105	120	147	139	160	156	130	107	111	119	119	125	102	108	108	106	096	116	116	110	123	119	115	113	000																								
Hydpsty 37	128	152	138	136	120	116	137	126	087	086	107	103	118	147	140	166	159	109	110	121	125	119	128	101	106	106	106	103	125	125	126	130	128	119	122	011	000																								
Chebbau 38	110	119	100	100	098	099	102	103	077	077	121	106	118	147	108	133	124	097	102	116	083	088	095	092	104	104	098	092	110	110	113	109	126	098	125	088	084	000																								
Spofbau 39	128	137	109	109	108	120	101	110	109	110	114	093	142	161	123	148	142	119	122	119	094	091	107	102	107	107	112	088	109	109	127	110	108	107	128	084	090	072	000																								
Fixjaca 40	116	136	138	134	120	126	125	128	119	118	143	136	143	199	112	137	133	121	119	120	113	113	117	101	131	131	110	108	118	120	151	122	110	106	127	091	098	107	103	000																								
Fixjbja 41	121	138	133	131	111	137	130	136	100	099	133	129	136	171	105	130	128	110	110	116	106	111	114	113	121	121	107	102	113	114	157	127	119	110	113	081	075	096	095	043	000																								
Fixjcma 42	138	147	138	136	112	135	133	147	117	117	147	128	172	214	121	150	140	124	124	132	108	095	115	101	130	130	112	118	124	126	157	132	107	124	160	104	095	090	094	051	057	000																								
Nodvbja 43	117	175	125	122	145	157	136	153	145	144	171	145	148	201	124	136	137	132	130	127	129	127	118	114	130	130	131	122	125	127	137	133	124	124	129	114	112	116	121	076	083	077	000																								
Dctdrla 44	128	155	147	144	116	138	113	120	098	099	143	114	137	197	103	123	121	112	108	105	113	111	114	108	120	120	107	124	113	113	152	129	119	123	137	103	101	091	110	070	067	070	091	000																								
Dctdrma 45	117	138	135	132	113	122	114	120	095	096	128	103	128	187	108	135	123	104	107	110	118	121	111	106	120	120	094	120	107	107	134	111	117	101	126	093	086	083	103	075	069	074	089	024	000																								
Pgtaaty 46	133	157	173	171	152	138	170	160	126	129	154	160	174	184	159	190	175	134	135	132	131	131	135	148	160	160	130	139	134	135	169	138	142	137	154	105	102	110	129	092	081	096	105	081	072	000																								
Mtrcaco 47	113	139	119	117	107	110	132	103	101	099	127	097	122	187	108	135	126	098	096	106	105	098	099	084	105	105	102	095	104	105	116	110	093	106	099	077	074	088	080	076	069	089	097	081	068	105	000																								
Mtrckpn 48	116	143	118	117	108	118	135	108	101	099	127	101	124	189	111	139	129	102	101	105	107	101	101	085	105	105	099	099	108	108	116	110	093	111	101	078	075	091	078	079	071	088	099	081	070	109	006	000																								
Mtrcpwu 49	107	150	124	122	103	118	131	105	101	099	133	102	133	181	105	131	124	096	095	104	106	107	115	088	101	101	103	101	094	095	137	117	099	111	111	085	080	096	080	090	073	082	105	084	077	102	020	021	000																								
Mtrcber 50	137	155	122	123	119	117	115	137	118	117	114	104	147	177	117	138	131	107	107	092	109	102	116	100	102	102	096	103	097	096	126	108	112	102	108	099	098	104	092	101	093	096	121	096	093	112	070	075	068	000																								
Mtrcmse 51	141	162	123	120	122	117	120	126	107	106	122	111	161	181	117	140	134	105	106	085	108	094	120	099	098	098	096	094	097	096	134	106	114	103	110	095	095	102	098	098	084	088	121	092	094	102	070	074	069	014	000																								
Mtrcrca 52	142	157	128	125	126	116	113	134	100	098	132	116	152	171	109	135	129	104	107	092	108	102	112	085	111	111	093	097	105	106	142	105	122	112	121	098	099	102	107	113	105	094	111	095	084	103	079	081	076	036	036	000																								
Mtrxaca 53	122	126	138	136	124	126	121	133	106	103	158	135	148	177	132	168	159	119	121	114	082	089	107	094	113	113	109	100	120	119	121	093	174	101	118	099	091	094	099	103	086	090	126	078	071	099	080	082	086	096	088	094	000																								
Pilrpae 54	135	128	121	118	116	130	146	133	118	116	136	104	137	172	131	163	140	117	117	110	111	097	118	103	110	110	111	105	108	109	133	099	127	115	128	096	089	098	093	102	105	099	121	097	096	106	074	075	080	090	085	095	079	000																								
Rcabaco 56	159	190	180	177	130	127	156	150	163	161	161	153	177	202	172	221	207	153	158	161	186	152	169	167	154	154	167	154	154	157	213	183	173	152	229	208	201	187	191	189	196	174	226	172	174	203	175	179	160	164	165	169	199	189	122	000

