UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Unidad Académica de los Ciclos Profesional
y de Posgrado del C.C.H.

ESTIMACION FILOGENETICA DE TRANSFERENCIA HORIZONTAL DE GENES EN BACTERIAS.

T E S I S

Que para obtener el título de

LICENCIADO EN INVESTIGACION BIOMEDICA BASICA

presenta

ANA MARIA VALDES FERNANDEZ

México, D.F. 1990.

A mis hermanas Marisa y María Teresa.
A mis padres.

Ash on an old man's sleeve
is all the ashes burnt roses leave
Dust in the air suspended
marks the place where a story ended.

T.S. Eliot

Para Jorge Sobrón
con mucho cariño
Ana María

AGRADECIMIENTOS.

Quiero agradecer muy especialmente al Dr. Daniel Piñero por su apoyo, su colaboración, su confianza y su asesoría en el desarrollo de esta tesis.

A Luis Eguiarte y Jorge Soberón por su paciencia, sus opiniones, por hacerme ver la Biología de otra forma y por su ayuda en todo momento.

A Exequiel Ezcurra por su ayudarme con la estadística y con algunos problemas computacionales.

A La Morsa *alias* Manuel López por echarme la mano con los programas, por ayudarme a imprimir esta tesis y, sobretodo, por su increible sentido del humor.

A Lorenzo Segovia por sus comentarios y críticas a este trabajo y por su apoyo durante la licenciatura.

A Roberto Hernández por darme la oportunidad de colaborar con él, por su amistad y su apoyo.

A Carmen Gómez por sus críticas y comentarios.

A Néstor Arvizu por su ayuda en toda clase de trámites.

A Georges Dreyfus por toda la paciencia que me tuvo y por enseñarme las cosas más elementales.

A mis amigas:

las del 322 Nora Vázquez, Mónica Montero y Silvia Deva por su ayuda en el laboratorio.

las del Centro de Ecología Gaby Jiménez, Pola Parlange y Ellen Gryj, por ser a todo dar.

las de la carrera Marina Chicurel, Magda Plebanski, Claudia González y Laura Velázquez, por las discusiones, por el apoyo y por ser compañeras del mismo dolor.

A Ana Mendoza, con todo mi cariño.

RESUMEN

La existencia de transferencia horizontal de genes entre especies o poblaciones bacterianas puede probarse comparando dos árboles filogenéticos derivados de distintos genes. En este trabajo se desarrolló, por medio de una simulación en computadora, un método cuantitativo para estimar el número de eventos de transferencia ocurridos en un grupo bacteriano usando el principio de congruencia filogenética. Se estudió el efecto de la transferencia de genes sobre la diferencia topológica entre dos árboles filogenéticos, uno derivado de un gen cromosomal -secuencia no transferible - y otro, de un gen plasmídico -secuencia transferible. Se encontró que la diferencia topológica entre dos árboles de este tipo es una función del número de eventos de transferencia de genes ocurridos en las poblaciones o especies analizadas. Los resultados pueden explicarse de manera satisfactoria usando un modelo logístico y el comportamiento es esencialmente el mismo bajo diferentes condiciones: tasas de sustitución variables entre linajes, diferente longitud de las ramas, diferente número de OTUs (unidades taxonómicas operacionales) en el árbol y diferentes topologías.

El método desarrollado se probó en datos empíricos tomados de la literatura (Young y Wexler, 1988) y se estimó la cantidad de transferencia de genes ocurrida entre plásmidos Sym en dos poblaciones agrícolas de *Rhizobium leguminosarum* biovar *viceae*. Al comparar árboles filogenéticos derivados del plásmido y del cromosoma, tomando en cuenta el error de reconstrucción filogenética involucrado, se encontró que entre el 20 y el 40% de todos los tipos genéticos involucrados, han estado implicados en eventos de transferencia horizontal de genes. Este valor disminuye hasta 1.5% cuando sólo se comparan árboles derivados de distintas regiones del plásmido.

INTRODUCCION

A nivel evolutivo, la consecuencia más importante de la reproducción sexual es la redistribución de genes (Felsenstein 1988). Este proceso implica que los genomas se mezclen y recombinen para dar lugar a nuevas combinaciones genéticas. Si bien este no es el lugar para discutir las ventajas adaptativas de la reproducción sexual, es decir, de la recombinación, es importante hacer notar que este fenómeno, como se discute en la primera sección, condiciona de que forma puede actuar la selección natural sobre una población.

En organismos procariontes no hay propiamente reproducción sexual. Sin embargo, si hay transferencia horizontal de genes mediada por elementos extracromosomales y existe toda una maquinaria molecular de recombinación (Low y Porter 1978).

La transferencia horizontal de genes es una forma de flujo génico y, por lo tanto, determina hasta que grado son independientes los cambios genéticos en diferentes poblaciones. El efecto evolutivo de la transferencia de genes va a depender de varios factores:

1) La estructura genética de las poblaciones en cuestión.

2) Las características adaptativas para las cuales codifiquen los genes intercambiados, es decir, los coeficientes de selección involucrados.

3) La magnitud y el alcance del proceso; es decir, si es muy frecuente y si se da entre linajes muy distintos como podrían ser organismos pertenecientes a diferentes géneros o familias.

Por otra parte, si hay intercambio genético y recombinación entre distintas especies o poblaciones, los patrones filogenéticos derivados de genes homólogos no van a representar las rutas de divergencia de dichos linajes. Sin embargo, los patrones filogenéticos pueden ser utilizados para estimar la magnitud de la transferencia de genes entre linajes.

El objetivo de esta tesis ha sido desarrollar un método cuantitativo para estimar a partir de la comparación de árboles filogenéticos la magnitud de la transferencia horizontal de genes entre poblaciones o especies bacterianas.

I. Estructura genética de poblaciones bacterianas.

i) Propiedades del desequilibrio por ligamiento en presencia de selección. (Felsenstein, 1988)

Imaginemos una población haploide con dos loci *A* y *B*, y que cada locus presenta 2 alelos, a_1 y a_2, y, b_1 y b_2 respectivamente. Supongamos que la población está en equilibrio por ligamiento, lo que quiere decir que no hay asociación entre la presencia de cualquier alelo en el locus *A* y la presencia de cualquier alelo en el locus *B*. De esta forma, la fracción de alelos b_1 entre individuos que son a_1 es la misma que la fracción de genotipos b_1 en toda la población. Si $P(a_1)$ es la fracción de todos los alelos a1 en la población y $P(b_1)$ de los alelos b_1 entonces

$$P(a_1b_1) = P(a_1)P(b_1)$$

Es decir, si la selección natural actúa sobre el locus *B* no va a modificar las frecuencias génicas en el locus *A*. Si en una generación hay selección por viabilidad primero en el locus *A* y luego en el locus *B*, el efecto es el mismo que si el regimen de selección impuesto de acuerdo a la adecuación de cada genotipo fuese el producto de la adecuación en el locus *A* por la adecuación en el locus *B*. Cuando la selección es multiplicativa, el desequilibrio por ligamiento no aparecera si no existia antes. Si las adecuaciones de a_1 y a_2 son, respectivamente, 1.0 y 0.9, y las adecuaciones de b_1 y b_2 son 1.0 y 0.8, la adecuación del genotipo a_2b_2 será de 0.9 x 0.8 = 0.72 y tendrá exactamente el mismo efecto que si ocurrieran dos eventos de selección sucesivos, uno en cada locus. De hecho, el efecto de la recombinación es disminuir el desequilibrio por ligamiento hasta cero creando gametos que contienen genes asociados al azar a partir de diferentes gametos en la generación previa.

Si hay desequilibrio por ligamiento esto afecta la tasa de respuesta en cada locus a la selección natural. Si en una población los alelos a_1 y b_1 están asociados de manera no aleatoria, lo mismo se aplica a los alelos a_2 y b_2. Debido a esta asociación, la selección natural, al eliminar a los alelos a_2 elimina al mismo tiempo a los alelos b_2, por lo que la selección sobre un locus cambia las frecuencias génicas en el otro.

Si los alelos favorecidos son a_1 y b_1 y estan asociados (desequilibrio por acoplamiento), la selección sobre cada uno acelera el cambio en las frecuencias génicas del otro. El desequilibrio por ligamiento acoplado aumenta la tasa a la que una población responde a la selección natural. El caso contrario, la asociación entre a_1 y b_2, (y por lo tanto, entre a_2 y b_1) provoca que la selección sobre ambos loci entre en conflicto (desequilibrio por repulsión). Desde luego, los términos "acoplamiento" y "repulsión" requieren que se especifique que alelos en cada locus son los que incrementan la adecuación. La manera más sencilla de visualizar el desequilibrio por ligamiento es imaginar al principio dos loci con dos alelos cada uno tal que las adecuaciones de a_1b_1, a_1b_2, a_2b_1 y a_2b_2 esten en proporciones $(1+s)^2$: $1+s$: $1+s$: 1. En este caso el alelo con subíndice '1' en cada locus está favorecido por un coeficiente de selección *s* y los dos loci no interactúan. Si las frecuencias alélicas fueran idénticas en ambos

loci, el caso más extremo de desequilibrio por acoplamiento sería que la población consistiera únicamente de genotipos a_1b_1 y a_2b_2. Sus adecuaciones estarían en proporción $(1+s)^2 : 1$. El desequilibrio por acoplamiento implica que los individuos con el alelo a_1 y aquellos con el alelo a_2 difieren en adecuación más de $(1+s):1$ y esto va a acelerar el cambio de frecuencias en el locus A. En el caso extremo de repulsión en que unicamente existen individuos a_1b_2 y a_2b_1 en la población, las adecuaciones de ambos genotipos son idénticas ($1+s$ en ambos casos) por lo que la selección favoreciendo al alelo a_1 se ve contrabalanceada por la selección que favorece al alelo b_1.

De todo lo anterior se concluye que, para que podamos observar algun efecto en la modificación de las frecuencias alélicas o en los equilibrios genéticos de una población debido a recombinación deben existir adecuaciones no multiplicativas, que es a final de cuentas lo que provoca el desequilibrio por ligamiento.

ii) Selección periódica en poblaciones bacterianas.

La selección periódica es, en realidad, un tipo de "cuello de botella" y se refiere a una de las principales consecuencias de la selección natural al actuar sobre una población de estructura esencialmente clonal. Es decir, puesto que en organismos asexuales existe un ligamiento practicamente absoluto entre genes, las frecuencias de determinados alelos para un cierto número de marcadores geneticos, aumentan abruptamente como resultado de la selección que opera sobre cualquier otro locus en el genoma del organismo (Hartl y Dykhuizen , 1984). Aunque el tamaño real de la población sea bastante grande, las clonas principales son reemplazadas periódicamente por clonas de mayor adecuación, por lo que el tamaño efectivo de la población disminuye considerablemente. Las clonas principales se originan como uno, o unos cuantos individuos, que se han derivado por mutación a partir de una clona ya existente. El resultado es una población mucho más monomórfica de lo que se esperaría si la población no fuera purgada de esta forma. Dado que no se requiere una contracción en el tamaño total de la población, las clonas de mayor adecuación podrían, durante el proceso de fijación, intercambiar genes con las clonas a las que van a reemplazar. Como resultado del intercambio genético y la recombinación, el efecto asociado de selección sobre adecuaciones no multiplicativas característico de la selección periodica se vería muy reducido (Levin, 1981). Las combinaciones de genes podrían romperse y las clonas dejarían de funcionar de manera individual y de evolucionar separadamente (Selander y Levin, 1980).

En la actualidad es ampliamente aceptado el hecho de que las poblaciones bacterianas sufren "selección periódica" (Selander y Levin, 1980; Selander et al., 1985; Milkman y Crawford, 1983). Muchos estudios indican que la estructura de las poblaciones bacterianas es fundamentalmente clonal con muy poco intercambio de genes cromosomales entre individuos mediado por recombinación. Las frecuencias observadas de los genotipos multilocus se desvian de

manera muy significativa de una asociación aleatoria de alelos, y la magnitud de los coeficientes de desequilibrio al comparar los loci por pares, no está aparentemente relacionada con la distancia física entre los loci. Más aún, al considerar clonas de *Escherichia coli* aisladas de un solo individuo, los genotipos observados no muestran evidencia de recombinación frecuente. (Selander y Levin, 1980; Caugant et al. 1981).

Sin embargo, ¿qué ocurriría si las distintas clonas de una población bacteriana intercambiaran material genético contínuamente? A nivel de la población el resultado inmediato sería un mucho mayor tamaño efectivo y por lo tanto una menor capacidad de responder al efecto de la deriva génica. Para que exista intercambio genético en bacterias, éste debe de ser mediado por elementos extracromosomales que por recombinación adquieran segmentos de DNA cromosomal, se transfieran a otro organismo, y nuevamente recombinen intercambiando el material genético del organismo donador. Es decir, la tasa de transferencia de elementos extracromosomales es la primera condición para que se rompa la estructura clonal de una población bacteriana.

Sin embargo, la tasa total de recombinación (intercambio cromosómico) debe ser una fracción de la tasa total de transferencia de elementos extracromosomales y por lo tanto es de gran importancia obtener estimadores confiables de esta última.

II. TRANSFERENCIA HORIZONTAL DE GENES EN BACTERIAS.

En 1955 Joshua Lederberg publicó un trabajo titulado "Recombinatión Mechanisms in Bacteria" (Lederberg, 1955). En este artículo Lederberg usó el término 'recombinación' para referirse a los procesos combinados de transferencia de DNA de una célula a otra y al subsecuente establecimiento de alguna de la información genética de la célula donadora en la célula receptora. La recombinación en bacterias puede verse desde varias perspectivas, pero en muchos casos debe ser analizada en relación a un sistema en particular de transferencia de genes con el cual esté asociado, como Lederberg implicó en su definición original de recombinación (Low y Porter 1978).

La transferencia de genes presenta ciertas características únicas en poblaciones bacterianas puesto que es mediado por elementos extracromosomales tales como plásmidos, virus y elementos transponibles. Estos elementos son muy comunes y diversos en la naturaleza (Eberhardt, 1989).