7. REFERENCIAS

1. Albright, L; Huala, E; y Ausubel, F; 1989 "Procaryotic signal transduction mediated by sensor and regulator protein pairs" *Annu.Rev.Genet.* **23**:311-336

2. Anba, J; Bidaud, M; Vasil, M; y Lazdunski, A; 1990 "Nucleotide Sequence of the *Pseudomonas aeruginosa phoB* Gene, the Regulatory Gene for the Phosphate Regulon" *J.bacteriol.* **172**:4685-4689

3. Anthamatten, D; y Hennecke, H; 1991 "The regulatory status of the *fixL*-and *fixJ*-like genes in *Bradyrhizobium japonicum* may be different from that in *Rhizobium meliloti*" *Mol Gen Genet* **225**:38-48

4. Amemura, M; Makino, K; Shinagawa, H; y Nakata, A; 1986 "Nucleotide Sequence of the *phoM* Region of *Escherichia coli*:Four Open Reading Frames May Constitute an Operon" *J.bacteriol.* **168**:294-302

5. Amemura, M; Makino, K; Shinagawa, H; y Nakata, A; 1990 "Cross Talk to the Phosphate Regulon of *Escherichia coli* by PhoM Protein: PhoM Is a Histidine Protein Kinase and Catalyzes Phosphorylation of PhoB and PhoM-Open Reading Frame 2" *J.bacteriol.* **172**:6300-6307

6. Ames, P; Parkinson, J. S. 1988 "Transmembrane signaling by bacterial chemoreceptors: *E.coli* transducers with locked signal output". *Cell.* **55**:59-65

7. Aricó, B; Miller, J; Roy, C; Stibitz, S; Monack, D; Falkow, S; Gross, R; y Rappuoli, R; 1989 "Sequences required for expression of *Bordetella pertussis* virulence factors share homology with prokaryotic signal transduction proteins" *Proc.Natl.Acad.Sci.USA* **86**:6671-6675

8. Aricó, B; Scarlato, V; Monack, D; Falkow, S; y Rappuoli, R; 1991 "Structural and genetic of the *bvg* locus in *Bordetella* species" *Molecular Microbiology* **5**:2481-2491

9. Arthur, M; Molinas, C; y Courvalin, P; 1992 "The VanS-VanR two-component Regulatory System Controls Synthesis of Depsipeptide Peptidoglycan Precursors in *Enterococcus faecium* BM4147" *J.bacteriol.* **174**:2582-2591

10. Chang, C; Kwok, S; Bleecker, A; y Meyerowitz, E; 1993 "*Arabidopsis* Ethylene-Response Gene *ETR1*:Similarity of Product to two-component Regulators" *Science* **262**:539-544

11. Chiang, R; Cavicchioli, R; y Gunsalus, R; 1992 "Identification and characterization of *narQ*, a second nitrate sensor for nitrate-dependent gene regulation in *Escherichia coli*" *Molecular Microbiology* 6:1913-1923