Recombinación y transferencia de genes en bacterias.

i) Elementos extracromosomales en bacterias

Las poblaciones bacterianas pueden ser parasitadas por una variedad de elementos genéticos independientes (Levin y Lenski, 1983). En esta discusión me referire a elementos extracromosomales (aún si están insertados en el cromosoma bacteriano) como a todos

aquellos elementos genéticos presentes en poblaciones bacterianas con las siguientes características: (a) dichos elementos generalmente no incluyen genes que sean necesarios de manera obligatoria para la reproducción del organismo que los acarrea; (b) son capaces ya sea de replicarse de manera autónoma o de sobre replicar su propio DNA en relación al DNA cromosomal típico de la célula (Campbell, 1981). Existen fundamentalmente 3 tipos de elementos extracromosomales a los que me referiré a continuación:

1.- Bacteriofagos o fagos son virus parásitos de las bacterias. Cuando se encuentran fuera de la célula, consisten de un genoma (DNA por lo general, pero a veces puede ser RNA) rodeado y protegido por una cápside protéica que ademas ayuda a introducir al virus dentro del huésped. Una vez dentro de la célula, el DNA del fago redirige el metabolismo del huésped hacia la síntesis de nuevas partículas víricas, las cuales son liberadas ocasionando la muerte de la célula huésped . A este proceso se le conoce como ciclo lítico . Los fagos virulentos únicamente presentan el ciclo lítico, mientras que los fagos temperados pueden insertarse en el cromosoma del huésped como un profago. Este luego se replica dentro del huésped sin que se produzcan nuevas partículas virales, y la célula huésped no es dañada. Una bacteria que acarrea un profago se dice que es lisogénica. Esta célula es inmune a nuevas infecciones por el mismo tipo de fago. Ocasionalmente el profago es inducido a entrar en fase lítica. (Maynard Smith, 1989)

2.- Plásmidos : son moléculas de DNA circular (que al menos en enterobacterias son físicamente independientes del cromosoma). Difieren de los fagos en que no presentan un estadio extracelular, y por lo tanto no codifican para proteínas de cápside. En la actualidad se conocen miles de plásmidos (Willets, 1985). El fenotipo de un plásmido no es un criterio confiable de clasificación porque, por ejemplo, puede estar determinado por un elemento transponible a mucho tipos de plásmidos. En cambio, la clasificación se basa en dos propiedades claves de los plásmidos: replicación y conjugación. Un plásmido conjugativo es aquel que provoca un contacto entre la célula huésped y otra células (receptoras), usualmente mediante la producción de pilum conjugativo que se extiende desde la pared celular de la célula conjugativa. La conjugación permite que un plásmido pase de una célula a otra. Algunos plásmidos son no conjugativos pero son movilizables por lo que pueden pasar a otra célula si la conjugación es causada por otro plásmido. A un plásmido que es tanto conjugativo como movilizable se le llama autotransferible (Maynard Smith 1989).

3.- Elementos transponibles o transposones: son pedazos de DNA que pueden transponerse de un sitio en un cromosoma o plásmido a otro. Cuando una copia de un transposón se inserta en un nuevo sitio, la copia original generalmente permanece en el antiguo sitio; la transposición es una forma de replicación. En segundo lugar, aunque la transposición involucra el rompimiento y religamiento del DNA, no requiere una secuencia homóloga entre el transposón y el cromosoma o plásmido: es decir, no es una recombinación homóloga. Los

transposones son generalmente secuencias de varias kilobases y frecuentemente acarrean genes de resistencia a antibióticos. Las secuencias de inserción son elementos transponibles más pequeños de aproximadamente unas 1000 bases (Maynard-Smith 1989).

La resistencia a antibióticos y a metales pesados; la capacidad de producir enzimas de restricción, toxinas, bactericinas y antibióticos; la habilidad de fermentar ciertas fuentes de carbono y la producción de estructuras para invadir habitats específicos son características codificadas por genes acarreados en plásmidos o en elementos transponibles. Aunque la abundancia y la diversidad de fenotipos bacterianos determinados por fagos es menor que la de fenotipos determinados por plásmidos o por elementos transponibles, hay algunos casos en los que la resistencia a antibióticos y la producción de toxinas se deben a genes acarreados en fagos.

ii) Papel evolutivo de la transferencia horizontal de genes

Los plásmidos conjugativos, los fagos y los elementos transponibles asímismo desempeñan un papel importante en la evolución y adaptación bacterianas al funcionar como vehículos de intercambio de material genético. En el curso de una transmisión infecciosa estos replicones pueden "tomar" genes cromosomales de una bacteria y transmitirla a otra. Puesto que el rango de huéspedes de los plásmidos y virus bacterianos generalmente sobrepasa las barreras de "especies" y dada la existencia de mecanismos de recombinación en ausencia de homología, el rango de intercambio genéticos mediado por elementos extracromosomales puede abarcar grupos de bacterias muy diversos filogenéticamente. (Levin y Lenski 1985).

Muchos autores han discutido sobre la trascendencia de la transferencia horizontal de genes en bacterias como una estrategia adaptativa (Reanny, 1978, Campbell, 1981, Levin y Lenski, 1985,; Evans, 1986; Eberhardt, 1989). Como Woese (1987) ha señalado, "en la situación extrema, los intercambios interespecíficos de genes serían tan comunes que una bacteria no tendría en realidad una historia evolutiva por sí misma; sería una quimera evolutiva".

El flujo génico, ya sea intra o interespecífico, determina hasta que grado los cambios genéticos en diferentes poblaciones son independientes (Slatkin 1985). Si bien puede inhibir la evolución a nivel genético al impedir que la selección natural y la deriva génica establezcan y mantengan diferencias genéticas locales, también puede desempeñar una función creativa al permitir que las poblaciones se muevan de un pico adaptativo a otro (Slatkin 1987). El flujo génico, como ya se ha mencionado, al ser mediado por elementos extracromosomales presenta características muy interesantes en poblaciones bacterianas.

Los genes bacterianos se pueden mover entre elementos extracromosomales, o entre estos y el cromosoma. Lo último es particularmente cierto para los transposones, que son capaces de

insertarse tanto en plásmidos como en el cromosoma. Los genes también pueden moverse de una cepa, especie o género bacteriano a otro y es este proceso lo que se conoce como transferencia horizontal de genes. En este trabajo me referiré específicamente a este tipo particular de flujo génico. Si este proceso fuera muy común entre grupos de orden taxonómico alto tales como especies, géneros o familias alteraría la tasa a la que evolucionan los genes homólogos de los linajes involucrados.

La transferencia de genes es, entonces, una forma de flujo génico mediada por elementos extracromosomales los cuales normalmente codifican para características altamente adaptativas bajo presiones de selección. Dado que puede afectar en gran medida la dinámica de las poblaciones bacterianas, la transferencia de genes puede resultar un mecanismo evolutivo muy importante en poblaciones procariontes. De igual manera, puede tener consecuencias inesperadas si cepas modificadas por ingeniería genética son liberadas en ambientes naturales.

III. ARBOLES FILOGENETICOS

1) TASAS DE EVOLUCION

Fenotípicas

La evolución a nivel de forma y función es deteminada por la selección natural darwiniana, i.e. independientemente de los mecanismos genéticos subyacentes, se van a seleccionar fenotipos con una alta adecuación relativa. Si analizamos cuidadosamente el registro fósil se hace evidente que la aparición de un tipo estructural completamente nuevo es seguida por un período de evolución "explosiva" y luego, fases de cambio muy lento pueden conservar ciertos tipos morfológicos por largos periodos de tiempo (Kurten 1959). En conclusión la evolución a nivel de forma y función no obedece un comportamiento constante con respecto al tiempo y por lo tanto la tarea de la taxonomía clásica de derivar las relaciones evolutivas entre especies cuantificando diferencias morfológicas se torna muy complicada (Kimura 1983).

Tasas de evolución a nivel molecular

En las dos últimas décadas se ha acumulado una enorme cantidad de datos de secuencias aminoácidas y nucleotídicas para un gran número de proteínas y de genes. Kimura (1983) analizó las secuencias de la cadena alfa de hemoglobina de humano, perro, canguro, equidna, pollo, salamandra acuática, carpa y tiburón, y comparó el porcentaje de diferencias aminoácidas entre pares de secuencias con el tiempo estimado de divergencia para estos organismos a partir del registro fósil. Kimura observó que el porcentaje de diferencia entre secuencias es proporcional al tiempo. El paralelismo es aún más evidente si, en lugar de utilizar el número de diferencias, se utiliza el número de sustituciones. Es decir, puesto que puede haber ocurrido más de un cambio en un sitio determinado, el número de

sustituciones puede ser mayor al número de diferencias observadas. Para corregir por las mutaciones superimpuestas es muy útil suponer que el proceso de sustitución aminoácida se comporta de acuerdo a la ley de Poisson. Sea p_d = num de diferencias/ num total de aminoácidos, y K_{aa} el promedio de sustituciones por sitio entre dos polipéptidos. Supóngase que las probabilidades de que ocurran 0,1,2,... sustituciones aminoácidas en un sitio en particular están dadas por la serie de Poisson:

$$p(i) = e^{-x}x^i/i!$$

donde $p(i)$ es la frecuencia relativa esperada de que ocurran i eventos y x es la probabilidad asociada, por lo tanto:

$$e^{-Kaa} + K_{aa}e^{-Kaa} + K_{aa}2e^{-Kaa}/2! + ... \text{ etc.}$$

Al utilizar un proceso de Poisson estamos suponiendo que los eventos de sustitución a lo largo de la secuencia son independientes y con probabilidades iguales y que la existencia de una sustitución aminoácida por sitio es un evento muy raro para cualquier período de tiempo, pero al extendernos sobre un período enorme, la probabilidad se vuelve apreciable. En concreto, la probabilidad de que dos sitios sean idénticos es e^{-Kaa}. Si igualamos esto a la fracción de sitios idénticos de una secuencia dada tendremos

$$e^{-Kaa} = 1 - p_d$$

Sacando logaritmo natural en ambos lados tenemos:

$$K_{aa} = -\ln(1 - p_d);$$

Dado que K_{aa} es el número de sustituciones aminoácidas ocurridas entre dos linajes actuales que divirgieron hace T unidades de tiempo (años o generaciones), es decir, el número de sustituciones ocurridas en $2T$ unidades de tiempo, la tasa de sustitución por unidad de tiempo es entonces:

$$k_{aa} = K_{aa}/2T$$

En la figura 1 el número estimado de sustituciones aminoácidas (K_{aa}) al ser graficado contra el tiempo de divergencia (puntos llenos) presenta un comportamiento claramente lineal cuya pendiente es $2K_{aa}$.

Tasas de sustitución nucleotídicas

El método más sencillo para estimar el número de sustituciones nucleotídicas es la fórmula de Jukes y Cantor (1969). Supongamos que las sustituciones nucleotídicas ocurren en cualquier sitio nucleotídico con igual probabilidad, y que en un sitio determinado, un nucleótido cambia a cualquiera de los otros tres tipos con una tasa ß por año. Consideremos dos secuencias (X y Y) que divergieron de una secuencia ancestral común hace t años. Denotamos por q_t la

Figura 1

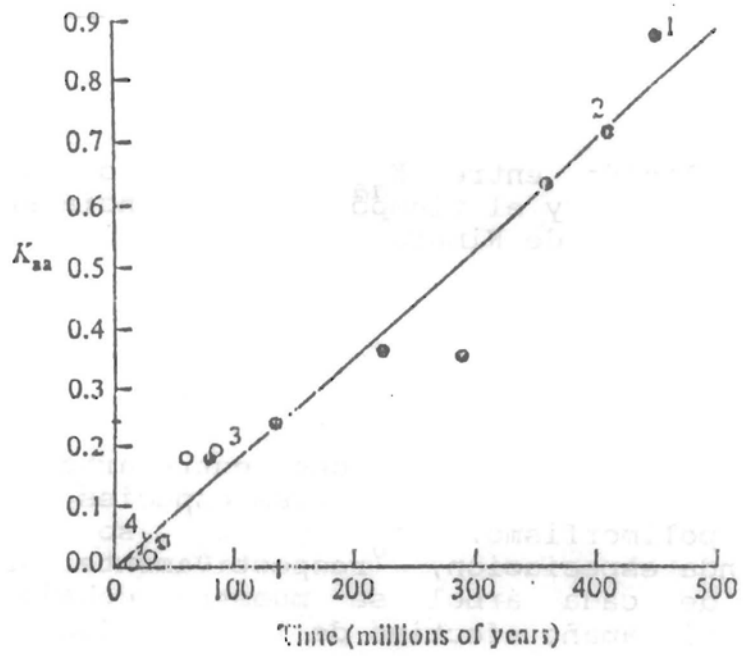


Figura 2

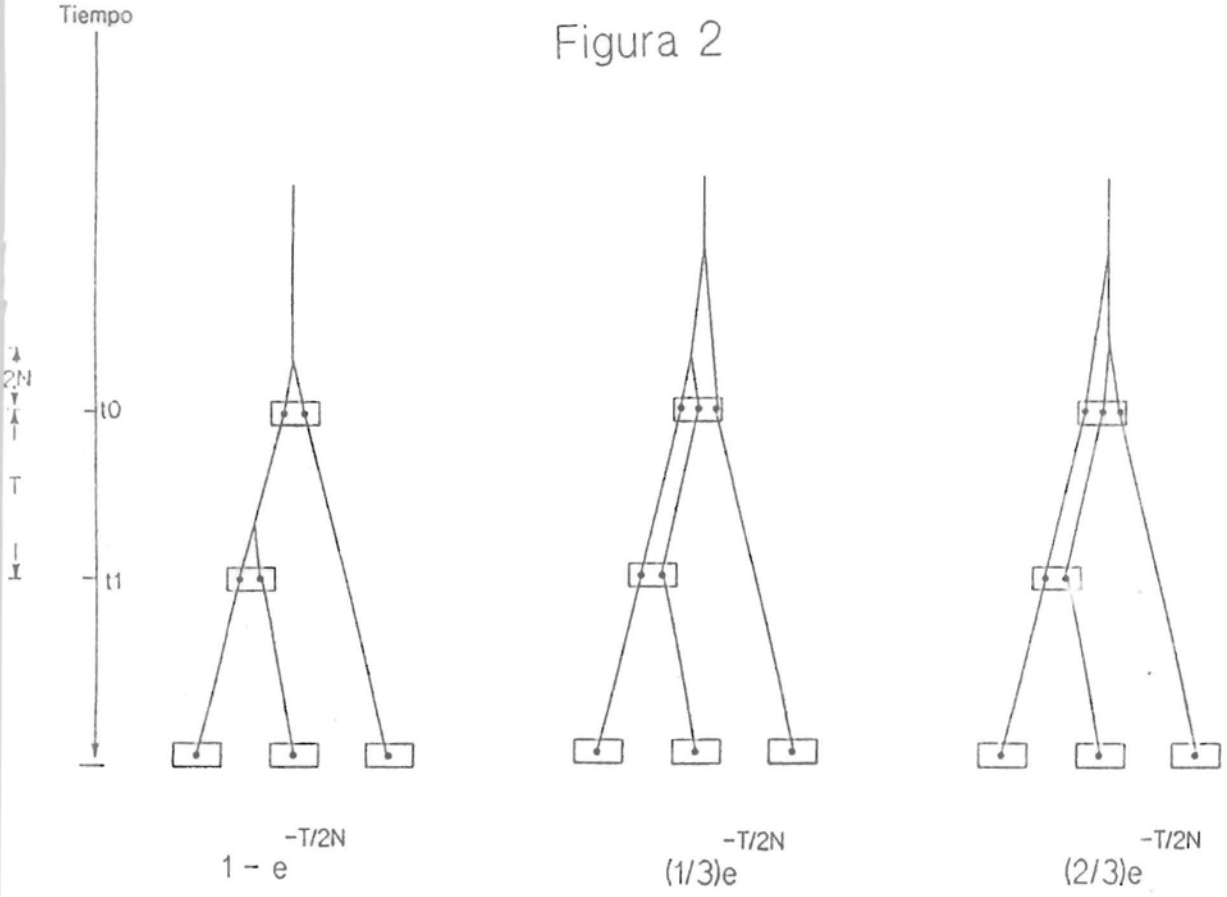


Figura 1. Relación entre K_{aa}, el número de sustituciones aminoácidas (ordenada) y el tiempo de divergencia en millones de años (abscisa). Tomado de Kimura, 1983.

Figura 2. Tres posibles relaciones entre árboles de especies y árboles de genes para el caso de tres especies (X,Y y Z) en presencia de polimorfismo. t_0 y t_1 son los tiempos de la primera y segunda especiación, respectivamente. La probabilidad de ocurrencia de cada árbol se muestra debajo del mismo. $T = t_1 - t_0$ y N es el tamaño efectivo de la población. Tomado de Nei, 1987.

proporción de nucleótidos idénticos entre X y Y, y $p_t = 1 - q_t$, la proporción de nucleótidos diferentes. En primer lugar, un sitio que fuera idéntico para X y Y al tiempo *t* permanecerá igual al tiempo *t+1* con probabilidad 1- 2ß. En segundo lugar, un sitio que tenga nucleótidos diferentes al tiempo *t* en X y Y, tendrá el mismo nucleótido al tiempo *t+1* con probabilidad 2ß/3 (en esta derivación los terminos al cuadrado o de orden superior han sido descartados).

Por lo tanto:

$$q_{t+1} = (1- 2ß)q_t + 2ß/3 \cdot (1 - q_t)$$

donde $ß = ß'dt$

Sea $q_{t+1} - q_t = dq_t/dt = (2ß - 8ßq_t)/3$;

Si partimos de las condiciones iniciales q=1, t=0 la solución de la ecuación diferencial es

$$q = 1 - (3/4)(1 - \exp(-8ßt/3))$$

es decir

$$ß = (-3/8t)[\ \ln\ (1 -4p/3)]$$

Puesto que el número esperado de sustituciones nucleotídicas por sitio (*d*) está dado por 2ßt, *d* se puede estimar de la siguiente manera:

$$\hat{d}=2t\ [(-3/8t)[\ \ln\ (1- 4p/3)]$$

$$\hat{d}= -(3/4)\ln(1 - (4p/3))$$

donde p= 1 - q es la proporción de nucleotidos diferentes entre X y Y.

Además del método de Jukes y Cantor existen otros métodos como el método de dos parámetros de Kimura (1980) que toma en cuenta que las tasas de transversiones y transiciones no sean iguales; o el método de Tajima y Nei (1984) que toma en consideración que las tasas de sustitución sean variables para diferentes posiciones de la secuencia.

Una vez que se tienen los elementos para decir que tan distinta es una secuencia de otra, se pueden construir relaciones filogenéticas entre organismos.

TIPOS DE ARBOLES FILOGENETICOS

ARBOLES DE ESPECIES Y ARBOLES DE GENES

Los evolucionistas generalmente están interesados en obtener un árbol filogenético de especies o poblaciones. Sin embargo, si obtenemos un árbol filogenético para un grupo determinado de

organismos con la información de una sola proteína o un solo gen, lo que en realidad estamos obteniendo es un árbol de genes el cual puede o no ser igual al árbol de especies. Hay muchos árboles de genes dentro de cualquier árbol de poblaciones o especies y, más aún, el árbol de poblaciones puede en cierto sentido representar una compilación de genealogías para varios genes. Sin embargo, la topología para un determinado árbol de genes puede diferir de la del árbol de poblaciones o especies debido a:

a) Errores de muestreo atribuibles a un número pequeño de nucleótidos o aminoácidos examinados (Saitou y Nei 1986)

b) Cuando el gen estudiado pertenece a una familia multigénica surge otro problema; no es fácil identificar los genes homólogos para todas las especies estudiadas. Cuando el número de genes duplicados es pequeño, la identificación es relativamente sencilla, pero cuando el número es mayor, como en el caso de las regiones variables de las inmunoglobulinas, resulta virtualmente imposible. (Nei, 1987)

c) Heterogeneidad en las tasas de evolución entre linajes . Vamos a suponer, en el caso más simple, que los genes de cada especie se separaron al mismo tiempo que las especies. Supongamos además que los patrones de separación en la historia evolutiva de los genes concuerdan con los patrones de separación de las especies bajo consideración. Aún así puede darse el caso de que los patrones de ramificación o topologías de los árboles filogenéticos, uno el de las especies, que en el caso ideal se deriva de todo el genoma, y otro, el de un gen particular, no coincidan. Esto se debe a que las sustituciones nucleotídicas o aminoácidas ocurren al azar (procesos estocásticos) por lo que el número de sustituciones en el linaje (organismo) Z puede ser menor que en los linajes X y Y. Para evitar este tipo de error es muy importante analizar un gran número de nucleótidos o aminoácidos.

d) Si existió polimorfismo en el momento de la separación de los genes, los tiempos de divergencia de los genes muestreados para diferentes especies serán mayores que los tiempos de divergencia de las especies. Más aún, la topología del árbol construido para un gen en particular puede ser diferente de la topología para el árbol de las especies a causa del polimorfismo ancestral (Fig. 2; Tajima, 1983; Neigel and Avise, 1986; Pamilo y Nei, 1988).

e) Flujo génico. Si hay intercambio genético y recombinacion entre distintos linajes (especies o poblaciones), los patrones filogenéticos derivados de genes homólogos, no van a representar la historia "real" de los linajes.

Las últimas tres posibilidades no son simples generadores de "ruido" en la estimación filogenética. Al contrario, son fenómenos reales y una parte importante de la historia evolutiva de un grupo de linajes (Avise, 1989). Por lo tanto, los árboles de genes pueden utilizarse no sólo para estimar el árbol de las especies, sino para

estudiar los procesos evolutivos anteriormente mencionados.

ARBOLES CON RAIZ Y ARBOLES SIN RAIZ

Las relaciones filogenéticas entre genes o organismos generalmente se presentan en forma de un árbol con raíz. Sin embargo es posible dibujar árboles sin raíz o redes (Figura 3). El número de posibles árboles con raíz para un número n de especies es

$$(2n - 3)!/ 2^{n-2}(n-2)!.$$ (Nei, 1987)

O sea que para n=10 el número posible de árboles con raíz es 34'459,429. El número de árboles sin raíz viene dado por

$$(2n-5)!/2^{n-3}(n-3)!$$ (2'027,025 para n=10)

Es decir, a medida que el número de especies de un árbol aumenta se vuelve cada vez más difícil encontrar el árbol verdadero. Aunque es posible, a veces, obtener una topología sin raíz que con una alta probabilidad es la verdadera, localizar la raíz en el árbol puede ser una tarea muy difícil debido a los errores estándar asociados con los puntos de ramificación en una filogenia.

METODOS DE RECONSTRUCCION

Existen varios métodos distintos que se utilizan para reconstruir árboles filogenéticos a partir de datos moleculares. Los métodos más populares son los de matriz de distancias y los métodos de máxima parsimonia.

En los métodos de matriz de distancias se calcula una distancia genética (proporción de sustituciones aminoácidas o nucleotídicas) para todos los pares de especies o poblaciones considerados y se construye un árbol filogenético considerando las relaciones entre estos valores de distancia. Una vez obtenidos los valores de distancia hay varias formas de construir el árbol.

En los métodos de máxima parsimonia, la secuencia ancestral d nucleótidos o aminoácidos se infiere a partir de las especie actuales y el árbol se construye minimizando el número de cambios evolutivos para el árbol entero.

En general es muy difícil reconstruir el árbol que represente la historia evolutiva verdadera de las especies o poblaciones actuales. Hay dos tipos de errores en la reconstrucción filogenética: errores topológicos y errores en la longitud de las ramas.

Los errores topológicos se refieren a errores en el patrón de ramificación de un árbol o topología.

La longitud de las ramas, en el caso de un árbol de especies,

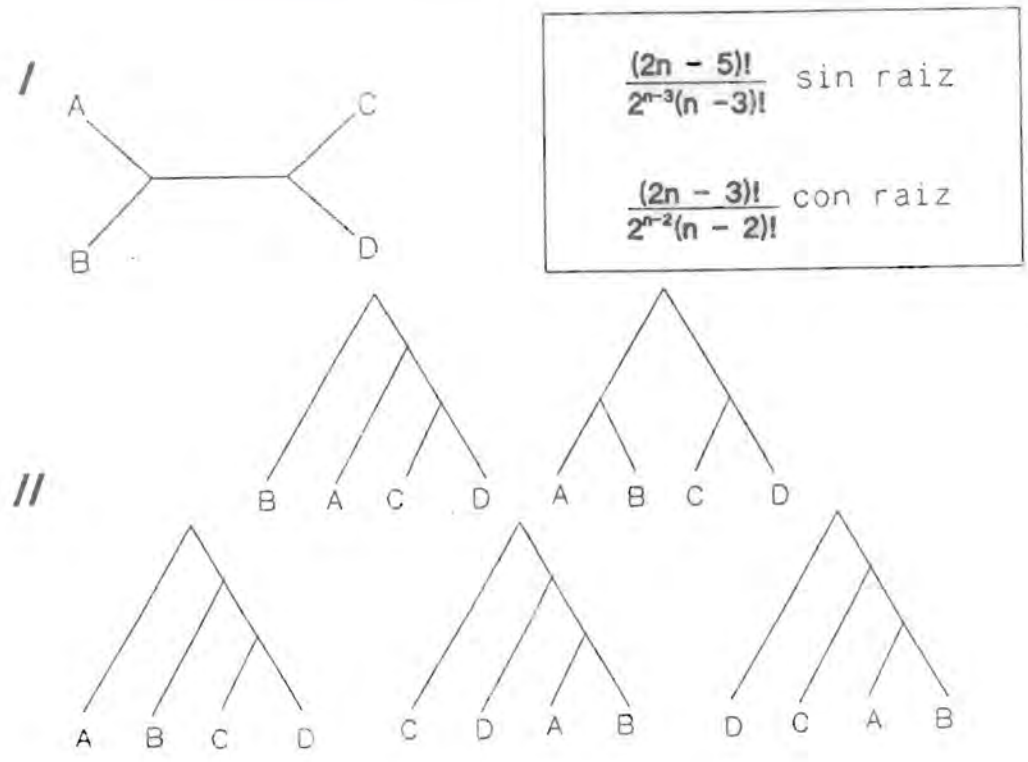


Figura 3. Cinco posibles árboles con raíz para un determinado árbol de 4 OTUs sin raíz. El número de posibles árboles de cada tipo se muestra en el recuadro, donde *n* es el número de OTUs en el árbol.

representa el intervalo de tiempo entre dos puntos de ramificación o entre un punto de ramificación (nodo) y una especie actual. En el caso de un árbol de genes, simplemente representa el número de sustituciones nucleotídicas ocurridas.

Obviamente sólo hay una topología verdadera para un grupo de organismos y la tarea de los evolucionistas es encontrar esa topología a partir de los datos existentes. En la práctica, no hay manera de saber si el árbol obtenido es el correcto. La única cosa que se puede hacer es usar un método confiable de reconstrucción filogenética. Desgraciadamente, la confiabilidad del método depende del tipo de datos usados y del propósito del investigador.

Métodos de matriz de distancias

Método de distancia promedio (UPGMA= unweighted pair group arithmetic average method)

Sea una matriz de distancias (figura 4). Se agrupan los dos organismos con la menor distancia entre ellos (La distancia para datos moleculares es la proporción de sustituciones nucleotídicas o aminoácidas). A este nuevo grupo se le calcula la distancia promedio con respecto a los organismos restantes y se vuelve a buscar la distancia más pequeña dentro de la matriz y se repite todo el proceso hasta que no queden más organismos en la matriz.
(figura 4)

Metodos de parsimonia

El principio de este método consiste en inferir la secuencia aminoácida o nucleotídica de la especie ancestral y escoger el árbol que requiera el mínimo número de cambios mutacionales. Al árbol obtenido por este método se le llama árbol de máxima parsimonia. Este método sirve sobre todo para obtener la topología, puesto que las longitudes de las ramas sólo se pueden obtener bajo ciertas suposiciones.

En primer lugar consideramos una topología en particular para un grupo de organismos e inferimos la secuencia ancestral para esta topología (figura 5).

Una vez que este número se obtiene se prueba otra topología y el número mínimo de sustituciones para esta nueva topología se determina. Este proceso se continúa para un número suficiente de topologías (el cual se determina de acuerdo a criterios estadísticos) y la topología que requiere el número mínimo de sustituciones es elegida como el árbol final.

IV. ESTIMACION DE TRANSFERENCIA DE GENES A PARTIR DE ARBOLES FILOGENETICOS.

Wilson y colaboradores (1977) han sugerido que la existencia de transferencia horizontal de genes puede ser evaluada comparando

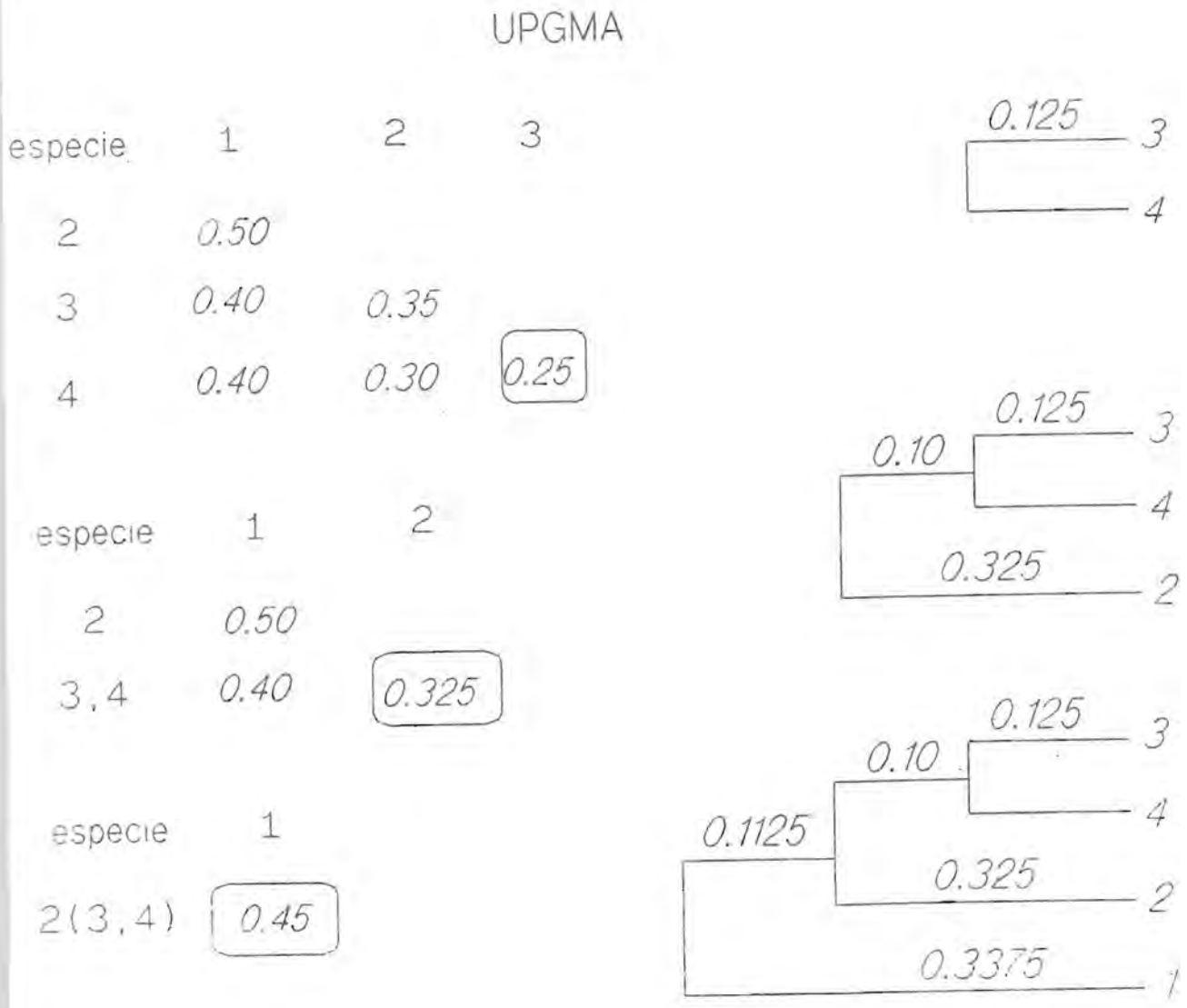


Figura 4. Ejemplo de reconstrucción filogenética por un método de matriz de distancias (UPGMA, ver texto) para cuatro especies.