12. Comeau, D; Ikenaka, K; Tsung, K; y Inouye, M; 1985 "Primary Characterization of the Protein Products of the *Escherichia coli ompB* Locus: Structure and Regulation of Synthesis of the OmpR and EnvZ Proteins" *J.Bacteriol* **164**:578-584

13. Darwin, C; 1859 "The Origin of Species" London: Murray

14. DeVault, J. D. *et al*, 1989 Bio/Technology 7:352-357

15. Deveraux, J; Haeberli, P; Smithies, O. 1984 "A comprehensive set of sequence analysis programs for the VAX". *Nucleic Acids Res.* **12**:387-395

16. Doolittle, R; 1992 "Reconstructing History with Amino Acid Sequences" *Protein Science* **1**:191-200

17. Drury, L; y Buxton, R; 1985 "DNA Sequence Analysis of the *dye* Gene of *Escherichia coli* Reveals Amino Acid Homology between the Dye and OmpR Proteins" *The Journal of Biological Chemistry* **260**:4236-4242

18. Eck, R; y Dayhoff, M; 1966 "Atlas of Protein Sequence and Structure" Silver Springs Md. Natl. Biomed. Res. Found.

19. Egan, S; y Stewart, V; 1991 "Mutational Analysis of Nitrate Regulatory Gene *narL* in *Escherichia coli* K-12" *J.bacteriol.* **173**:4424-4432

20. Engelke, T; Jording, D; Kapp, D; y Pühler, A; 1989 "Identification and Sequence Analysis of the *Rhizobium meliloti dctA* Gene Encoding the C4-Dicarboxylate Carrier" *J.Bacteriol.* **171**:5551-5560

21. Felsenstein, J; 1988 "Phylogenies from Molecular Sequences: inference and reliability" *Annu.Rev.Genet.* **22**:521-565

22. Ferrari, F; Trach, K; LeCoQ, D; Spence, J; Ferrari, E; y Hoch, J; 1985 "Characterization of the *spo0A* locus and its deduced product" *Proc.Natl.Acad.Sci.USA* **82**:2647-2651

23. Fitch, W; Margoliash, E; 1967 "Construction of Phylogenetic Trees" *Science* **155**:279-284

24. Futuyma, D; 1986 "Evolutionary Biology" 2a. ed. Sinauer Associates, Inc. Sunderland, Massachusetts pp. 600

25. Genetics Computer Grup, Inc. versión 7.2; 1993. University Research Park 575 Science Drive, Suite B Madeson, Wisconsin 53711.

26. Göttfert, M; Grob, P; y Hennecke, H; 1990 "Proposed regulatory pathway encoded by the *nod*V and *nod*W genes, determinants of host specificity in *Bradyrhizobium japonicum*" *Proc.Natl.Acad.Sci.USA* **87**:2680-2684

27. Gross, R; Aricó, B; y Rappuoli, R; 1989. "Families of bacterial signal-transducing proteins" *Molecular Microbiology* 3:1661-1667

28. Hess, J; Oosawa, K; Kaplan, N; y Simon, M; 1988 "Phosphorylation of Three Proteins in the Signaling Pathway of Bacterial Chemotaxis" *Cell* **53**:79-87

29. Hoch, J; Trach, K; Kawamura, F; y Saito, H; 1985 "Identification of the Transcriptional Supressor *sof*-1 as an Alteration in the *spo0*A Protein" *J.bacteriol.* **161**:552-555

30. Hrabak, E; y Willis, D; 1992 "The *lem*A Gene Required for Pathogenicity of *Pseudomonas syringae* pv. syringae on Bean Is a Member of a Family of two-component Regulators" *J.bacteriol.* **174**:3011-3020

31. Ishizuka, H; Horinouchi, S; Kieser, H; Hopwood, D; y Beppu, T; 1992 "A putative two-component Regulatory System Involved in Secundary Metabolism in *Streptomyces spp*" *J.Bacteriol.* **174**:7585-7594