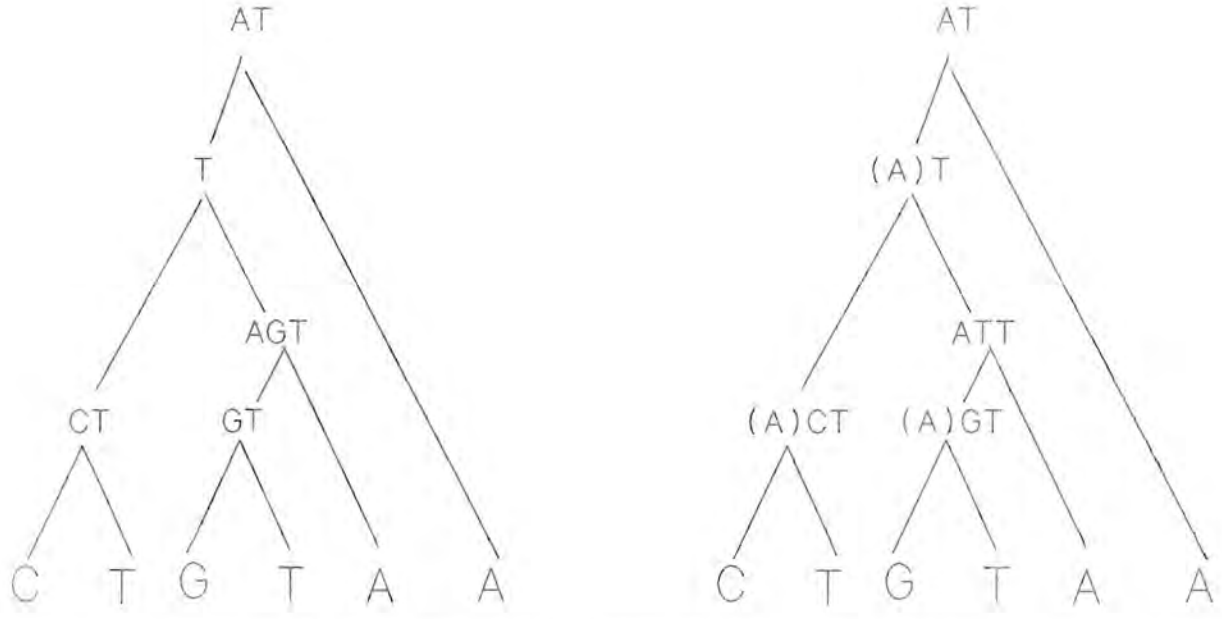


Figura 5. Nucleotidos en seis especies actuales y los posibles nucleotidos en cinco especies ancestrales (Fitch, 1971).

árboles filogenéticos estimados a partir de las secuencias de diferentes macromoléculas (por ejemplo 16S ribosomal y citocromo C). Si ambos árboles presentaran la misma topología, ello implicaría que ninguno de los dos genes ha estado involucrado en un proceso de transferencia interespecífica de genes. A esta prueba Wilson y colaboradores le llamaron "prueba de congruencia filogenética". Siguiendo esta línea, Woese y colaboradores (1980) compararon dos árboles filogenéticos obtenidos a partir de diferentes macromoléculas para ciertas especies de la subdivisión alfa de bacterias púrpuras y encontraron que ambos árboles presentaban prácticamente la misma topología.

El uso de enfoques filogenéticos puede, por lo tanto, resultar sumamente útil para esclarecer hasta que grado los mecanismos que rompen la clonalidad en poblaciones y especies bacterianas ocurren en la naturaleza. Aunque este enfoque puede resultar muy poderoso, ha sido utilizado muy poco debido, sobre todo, a la escasez de filogenias detalladas para muchos grupos de organismos.

Los enfoques filogenéticos han sido utilizados como evidencia para apoyar la existencia de transferencia intraespecífica de genes (Schoefield et al., 1987; Young y Wexler, 1988), para cuantificar recombinación intraespecífica en *E. coli* (DuBose et al. 1988) y para fechar eventos de duplicación génica (Beintema y Campagne, 1987; Nelson y Strobel, 1987).

Este tipo de mecanismos pueden ser estudiados, en principio, de manera cuantitativa comparando las topologías de diferentes árboles. Por ejemplo, si comparamos la filogenia de un grupo de taxa (poblaciones p. ej.) con su distribución geográfica se pueden derivar conclusiones relevantes a las frecuencias de modos alternaltivos de especiación (Lynch, 1989). Slatkin y Madison (1989) desarrollaron un metodo para estimar $4Nm$ (donde N= tamaño efectivo y m= tasa de migración) comparando filogenias de alelos con sus distribuciones geográficas. En este trabajo se ha desarrollado un método cuantitativo para estimar la magnitud de la transferencia horizontal de genes en un grupo bacteriano usando el principio de congruencia filogénetica. Este método se aplicó a datos tomados de la literatura (Young y Wexler 1988) y se estimó la magnitud de transferencia de genes ocurrida entre plásmidos Sym en poblaciones de dos campos agrícolas británicos de *Rhizobium leguminosarum* biovar *viceae*. El método propuesto es insensible a la longitud de las ramas en los árboles comparados por lo que es igualmente adecuado para estimar transferencia de genes entre poblaciones cercanas genéticamente o entre grupos más distantes tales como especies o géneros bacterianos.

METODOS

Este trabajo parte de la hipótesis de que la distancia topológica entre un árbol filogenético obtenido a partir de una secuencia cromosomal (no transferible) y otro, derivado de una secuencia plasmídica (transferible), debe ser función del número de eventos de transferencia genética ocurridos en la historia de un grupo de poblaciones o grupos de orden taxonómico superior. Con esta idea en mente, se simuló el proceso de transferencia horizontal de genes en un grupo bacteriano compuesto por 30 linajes diferentes. De estos 30, un número constante, *n*, de unidades taxonómicas operacionales (OTUs) se muestreaba al final de 500 generaciones. La idea era imitar un grupo bacteriano en el que un gran número de linajes son capaces de intercambiar material genético, pero sólo unos cuantos linajes son realmente muestreados. El proceso de simulación se ilustra gráficamente en la figura 6.

Se generó en una computadora un gene ancestral de 300 pares de bases usando números pseudoaleatorios y asumiendo que los cuatro tipos de nucleótidos se encontraban en proporciones iguales. Se diseñaron una serie de un árboles filogenéticos que sirvieran como modelo de la evolución sufrida por los linajes bacterianos en cuestión (figura 7 y figura 12). La secuencia ancestral se duplicaba en cada punto de ramificación y se sometía a una tasa de sustitución nucleotídica constante (M o L sustituciones esperadas por rama) usando el mismo tipo de simulación que Tateno et al. (1982), Tateno y Tajima (1986) y Nei y Sourdis (1988) . Las secuencias de los 30 - *n* linajes restantes se derivaron del gen ancestral y se sometieron a mutación esperando 2.5L sustituciones nucletídicas en cada una. Las *n* secuencias muestradas se compararon entonces por pares y se generó una matriz de diferencias nucleotídicas. Esta matriz fue corregida para sustituciones múltiples usando la fórmula de Jukes y Cantor (1969) y las distancias genéticas se usaron para la reconstrucción de árboles filogenéticos. Los árboles obtenidos de esta forma fueron considerados como árboles cromosomales.

En este trabajo los métodos de reconstruccion empleados fueron UPGMA (unweighted pair group arithmetic average method; Sokal y Michener, 1958), el método de Farris modificado (Modified Farris, MF, Tateno et al. 1982), y el método de unir vecinos (Neighbor joining, NJ, Saitou y Nei, 1987).

Las 30 secuencias nucleotídicas se duplicaron y las consideramos como el gen transferible (plásmido) acarreado por cada linaje. En este punto simulamos la transferencia horizontal entre bacterias suponiendo que cada célula de un determinado linaje tenia la misma probabilidad de actuar como receptora o como donadora de un gen transferible. La proporción de encuentros resultantes en transferencia de una copia del plásmido de una célula a otra que pertenecía a un linaje diferente estaba determinado por un parámetro *tr* que designa tanto la tasa de transferencia por unidad

Figura 6

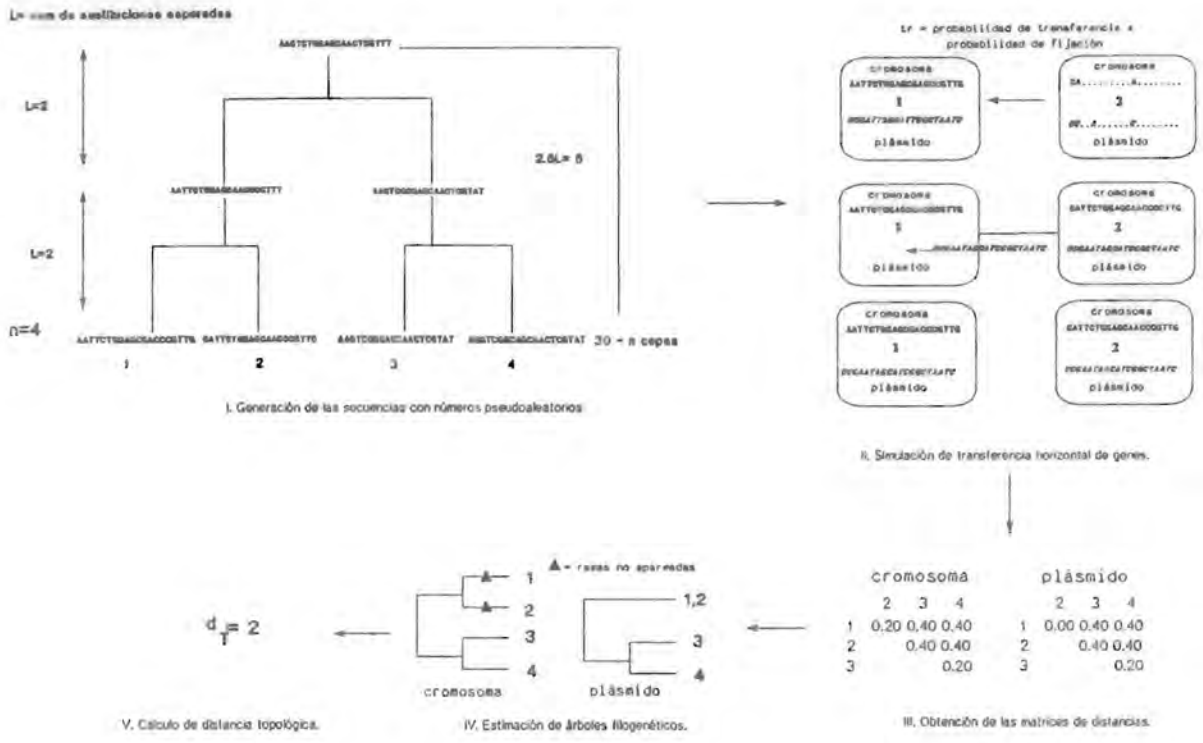
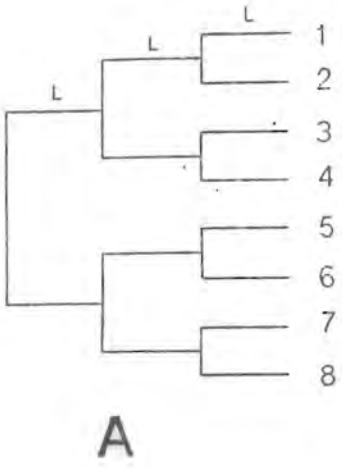
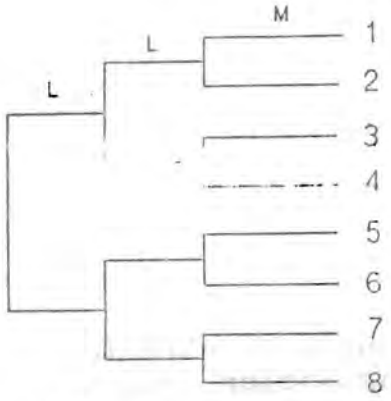


Figura 6. Diagrama de los pasos seguidos durante la simulación en computadora de transferencia horizontal de genes. En este caso unicamente se muestra la generación de las secuencias cromosomales de 20 bases para 4 OTUs esperando 2 sustituciones en cada paso de ramificiación (I). Se ejemplifica un solo evento de transferencia y su efecto en las topologias resultantes.