32. Iuchi, S; Matsuda, Z; Fujiwara, T; y Lin, E; 1990 "The *arc*B gene of *Escherichia coli* encodes a sensor-regulator protein for anaerobic repression of the *arc* modulon" *Molecular Microbiology* **4**:715-727

33. Jones, R; y Haselkorn, R; 1989 "The DNA sequence of the *Rhodobacter capsulatus ntr*A, *ntr*B y *ntr*C gene analoges required for nitrogen fixation" *Mol Gen Genet* **215**:507-516

34. Kimura, M; 1981 "Estimation of evolutionary distances between homologous nucleotide sequences" *Proc. Natl. Acad. Sci.* **78**:454-458

35. Kimura, M; y Ohta, T; 1974 "On some principles governing molecular evolution" *Proc. Natl. Acad. Sci. USA* **71**:2848-2852

36. Kleckner , N; 1981 "Transposable elements in prokaryotes" *Ann. Rev. Genet.* **15**:341-404

37. Koshland, D; 1993 "The two-component Pathway Comes to Eukaryotes" *Science* **262**:532

38. Kunst, F; Debarbouille, M; Msadek, T; Young, M; Mauel, C; Karamata, D; Klier, A; Rapoport, G; y Dedonder, R; 1988 "Deduced Polypeptides Encoded by the *Bacillus subtilis sac*U Locus Share Homology with two-component Sensor-Regulator Systems" *J.bacteriol.* **170**:5093-5101

39. Li, W; Graur, D; 1991 "Fundamentals of Molecular Evolution" Sinauer Sociates Inc. Publishers Sunderland, Massachusetts. pp 284

40. Makino, K; Shinawa, H; Amemura, M; y Nakata, A; 1986 "Nucleotide Sequence of the *pho*B Gene, the Positive Regulatory Gene for the Phosphate Regulon of *Escherichia coli* K-12" *J.Mol.Biol.* **190**:37-44

41. Makino, K; Shinawa, H; Amemura, M; y Nakata, A; 1986 "Nucleotide Sequence of the *pho*R Gene, a Regulatory Gene for the Phosphate Regulon of *Escherichia coli*" *J.Mol.Biol* **192**:549-556

42. Margoliash, E; 1963 "Primary structure and evolution of cytochrome c." *Proc. Natl. Acad. Sci. USA* **50**:672-679

43. Margulis, L. 1970 "Origin of eucaryotic cells" Yale University Press, New Haven, Conn.

44. Melchers, L; Thompson, D; Idler, K; Schilperoort, R; y Hooykaas, P; 1986 "Nucleotide sequence of the virulence gene *virG* of the *Agrobacterium tumefaciens* octopine Ti plasmid: significant homology between *virG* and the regulatory genes *omp*R,*pho*B and *dye* of *E.coli*" *Nucleic Acids Research* **14**:9933-9942

45. Miller, S; Kukral, A; y Mekalanos, J; 1989 "A two-component regulatory system (*phoP phoQ* controls *Salmonella typhymurium* virulence" *Proc.Natl.Acad.Sci.USA* **86**:5054-5058

46. Mizuno, T; Wurtzel, E; y Inouye, M; 1982 "Osmoregulation of gene Expression" *The Journal of Biological Chemistry* **257**:13692-13698

47. Morett, E; y Segovia, L; 1992 "The σ^{54} Bacterial Enhancer-Binding Protein Family: Mechanism of Action and Phylogenetic Relationship of Their Functional Domains" *J. Bacteriol.* **175**:6067-6074

48. Mutoh, N; y Simon, M; 1986 "Nucleotide Sequence Corresponding to Five Chemotaxis Genes in *Escherichia coli*" *J.Bacteriol.* **165**:161-166

49. Nagasawa, S; Tokishita, S; Aiba, H; y Mizuno, T; 1992 "A novel sensor-regulator protein that belongs to the homologous family of a signal-transduction proteins involved in adaptative responses in *Escherichia coli*" *Molecular Microbiology* **6**:799-807