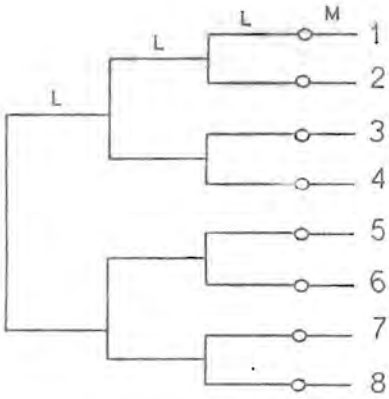
Figura 7



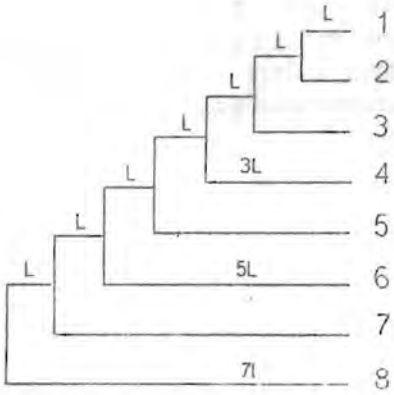
A



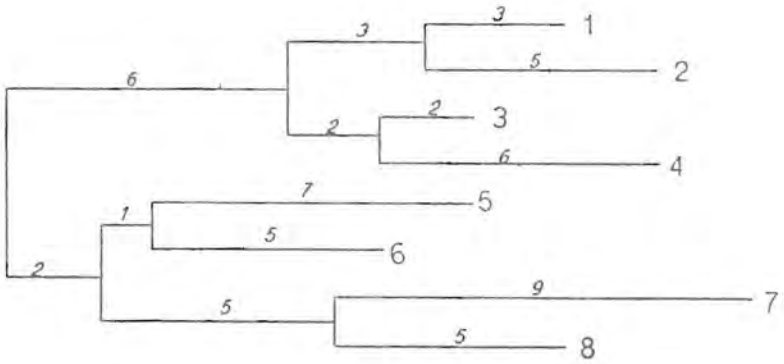
B



C



D



E

Figura 7. Arboles modelo utililzados en los experimentos de simulación muestrando ocho OTUs cada vez de un grupo de linajes bacterianos. M y L, o los números en itálicas (árbol E) denotan la longitud de las ramas y representan el número esperado de sustituciones nucleótidicas por gen, suponiendo un proceso de Poisson. En cada caso, con excepción del árbol C, se simuló que la transferencia horizontal de genes cuando ocurría, lo hacía al final de las ramas (i.e. los linajes bacterianos eran muestrados inmediatamente después de haber intercambiado material genético). En el árbol C, los círculos abiertos representan el lpunto en el que la transferencia horizontal de genes tuvo lugar, esperando que ocurrieran M sustituciones tanto en el árbol del plasmido como en el arbol del cromosoma en las ocho secuencias muestreadas.

de tiempo como la probabilidad de fijación, suponiendo que esta probabilidad es constante para todos los linajes.

Después de 500 unidades de tiempo (generaciones) calculamos las distancias geneticas para las *n* especies muestreadas y estimamos los árboles filogenéticos correspondientes. En este trabajo unicamente trabajamos con árboles sin raíz. Estos árboles fueron comparados con aquellos derivados del "cromosoma" y se obtuvo una medida de distancia topológica.

Distancia topológica

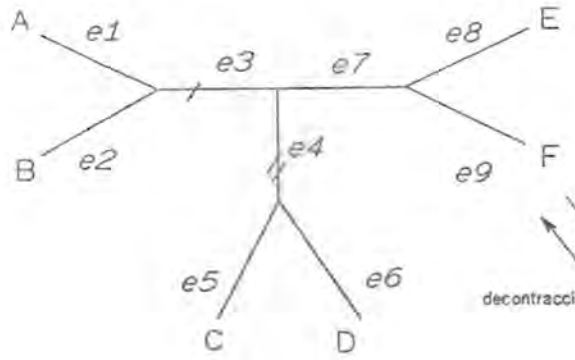
Para obtener una medida de distancia topológica usamos el indice de distorsión (d_T) de Robinson y Foulds (1981) (tambien llamado metrica de partición). A continuación se explican algunos conceptos topológicos - ya definidos en otros trabajos (Robinson and Foulds, 1981; Hendy et al. , 1984)- útiles para entender dicha medida de distancia topológica.

Sea L_n un conjunto de *n* etiquetas *[1,2,...n]*, donde *n* es el número de unidades taxonómicas operativas (OTUs) o linajes en el árbol. Es decir, cada etiqueta es un linaje o OTU. Un árbol filogenético es un árbol que tiene a lo más *n* nodos de grado 1, aunque puede tener menos. El grado de un nodo esta dado por el número de ramas unidas a él. Cada miembro de L_n aparece en solamente un nodo. A estos nodos se les conoce como nodos terminales y estan "etiquetados" por uno o más elementos de Ln, y a las ramas unidas a ellos se les conoce como ramas terminales. Los nodos restantes (internos) son cada uno de grado 3. Sean *T1* y *T2* dos árboles filogenéticos. El índice de distorsión d(T1,T2), - d_T para abreviar - se define como el número mínimo de transformaciones necesarias para convertir *T1* a *T2*. Dicha medida es equivalente a contar el número de ramas no apareadas en ambos árboles. Una rama no apareada es aquella que al ser partida da dos conjuntos no vacios de OTUs Se' y Se'', tal que ninguna otra rama del otro árbol pueda dar, al ser particionada, los mismos conjuntos. El índice de distorsión adopta valores entre 0 (árboles idénticos) y $3n - 6$ (ninguna partición en común, inclyendo ramas terminales). (figura 8)

Los árboles modelo fueron probados para diferentes valores de la tasa de transferencia. Para cada valor de tasa de transferencia se calculó el número de eventos de transferencia durante la simulación en todo el grupo bacteriano. Las diferencias entre el número observado promedio de eventos de transferencia y los números esperados fueron muy pequeñas. Por esta razón el análisis se hizo en función del número esperado de eventos de transferencia.

Figura 8

T_1



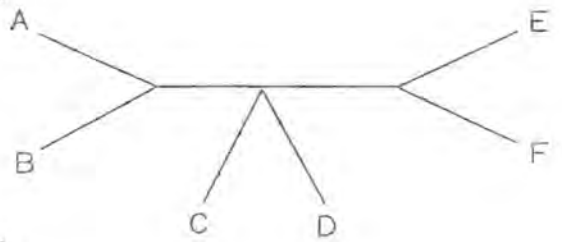
Partiendo e3 en T1 se obtienen los siguientes conjuntos: $Se3' = \{A,B\}$, $Se3'' = \{C,D,E,F\}$

y partiendo e3 en T2 : $Se3' = \{A,B\}$, $Se3'' = \{C,D,E,F\}$

por lo tanto e3 en T1 y e3 en T2 estan apareadas.

contraccion

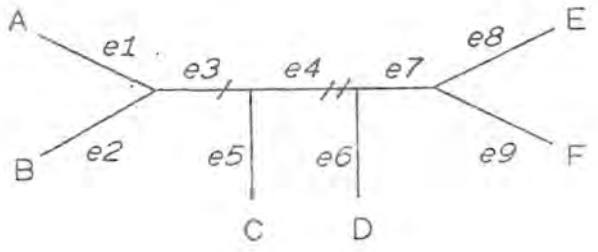
decontraccion



contraccion

decontraccion

T_2



Partiendo e4 en T1 se obtienen los siguientes conjuntos: $Se4' = \{C,D\}$, $Se4'' = \{A,B,E,F\}$

tal que ninguna rama en T2 da los mismos conjuntos

y partiendo e4 en T2 : $Se4' = \{A,B,C\}$, $Se4'' = \{D,E,F\}$

tal que ninguna rama en T1 da los mismos conjuntos

por lo tanto e4 en T1 es no apareada con respecto a T2 y e4 en T2 es no apareada con respecto a T1.

Figura 8. Ejemplo de el cálculo del indice de distorsión (d_T) contando el número de ramas no apareadas entre dos árboles. Puesto que el número de ramas no apareadas para este caso en particular es 2, entonces $d_{(T1,T2)} = 2$.

$$d_c = d_m / (1 + e^{-z})$$

$$z = a + b' \ln x = a + \ln x^{b'}$$

$$d_c = d_m / (1 + e^{-a} \cdot e^{-\ln x^{b'}})$$

$$= d_m / (1 + k\, x^{-b'}) = d_m / (1 + k/x^{b'})$$

$$\boxed{d_c = \frac{d_m\, x^{b'}}{k + x^{b'}}}$$

$$\frac{k + x^{b'}}{d_m\, x^{b'}} = 1/d_c \quad ; \quad \frac{k}{d_m} \cdot \frac{1}{x^{b'}} + \frac{1}{d_m} = 1/d_c$$

$$k/x^{b'} + 1 = d_m/d_c \rightarrow$$

$$k/x^{b'} = d_m/d_c - 1 = \frac{d_m - d_c}{d_c}$$

$$\frac{k\, d_c}{d_m - d_c} = x^{b'} \qquad x = \left[\frac{k\, d_c}{d_m - d_c}\right]^{1/b'}$$

RESULTADOS

Hay una clara relación entre el número de eventos de transferencia en una población y la diferencia topológica entre las filogenias plasmídicas y cromosomales como muestra un ejemplo de simulación usando el árbol modelo B de la figura 7 con L=4 y el método de unir vecinos (neighbor joining) para reconstrucción filogenética (figura 9). El número esperado de eventos de transferencia en una escala logarítmica ($\log_{10}$) al ser graficado contra el índice de distorsión promedio (d_T) (figura 10) muestra claramente una forma logística usando tres métodos distintos de reconstrucción (UPGMA, MF y NJ). Por lo tanto usamos el siguiente modelo logístico:

$$d_T = d_{Tmax} / (1 + e^{-z}) \qquad (1)$$

donde $z = a + b\log_{10}x$, x es el número esperado de eventos de transferencia dado un valor determinado de tr (tasa de transferencia y subsecuente fijación). El parámetro a define el valor de intersección de d_T cuando $\log_{10}x = 0$; el parámetro b determina la pendiente de la curva logística entre las dos asíntotas y d_{Tmax} define la cota superior de d_T.

A fin de comparar estadísticamente las curvas obtenidas a partir de datos de simulación se hicieron regresiones no lineales (procedimiento de Marquardt, Hewlett Packard series 98820A, Statistical Library Series 9000) con los datos obtenidos de las simulaciones. De este modo también evaluamos qué tan bien pueden ser explicados nuestros resultados por la ecuación (1). Sin embargo, debido a que los intervalos de confianza obtenidos con este análisis de regresión suponen una distribución normal, se utilizaron las medias de 20 replicaciones para probar nuestro modelo.

Las regresiones no lineales que se hicieron con cada grupo de datos usando la ecuación (1) se muestran como lineas en la figura 3. Las curvas así obtenidas son muy similares entre los diferentes métodos. Es importante hacer notar que las incongruencias filogenéticas pueden deberse a otros factores además de la transferencia de genes. El más común de estos factores es el de una reconstrucción filogénetica incorrecta. Por lo tanto, en primer lugar se explorará el efecto que cierta magnitud de error de reconstucción tendría en la estimación de la transferencia de genes.

Errores de reconstrucción

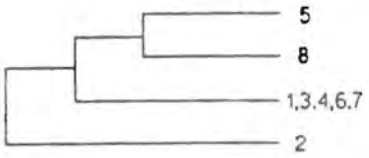
Existen varios factores, más o menos comunes a todas las metodologías, que pueden resultar en una estimación incorrecta de la topologia, las longitudes de las ramas, o ambas (ver Holmquist et al. [1988] para una revisión). El enfoque que ha sido utilizado para estudiar este problema consiste en fijar un árbol modelo, simular los cambios evolutivos en frecuencias alélicas, o secuencias aminoácidas/nucleotídicas siguiendo este árbol modelo. Si esta

Figura 9

$tr = 2.5 \times 10^{-4}$

$x_{obs} = 97$

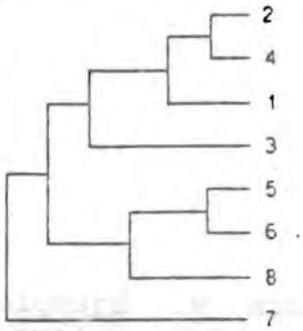
$d_T = 12$



$tr = 1.66 \times 10^{-5}$

$x_{obs} = 9$

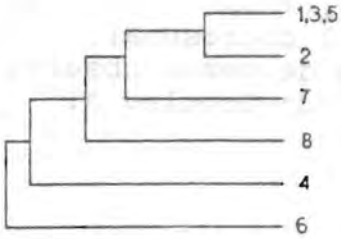
$d_T = 6$



$tr = 1.25 \times 10^{-4}$

$x_{obs} = 46$

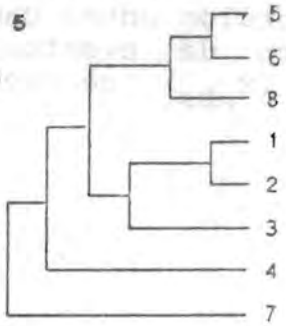
$d_T = 12$



$tr = 8.33 \times 10^{-6}$

$x_{obs} = 5$

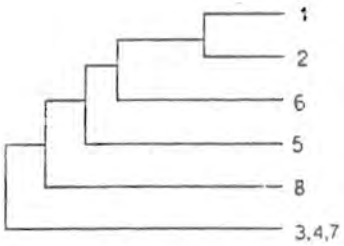
$d_T = 6$



$tr = 6.25 \times 10^{-5}$

$x_{obs} = 28$

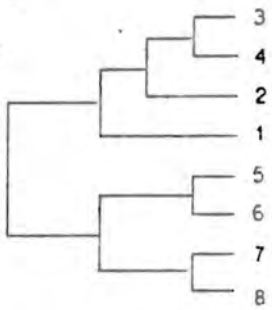
$d_T = 10$



$tr = 4.16 \times 10^{-6}$

$x_{obs} = 4$

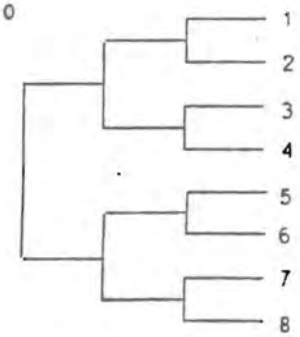
$d_T = 2$



$tr = 4.16 \times 10^{-7}$

$x_{obs} = 1$

$d_T = 0$



$tr = 3.33 \times 10^{-5}$

$x_{obs} = 15$

$d_T = 8$

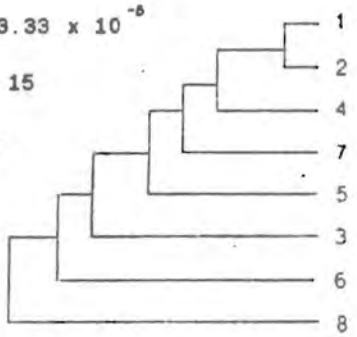


Figura 9. Ejemplo de una simulación en computadora usando diferentes valores de la tasa de transferencia (*tr*). Las topologías se estimaron usando el método de unir vecinos (neighbor joining). En este caso, la topología del árbol cromosomal obtenido resultó idéntica al árbol modelo usado (arbol modelo B en la figura 7; L=4, M=6). Los valores del índice de distorsión entre cada árbol y el árbol cromosomal, así como el número de eventos de transferencia de genes observados en cada caso (x_{obs}) se muestran al lado de cada topología.

Figura 10

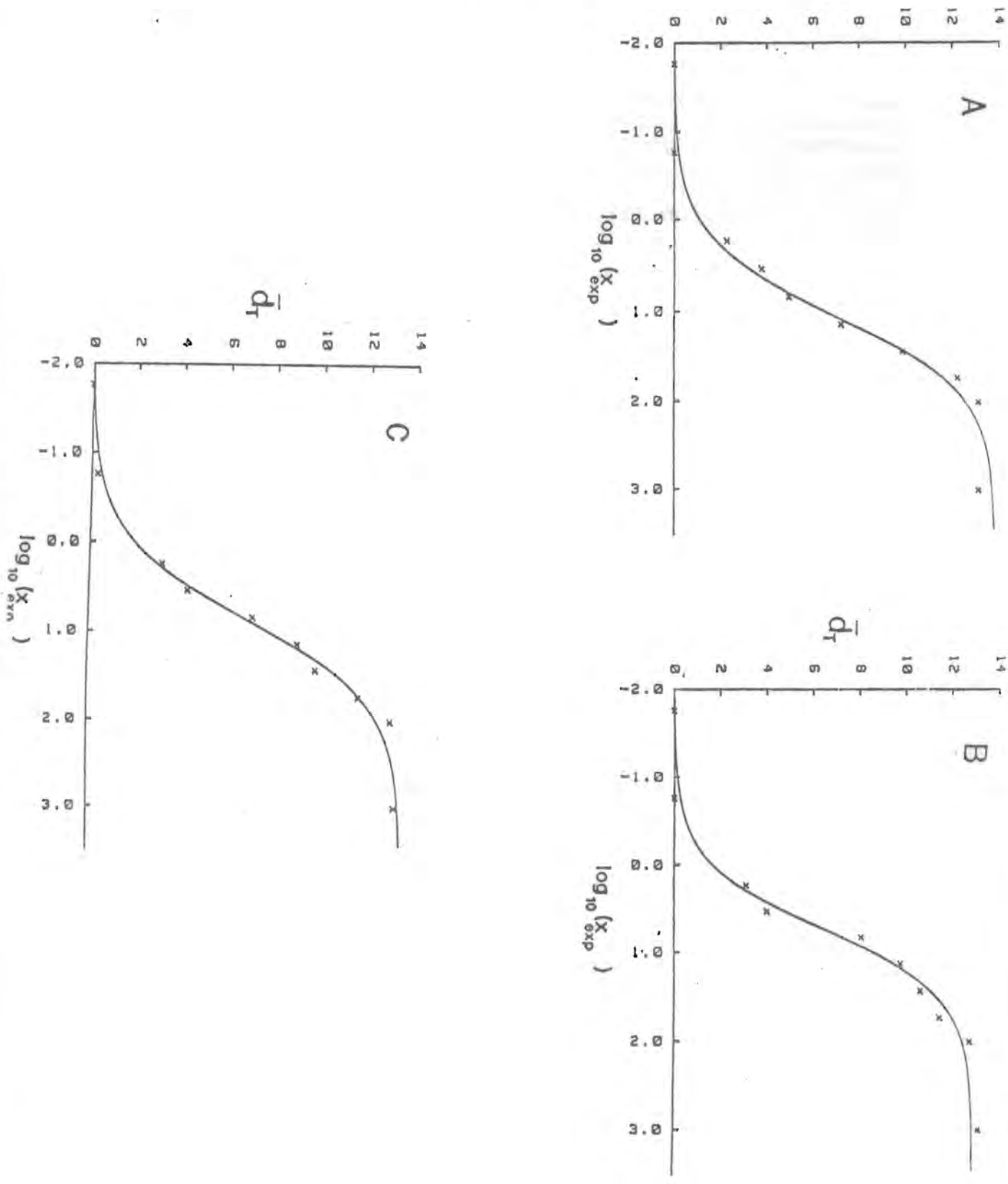


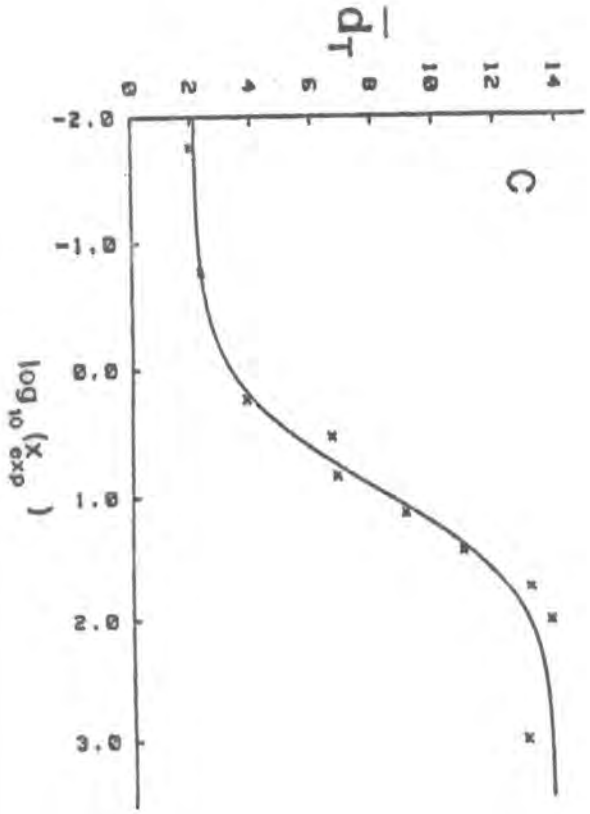
Figura 10. Indice de distorsión promedio (d_T) de 20 replicaciones, como función del numero esperado de eventos de transferencia de genes ($\log_{10}[x_{exp}]$). Las topologias se estimaron usando UPGMA (A), el método de Farris modificado (B) y el método de unir vecinos (C). Las regresiones no lineales que se ajustaron usando la ecuación (1) se muestran en la figura.

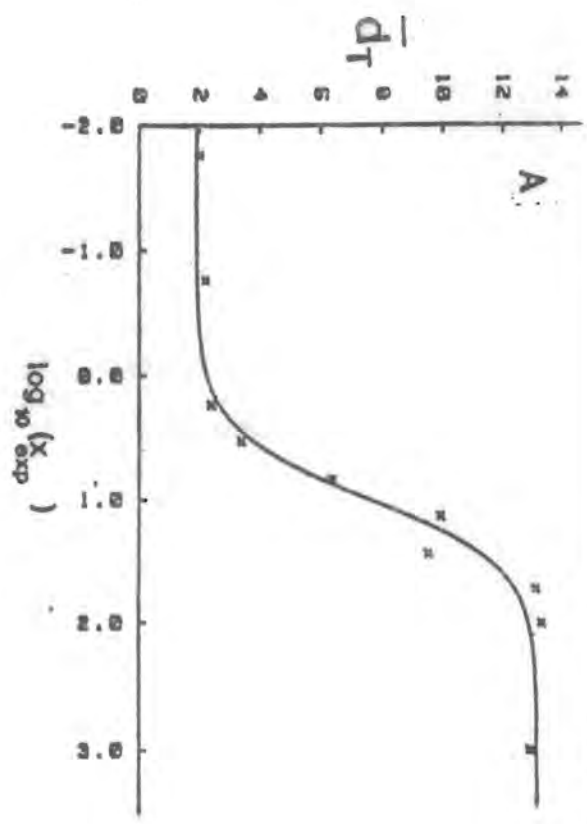
simulación se repite muchas veces es posible evaluar la probabilidad de obtener el árbol correcto para cualquier método dado de reconstrucción filogenética (Sourdis y Nei, 1988). Este enfoque ha sido utilizado para probar la eficiencia de varios métodos basados en matrices de distancias (Blanken et al. 1982; Tateno et al. 1982; Tateno y Tajima, 1986); Sourdis y Krimbas, 1987; Saitou y Nei, 1987). Tambien ha sido usado para evaluar métodos de máxima parsimonia (Sourdis y Nei, 1988) el método de máxima verosimilitud de Felsenstein (1981) (Rohlf y Wooten, 1988) y el método de parsimonia evolutiva de Lake (1987) (Li et al, 1987).En general, los distintos métodos funcionan de manera diferente dependiendo de las condiciones. Por ejemplo el método de Lake (1987) de parsimonia evolutiva da muy buenos resultados bajo condiciones extremas (longitudes de ramas muy distintas entre linajes - 10 veces más largas -) pero es bastante ineficiente en condiciones más moderadas (ramas 2 o 3 veces más largas.

El objetivo de este trabajo no es evaluar las eficiencias de los distintos métodos sino el efecto que diferentes eficiencias de reconstrucción tendrían en evaluar el número de eventos de transferencia. Por lo tanto, se exploró el efecto en la distancia topológica como función del número de eventos de transferencia cuando hay un cierto error de reconstrucción involucrado. Se escogieron tres métodos de matrices de distancias con distintas eficiencias de reconstrucción bajo distintas circunstancias. Los métodos de matriz de distancias consumen mucho menos tiempo que los métodos de máxima parsimonia o de máxima verosimilitud y además dan una sola topología final lo cual hace que los datos sean mucho más sencillos de analizar. Los factores inherentes a los métodos de matriz de distancias que los hacen susceptibles de errores son: *a*) Hay una gran pérdida de información al convertir los datos de secuencia a valores de distancia. *b*) Estos métodos son sensibles a cambios pequeños en los valores de distancia. Esto es particularmente cierto cuando los valores son pequeños. (Holmquist et al 1988).

Para poder medir el efecto del error de reconstruccion usamos u árbol modelo que fuera estimado de manera incorrecta parte de la veces y con un error asociado de reconstrución moderado. La reconstruccion filogenetica es extremadamente difícil cuando las tasas de evolución varían mucho entre linajes (Li et al 1987). Es importante, entonces, estudiar la utilidad de nuestro modelo cuando las tasas de sustitución son variables. El árbol modelo E en la figura 7 fue reconstruido correctamente *P*% de las veces con un error de reconstrucción asociado d_T (tabla 1) por tres métodos de matriz de distancia: UPGMA, MF y NJ. Los experimentos de simulación han demostrado que estos tres métodos tienen diferentes eficiencias de reconstrucción (Saitou y Nei 1987) lo cual puede verse en la tabla 1. Las eficiencias de reconstrucción obtenidas en este trabajo no son estrictamente consistentes con las de Saitou y Nei (1987). Una posible explicación es que nosotros usamos un árbol con longitudes de ramas variables a diferencia de Saitou y Nei (1987).

Figura 11


C
$\frac{1}{d_T}$
$\log_{10}(X)_{exp}$


A
$\frac{1}{d_T}$
$\log_{10}(X)_{exp}$

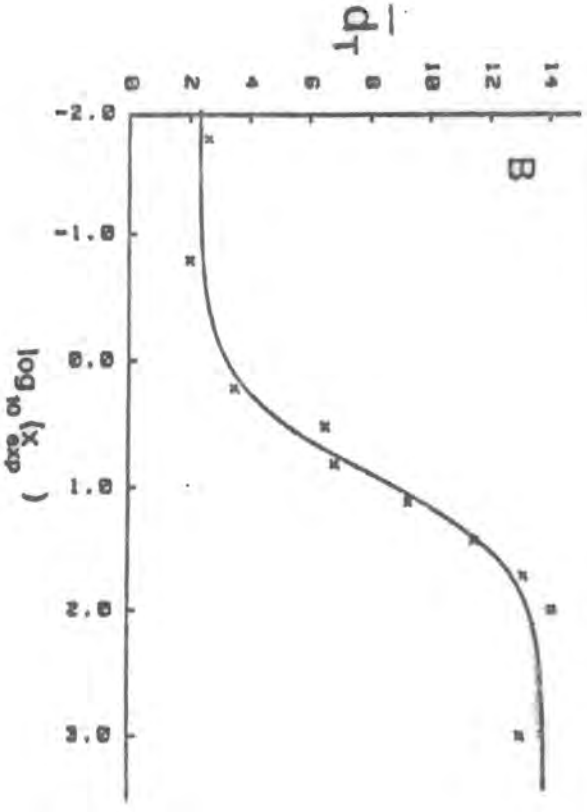

B
$\frac{1}{d_T}$
$\log_{10}(X)_{exp}$

Figura 11. Indice de distorsión promedio (d_T) de 20 replicaciones como función del número de eventos esperados de transferencia de genes ($\log_{10}[x_{exp}]$). El árbol E (figura 7) fue usado como modelo. Los métodos de reconstrucción empleados fueron: UPGMA (A), el método de Farris modificado (B) y el método de unir vecinos (C). Se muestran las regresiones no lineales ajustadas con la ecuación (2).

Tabla 1.

Parámetros estimados de la regresión (ec. 2) del índice de distorsión como función del número de eventos de transferencia de genes tomando en cuenta el error de reconstrucción.

	a	b	c	d_{Tmax}	r^2	P[1]	d_T[1]
UP	-3.4696 (-5.65,-1.28)*	3.5013 (1.34 , 5.65)	1.8976 (0.36 , 3.40)	13.2440 (11.57,14.91)	0.9421	52.0	1.72 (± 2.13)
MF	-2.5222 (-3.89,-1.15)	2.6885 (1.28 , 4.09)	2.3122 (0.87 , 3.75)	13.9372 (12.25,15.61)	0.9799	20.0	2.56 (± 2.06)
NJ	-2.1971 (-3.37,-1.01)	2.3756 (1.15 , 3.59)	2.1231 (0.65 , 3.58)	13.9305 (12.15,15.70)	0.9803	28.0	2.00 (± 1.52)

*Los límites superior e inferior de los intervalos de confianza se muestran entre paréntesis.

1. Los resultados se basan en 50 replicas.

a, b, c y d_{Tmax} son parametros de la regresión (ec. 2); r^2 es el coeficiente de correlación al cuadrado. P es el porcientaje de veces que cada método reconstruyó correctamente el árbol modelo E (fig. 7) y d_T es el error promedio de reconstrucción.

Las secuencias cromosómicas y plasmídicas se simularon como distintas al principio del experimento de simulación en cada linaje. Dada la magnitud del error de reconstrucción involucrado no era de esperarse que ambos árboles (el cromosómico y el plasmídico) fueran iguales aunque el número esperado de sustituciones en las dos secuencias de cada linaje fuera el mismo. Para modelar esta situación supusimos lo siguiente:

$$d_T = c + ([d_{Tmax} - c]/ (1 + e^{-z}) \qquad (2)$$

donde c es un parámetro que depende del error de reconstrucción y $z = a + b\log_{10}x$. Los resultados de esta simulación se muestran en la figura 10 y los parámetros estimados se muestran en la tabla 2. Es interesante que c no es significativamente diferente de el error de reconstrucción observado y que los otros tres parametros (a, b y d_{Tmax}) son muy semejantes entre los diferentes métodos utilizandos. De esta manera, es posible usar un valor aproximado al error de reconstrucción dados dos árboles a comparar y usar este valor como un estimador de c. Esto sugiere que para nuestros propósitos no es importante qué método de reconstrucción filogenética se use para estimar el número de eventos de transferencia.

Sensibilidad a longitud de las ramas, topología del árbol y número de OTUs.