50. Nei, M; 1987 "Molecular Evolutionary Genetics" Columbia University Press New York Guildford, Surress. pp 512

51. Ninfa, A. J; Ninfa, E. G; Lupas, A. N; Stock, A; Magasanik, B; Stock, J; 1988. "Crosstalk between bacterial chomotaxis signal transduction proteins and regulators of transcription of the Ntr regulon: Evidence that nitrogen assimilation and chemotaxis are controlled by a common phosphotransfer mechanism" *Proc.Natl.Acad.Sci.USA.* **85**:5492-5496

52. Ninfa, A. J; Ueno-Nishio, S; Hunt, T. D; Robustell, B; y Magasanik, B; 1986. "Purification of nitrogen regulator II, the product of the *glnL* (*ntrB*) gene of *Escherichia coli*". *J.Bacteriol.* **168**:1002-1004

53. Nixon, B. T; Ronson, C. W; y Ausubel, F. M; 1986. "two-component regulatory system responsive to enviromental stimuli share strongly conserved domains with the nitrogen assimilation regulatory genes *ntrB* y *ntrC*" *Proc.Natl.Acad.Sci.USA.* **83**:7850-7854

54. Olsen, G, J; D. J. Lane, S. J. Giovannoni, y N, R, Pace. 1986 "Microbial ecology and evolution: a ribosomal RNA approach" *Annu. Rev. Microbiol.* **40**:337-355

55. Olsen, G, J; Woese, C, R; y Overbeek, R; 1994 "The Winds of (Evolutionary) Change: Breathing New Life into Microbiology" *J.Bacteriol.* **176**:1-6

56. Ota, I; y Varshavsky, A; 1993 "A Yeast Protein Similar to Bacterial two-component Regulators" *Science* **262**:566-569

57. Parkinson, J. S; y Kofoid, E. C; 1992. "Communication modules in bacterial signaling proteins" *Annu. Rev. Genet.* **26**:71-112

58. Pawlowski, K; Klosse, U; y Bruijn, F; 1991 "Characterization of a novel *Azorhizobium caulinodans* ORS571 two-component regulatory sistem, NtrY/NtrX, involved in nitrogen fixation and metabolism" *Mol Gen Genet* **231**:124-138

59. Pearson y Lipman 1988 "Improved tools for biological sequence comparison" *Proc Natl. Acad. Sci.* 85:2444-2448

60. Perego, M; Cole, S; Burbulys, D; Trach, K; y Hoch, J; 1989 "Characterization of the Gene for a Protein Kinase Wich Phosphorylates the Sporulation-Regulatory Protein Spo0A and Spo0F of *Bacillus subtilis*" *J.bacteriol.* **171**:6187-6196

61. Phylip Inference Package, versión 3.4c 1993. Joseph Felsenstein. Frederick Biomedical Supercomputer Center NCI-FCRDC.

62. Rabin, R; y Stewart, V; 1992 "Either of two functionally redundant sensor proteins, NarX and NarQ, is sufficient for nitrate regulation in *Escherichia coli* K-12" *Proc.Natl.Acad.Sci.USA* **89**:8419-8423

63. Ronson, C. W; Nixon, B. T; y Ausubel, F. M; 1987. "Conserved Domains in Bacterial Regulatory Proteins That Respond to Enviromental Stimuli" *Cell* **49**:579-581

64. Seki, T; Yoshikawa, H; Takahashi, H; y Saito, H; 1988 "Nucleotide Sequence of the *Bacillus subtilis phoR* Gene" *J.bacteriol.* **170**:5935-5938

65. Stewart, V; Parales, J; y Merdel, S; 1989 "Structure of Genes *narL* and *narX* of the *nar* (Nitrate Reductase) Locus in *Escherichia coli* K-12" *J.bacteriol.* **171**:2229-2234

66. Stock, A; Koshland, D; y Stock, J; 1985 "Homologies between the *Salmonella typhimurium* CheY protein and proteins involved in the regulation of chemotaxis, membrane protein synthesis, and sporulation" *Proc.Natl.Acad.Sci.USA* **82**:7989-7993