Para simplificar nuestro análisis, en las siguientes simulaciones se ha manejado una situación en la cual no hay error de reconstrucción involucrado. Para evitar el efecto debido a errores de reconstrucción se hizo que las secuencias del cromosoma y el plásmido acarreadas por cada linaje fueran idénticas al principio de la simulación. Por lo tanto, en ausencia de transferencia de genes, los árboles filogenéticos correspondientes serían idénticos, independientemente de su similitud con la topología verdadera.

Se probó hasta que punto podía el modelo logístico (ecuación 1) funcionar de manera satisfactoria bajo distintas condiciones tales como longitudes distintas de las ramas, distinta topología y variando el número de especies en el árbol. La primera de tales condiciones que se probó fue el número de sustituciones nucleotídicas por rama. Por lo tanto se usó el árbol modelo A (figura 7) con dos diferentes valores de L: L=2 y L=15. Las curvas obtenidas resultaron muy parecidas a las de la figura 10. Los resultados se resumen en la tabla 2: (I: L=4; II: L=2; III : L=15) y es notable que los parámetros estimados (a, b y d_{Tmax}) al ser comparados entre las diferentes curvas no son estadísticamente diferentes con un nivel de significancia del 95%, como lo muestran los intervalos de confianza. A fin de probar si había un efecto debido a que las ramas teminales fueran más largas que el resto usamos el árbol modelo B (figura 7; L=4, M=6) y los parámetros que se ajustaron se muestran en la tabla 2 (IV) que son muy similares a los antes mencionados. Es interesante notar que los parámetros a, b y d_{Tmax} en estas simulaciones no son significativamente distintos de

Tabla 2. Parámetros estimados de la regresión (eq. 1) para el indice de distorsión como función del número de eventos de transferencia genes.

	METODO (1)	a		b		d_{1max}		r^2
I (2)	UP	-2.4522	(-3.20,-1.70)	2.3489	(1.52 , 3.17)	13.8619	(12.11,15.61)	0.9925
	MF	-1.9040	(-2.45,-1.35)	2.5691	(1.79 , 3.34)	12.8746	(11.83,13.90)	0.9909
	NJ	-1.7874	(-2.13,-1.44)	2.0881	(1.63 , 2.53)	13.5169	(12.53,14.49)	0.9943
II	UP	-3.3743	(-4.68,-2.39)	3.5326	(2.75 , 4.31)	13.4420	(12.67,14.21)	0.9979
III	UP	-3.5999	(-5.25,-1.94)	4.0379	(2.10 , 5.97)	13.1327	(11.70,14.55)	0.9911
IV	UP	-2.8704	(-3.52,-2.21)	3.2164	(2.43 , 3.99)	13.5427	(12.67,14.41)	0.9938
	MF	-2.4608	(-2.71,-2.20)	3.0364	(2.70 , 3.37)	13.3075	(12.92,13.68)	0.9995
	NJ	-2.7041	(-3.48,-1.92)	2.9431	(2.02 , 3.86)	13.4148	(12.21,14.61)	0.9893
V	UP	-2.4895	(-3.24,-1.73)	3.3036	(2.19 , 4.41)	9.3313	(8.45,10.20)	0.9917
	MF	-2.3461	(-3.25,-1.44)	3.0361	(1.68 , 4.38)	9.5808	(8.29,10.86)	0.9855
	NJ	-2.1769	(-3.07,-1.27)	2.8644	(1.46 ,4.26)	9.6125	(8.15,11.07)	0.9833
VI	UP	-2.4322	(-3.71,-1.15)	2.8786	(1.24 , 4.50)	13.7176	(11.61,15.83)	0.9626
	MF	-1.8169	(-2.97,-0.65)	2.2617	(0.69 , 3.82)	14.0266	(11.06,16.99)	0.9474
	NJ	-2.3624	(-3.60,-1.12)	2.7675	(1.19 , 4.33)	13.9116	(11.69,16.13)	0.9613

* Los limites inferior y superior de los intervalos de confianza al 95% se muestran entre parentesis.

(1):Los metodos de reconstruccion utilizados fueron: UP, UPGMA; MF, Modified Farris; NJ, Neighbor Joining method. method;

(2): Los numeros romanos I- IV se refieren a los distintos arboles modelo en la figura 6 o diferentes longitudes de las ramas de la siguiente forma:

I. Arbol modelo A, L=4; II, arbol modelo A, L=2; III, arbol modelo A, L=15;IV, arbol modelo B, L=4,M=6; V.Arbol modelo C, L=4,M=2; VI, arbol modelo D, L=4;

los mismos parámetros presentados en la tabla 1, cuando habia un error de reconstrucción involucrado.

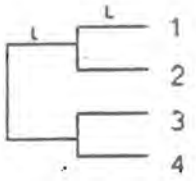
Hasta ahora se ha supuesto que la transferencia de genes ocurría inmediatamente antes de que las OTUs de muestra fueran analizadas. Es posible, sin embargo, que cierta cantidad de mutación haya tenido lugar desde la transferencia. El árbol C (figura 7) representa esta situación, donde los círculos marcan el punto en que la transferencia de genes ocurrió y la longitud de la rama después de este punto representa el número esperado de sustituciones después de la transferencia. Los resultados de las regresiones no lineales se muestran en la tabla 2 (V) y el efecto observable es una disminución en el parámetro d_{Tmax}. Esto se debe a que, puesto que ha habido mutación, independientemente de la cantidad de transferencia de genes, todas las OTUs involucradas son distinguibles; por lo tanto, todas las ramas terminales en ambos árboles estan apareadas. Cuando todas las ramas terminales estan apareadas, la máxima distorsión para un árbol de n OTUs es $2n - 6$ (10 en este caso en particular). Es interesante hacer ver que los valores de dTmax estimados no son significativamente distintos de 10 lo que indica que el máximo real para linajes que han divergido a partir del intercambio genético viene dado por $2n - 6$.

A fin de explorar el efecto de la topología del árbol en el comportamiento de d_T como función del número de eventos de transferencia génica, se uso el árbol modelo D (figura 7) el cual presenta una topología bastante diferente a la hasta ahora utlilizada. Todos los parámetros son notablemente similares a los de las otras simulaciones (I-IV) no habiéndose encontrado diferencias al nivel de confianza del 95% (tabla 2, VI).

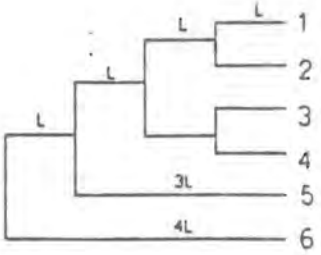
Finalmente, se exploró la dependencia de d_T con respecto al número de OTUs en el árbol y los árboles modelo utilizados se presentan en la figura 12. A partir de los parámetros ajustados y los coeficientes de correlación sobre las regresiones no lineales mostrados en la tabla 3 se concluye que el número de OTUs en el árbol sólo se refleja en el parámetro d_{Tmax} y no en la forma de la curva. Existe una razón teórica para esto: a medida que la cantidad de transferencia génica aumenta, todas las secuencias plasmídicas van a tender a ser idénticas (un solo tipo de secuencia se fija en todos los linajes) y por lo tanto el árbol filogenético de los plásmidos va a estar dado por un solo punto. El índice de distorsión será igual al número de ramas en el árbol cromosomal (todas las ramas son no apareadas puesto que no hay ramas en el árbol del plásmido) que es igual a $2n - 3$ para un árbol sin raíz. La relación observada entre n y el valor estimado de d_{Tmax} se muestra en la figura 13 junto con la linea recta $2n - 3$.

Con base en los resultados presentados se concluye que el modelo logístico propuesto es una buena aproximación de el comportamiento de d_T como función del número de eventos de transferencia de genes y puede por lo tanto ser usado para estimar los niveles de transferencia de genes en poblaciones bacterianas bajo una gran

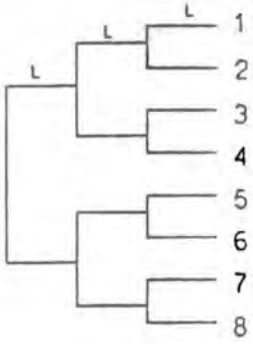
Figura 12



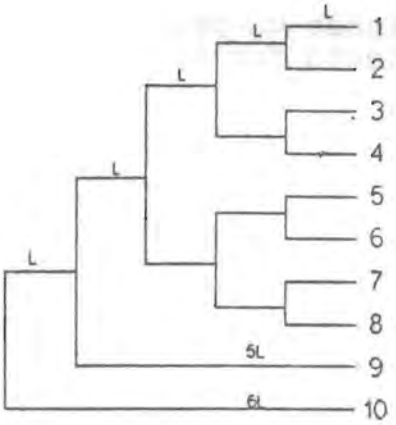
A



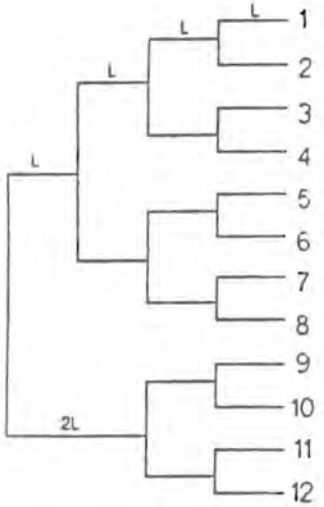
B



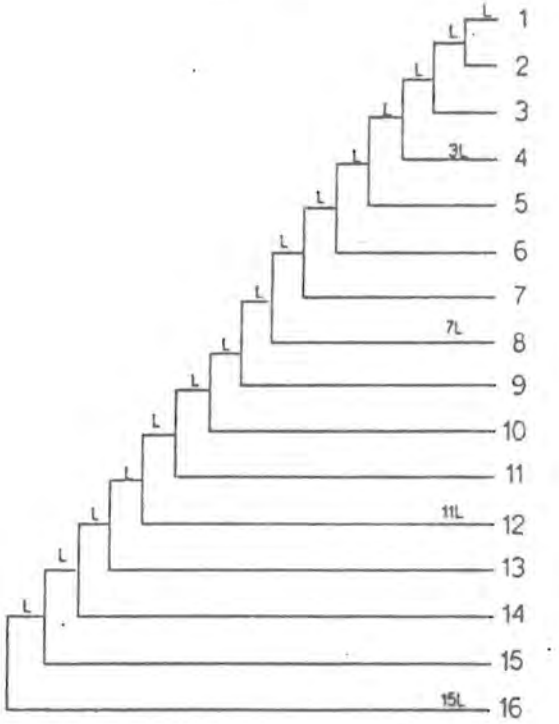
C



D



E



F

Figura 12. Arboles modelo utilizados en los experimentos de simulación muestreando diferente número de OTUs. L representa el número esperado de sustituciones nucleotídicas por rama suponiendo un proceso de Poisson. En todos los casos el valor de L empleado fue 15.

Tabla 3. Parametos estimados de la regresión (ec.1) de el índice de distorsión como funcion del número de eventos de transferencia variando el número de OTUs en el árbol.

NUMERO DE OTUs	a	b	d_{Tmax}	r^2
4	-4.1173 (-5.50,-2.73)	2.1422 (1.17,3.11)	5.3753 (4.13,6.61)	0.9896
6	-2.2597 (-3.12,-1.38)	1.9527 (1.03,2.87)	9.2620 (7.40,11.11)	0.9866
10	-2.6714 (-4.05,-1.80)	3.0085 (1.30,4.69)	18.5177 (15.71,21.34)	0.9843
12	-3.3724 (-4.94,-1.80)	3.8943 (1.99,5.78)	23.1439 (20.57,25.71)	0.9774
16	-0.9634 (-2.34,0.41)	2.7932 (-0.95,4.63)	33.0904 (27.34,38.83)	0.9691

Los límites inferior y superior de los intervalos de confianza al 95% se muestran entre parentesis. a,b y d_{Tmax} son parámetros de la regresión (ec. 1); r^2 es el coeficiente de correlación al cuadrado.

Figura 13

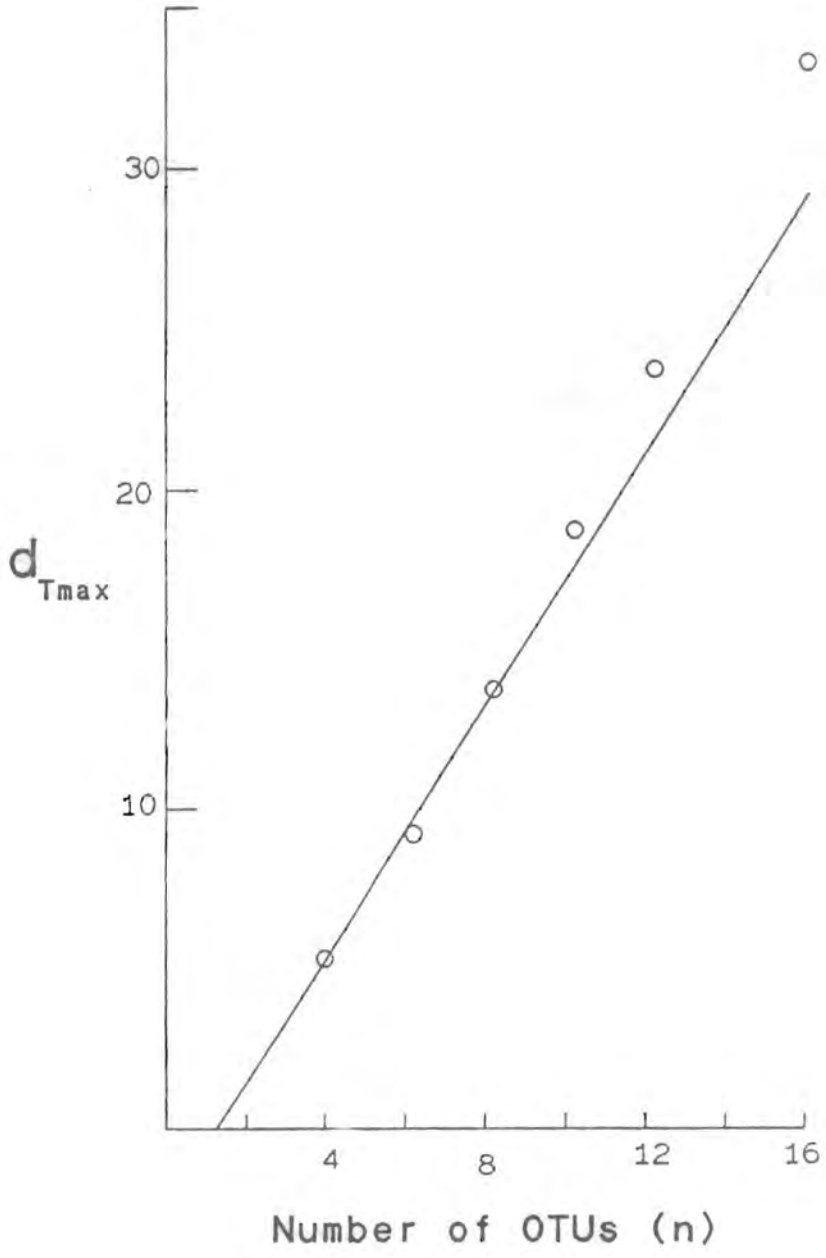


Figura 13. Valor estimado del índice de distorsión máximo (d_{Tmax}) como función del número de OTUs (n) en el árbol. Los círculos abiertos son los estimados obtenidos a partir de las regresiones no lineales. El mejor ajuste para los estimados de d_{Tmax} (r^2 =0.9970) resultó ser dTmax = 2.3317n - 4.680. Los límites inferior y superior de los intervalos de confianza al 95% fueron, respectivamente, [1.9791, 2.5073] y [-6.4592, -2.9007]. La linea graficada es $d_{Tmax} = 2n - 3$ y representa el comportamiento esperado teóricamente.

variedad de condiciones.