67. Stock, A. M; Chen, T; Welsh, D; y Stock, J; 1988. "CheA protein, a central regulator of bacterial chemotaxis, belongs to a family of proteins that control gene expression in response to changing enviromental conditions". *Proc.Natl.Acad.Sci.USA* **85**:1403-1407

68. Stock J. B; Ninfa, A. J; y Stock, A. M; 1989. "Protein Phosphorylation and Regulation of Adaptive Responses in Bacteria" *Microbiological Reviews* **53**:450-490

69. Stock, J. B; Stock, A. M; y Mottonen, J. M; 1990 "Signal transduction in bacteria" *Nature* **344**:395-400

70. Stoker, K; Reijnders, W; Oltmann, F; y Stouthamer, A; 1989 "Initial Cloning and Sequencing of *hyd*HG, an Operon Homologous to *ntr*BC and Regulating the Labile Hydrogenasa Activity in *Escherichia coli* K-12" *J.Bacteriol.* **171**:4448-4456

71. Stout, V; y Gottesman, S; 1990 "RcsB and RcsC:a two-component Regulator of Capsule Synthesis in *Escherichia coli*" *J.bacteriol.* **172**:659-669

72. Tanaka, T; y Kawata, M; 1988 "Cloning and Characterization of *Bacillus subtilis iep*, Wich Has Positive and Negative Effects on Production of Extracellular Proteases" *J.bacteriol.* **170**:3593-3600

73. Trach, K; Chapman, J; Piggot, P; LeCOQ, D; y Hoch, J; 1988 "Complete Sequence and Transcriptional Analysis of the *spo0*F Region of the *Bacillus subtilis* Chromosome" *J.bacteriol.* **170**:4194-4208

74. Walderhaug, M; Polarek, J; Voelkner, P; Daniel, J; Hesse, J; Altendorf, K; y Epstein, W; 1992 "KdpD and KdpE, Proteins That Control Expression of the *kdp*ABC Operon, Are Members of the two-component Sensor-Effector Class of Regulators" *J.bacteriol.* **174**:2152-2159

75. Wanner, B. L; 1992 "Is Cross Regulation by Phosphorylation of two-component Response Regulator Proteins Important in Bacteria?" *J.Bacteriol* **174**:2053-2058

76. Weinrauch, Y; Guillen, N; y Dubnau, D; 1989 "Sequence and Transcription Mapping of *Bacillus subtilis* Competence Genes *com*B and *com*A, One of Wich Is Related to a Family of Bacterial Regulatory Determinants" *J.Bacteriol.* **171**:5362-5375

77. Wilbur y Lipman 1983 "Rapid similarity searches of nucleic aceid and protein data banks" *Proc. Natl. Acad. Sci.* 80:726-730

78. Winans, S; Eberty, P; Stachel, S; Gordon, M; y Nester, E; 1986 "A gene essential for *Agrobacterium* virulence is homologous to a family of positive regulatory loci" *Proc. Natl.Acad.Sci.USA* **83**:8278-8282

79. Woese, C; 1987 "Bacterial Evolution" *Microbiological Reviews* **51**:221-271

80. Yoshikawa, H; Kazami, J; Yamashita, S; Chibazakura, T; Sone, H; Kawamura, F; Oda, M; Isaka, M; Kovayashi, Y; y Saito, H; 1985 "Revised assignment for the *Bacillus subtilis spo0*F gene and its homology with *spo0*A and with two *Escherichia coli* genes" *Nucleic Acids Research* **14**:1063-1072

81. Young, J. P. W. 1992 "Phylogenetic classification of nitrogen-fixing organisms, p.43-86. In G. Stacy, R. Burris, and H. Evans (ed.), Biological nitrogen fixation. Chapman Hall, New York.

82. Zuckerkandl, E; y Pauling, L; 1962 "Molecular disease, evolution, and genetic heterogeneity. In M. Kasha and B. Pullman, eds; *Horizons in Biochemistry*, pp. 189-225. New York: Academic Press.

83. Zuckerkandl, E; y Pauling, L; 1965 "Molecules as documents of evolutionary history" *J. Theor. Biol.* **8**:357-366