UNA APLICACION DEL METODO: Transferencia de genes entre muestras de *Rhizobium leguminosarum* biovar *viceae*.

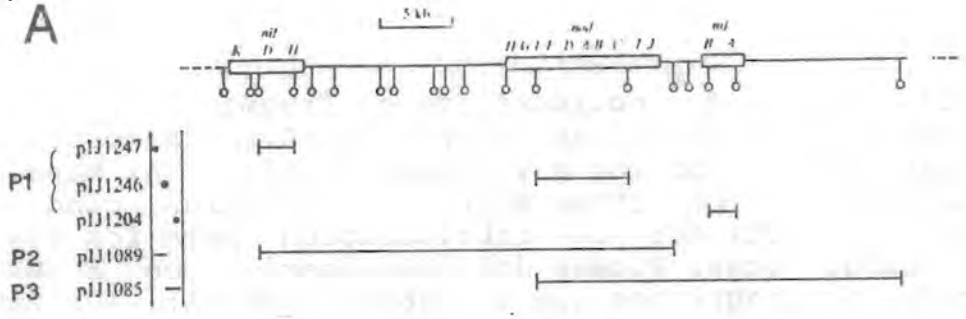
Young y Wexler (1988) reportaron datos de fragmentos de restricción (RFLPs) para 85 aislados de entre los cuales ellos distinguieron 50 tipos genéticos distintos gracias a las sondas utilizadas. La primera de estas sondas, un fragmento de 26.2 kb contiene el gen estructural para la beta-galactosidasa y las otras cinco sondas contenían secuencias del plásmido Sym. De estas cinco, tres consistían unicamente de genes estructurales involucrados en la nodulación y la fijación de nitrógeno. Estas tres sondas se encotraban flanqueadas por sitos *Eco*RI y la suma de los tamaños las tres era de 10.85 kb. Para el análisis se combinaron los tres y se les llamo *P1*. Las otras dos sondas con secuencias del plásmido se superponían en parte con uno de los primeros tres y eran de 25.6 y 29.65 kb. Los llamamos *P2* y *P3* respectivamente (figura 14).

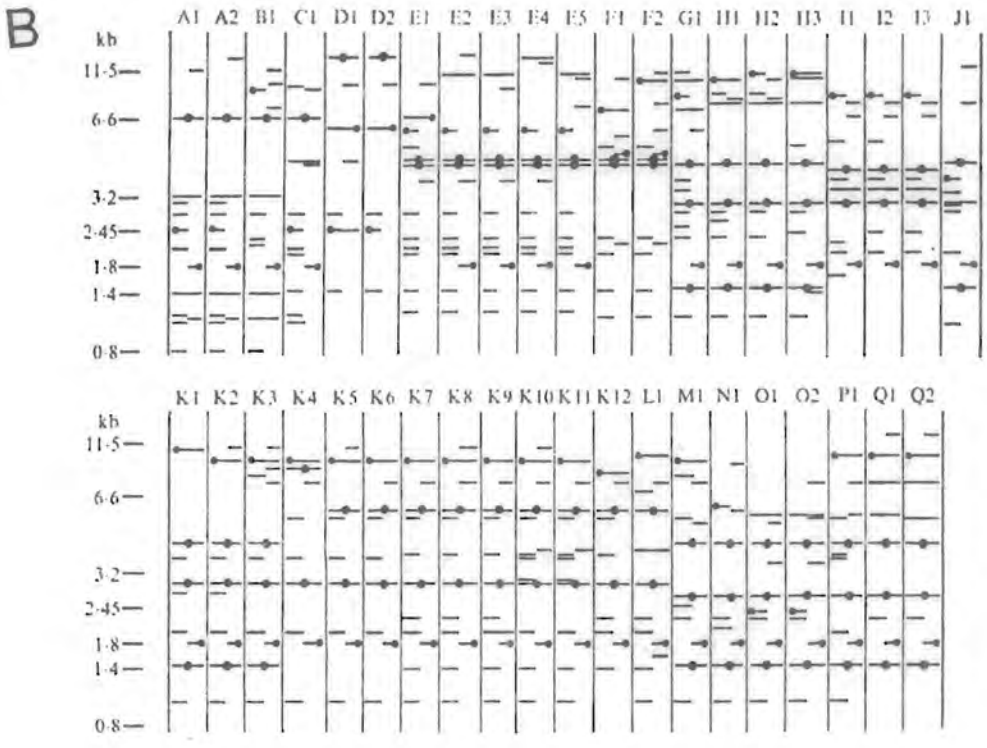
Young y Wexler (1988) concluyeron que ciertos tipos cromosomales iban asociados con ciertos tipos de Sym. Esto quiere decir que el intercambio genético entre cepas se restringe a algunos tipos genéticos y Young y Wexler propusieron varias alternativas para explicar este hecho. Es bastante claro a partir de sus datos que algunos tipos con el mismo cromosoma acarrean diferentes marcadores plasmídicos y que aislados idénticos si solo se considera el plásmidos se encontraban en distintos tipos cromosomales. Estos hallazgos les permitieron concluir que "en el caso de *Rhizobium* tenemos .... evidencia de que cierta cantidad de intercambio de plásmidos Sym debe haber ocurrido...". A una conclusión similar llegaron Schofield et al. (1987) usando otros marcadores geneticos en *Rhizobium leguminosarum* biovar *trifolii*. Estos resultados son apoyados por el hecho de que no hay, aparentemente, límites fisiológicos a la transferencia de plásmidos entre cepas y especies de *Rhizobium* bajo condiciones de laboratorio (Johnston et al. 1978) Por todas estas razones *Rhizobium* nos pareció un sistema muy conveniente para estudiar la cantidad de transferencia de genes que ha ocurrido naturalmente en la historia de una población.

De los 50 tipos genéticos reportados por Young y Wexler (1988) se escogieron 36 distinguibles usando el marcador cromosomal y los probes estructurales del Sym. Para estos 36 aislados se reconstruyó una filogenia sin raíz usando el algoritmo UPGMA a partir de matrices de distancias. Las matrices de distancias se construyeron utilizando el estimador de similitud para datos de fragmentos de restricción de Nei y Li (1979) y transformándolo a sustituciones nucleotídicas. Se obtuvieron árboles filogenéticos para el marcador cromosomal y para cada uno de los marcadores plasmídicos (*P1*, *P2* y *P3*, figura 15).

Las filogenias reconstruidas a partir de *P1*, *P2* y *P3* son muy diferentes de la filogenia cromosomal. De hecho, los índices de distorsión fueron 66,74 y 73 respectivamente (con respecto al árbol

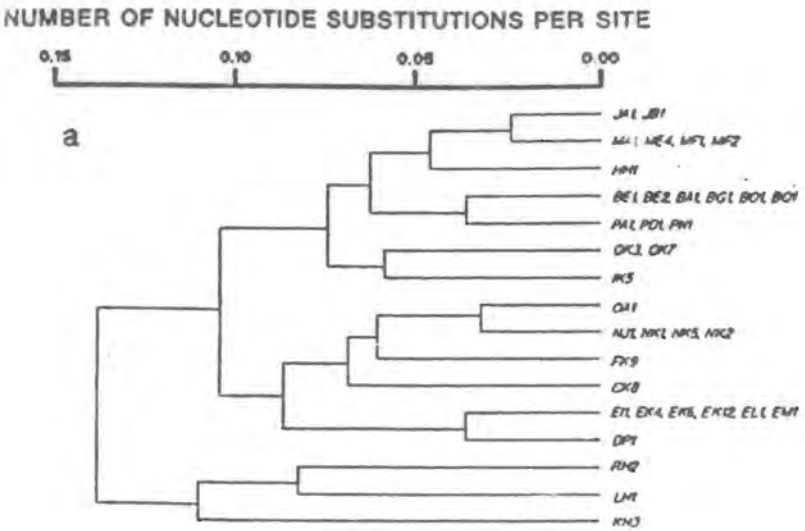
Figura 14

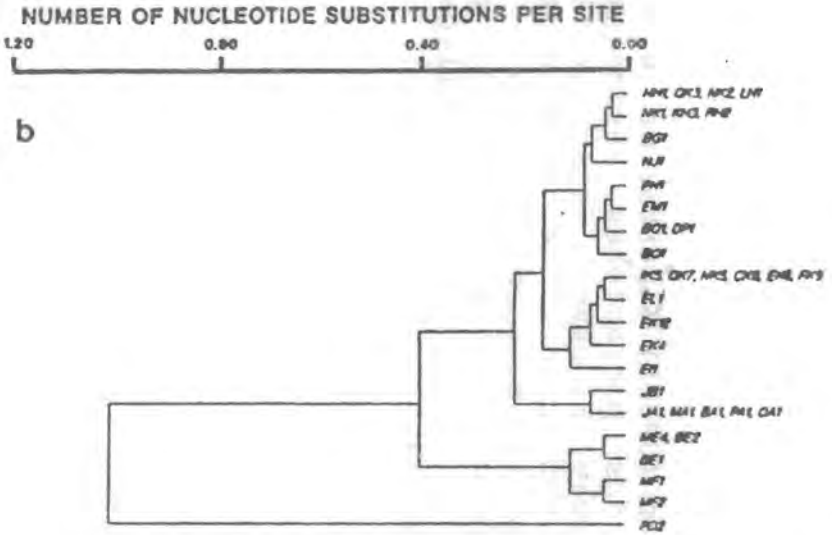

A
5 kb
P1
pIJ1247
pIJ1246
pIJ1204
P2
pIJ1089
P3
pIJ1085

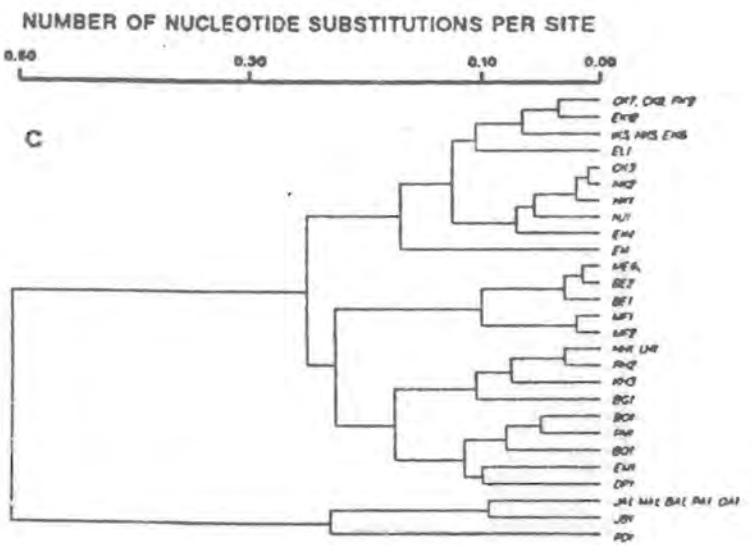

B
A1 A2 B1 C1 D1 D2 E1 E2 E3 E4 E5 F1 F2 G1 H1 H2 H3 I1 I2 I3 J1
kb
11·5
6·6
3·2
2·45
1·8
1·4
0·8
K1 K2 K3 K4 K5 K6 K7 K8 K9 K10 K11 K12 L1 M1 N1 O1 O2 P1 Q1 Q2
kb
11·5
6·6
3·2
2·45
1·8
1·4
0·8


C
kb
O B P J M K R L N H I F E D C Q
7·0
5·3
4·2
2·3
1·7
1·2

Figura 14. Datos de polimorfismo en fragmentos de restricción (RFLPs) para los aislados de campo de *R.leguminosarum* biovar *viceae* reportados por Young y Wexler (1988). (A) Mapa de parte del plásmido Sym. Los sitios EcoRI se muestran como círculos abiertos. (B) Patrones de hibridización para los aislados de *R.leguminosarum* biovar *viceae* del DNA digerido con *Eco*RI usando como sondas los fragmentos que se muestran en (A). (C) Patrones de hibridización para los aislados de *R.leguminosarum* biovar *viceae* del DNA digerido con *Eco*RI usando como sonda el plásmido plac12 que acarrea un segmento cromosomal clonado el cual incluye el gen de beta-galactosidasa.


NUMBER OF NUCLEOTIDE SUBSTITUTIONS PER SITE
0.15
0.10
0.05
0.00
a


NUMBER OF NUCLEOTIDE SUBSTITUTIONS PER SITE
1.20
0.80
0.40
0.00
b


NUMBER OF NUCLEOTIDE SUBSTITUTIONS PER SITE
0.60
0.30
0.10
0.00
c

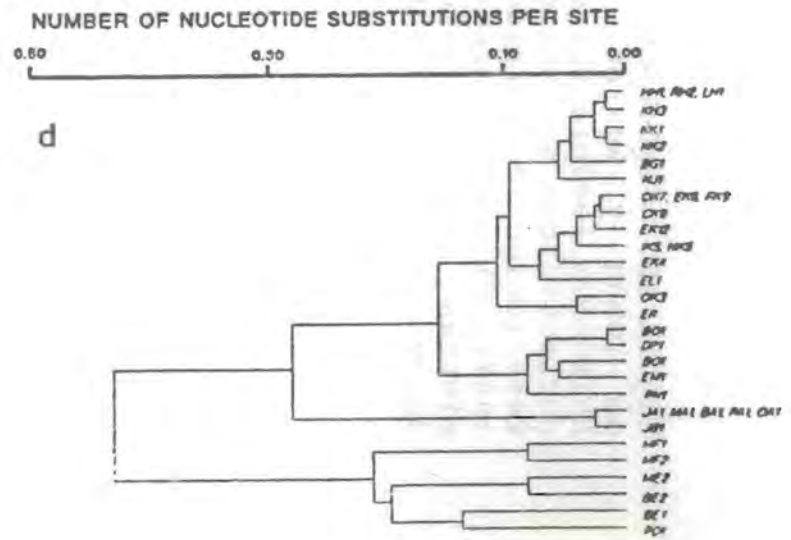

NUMBER OF NUCLEOTIDE SUBSTITUTIONS PER SITE
0.60
0.30
0.10
0.00
d

Figura 15

Figura 15. Toplogias reconstruídas (UPGMA) a partir de los datos de Young y Wexler (1988) de dos poblaciones agrícolas de *Rhizobium leguminosarum* biovar *viceae*. El número de sustituciones nucleotídicas se estimó usando la proporcio2n de fragmentos de restricción compartidos siguiendo los métodos descritos en Nei, 1987. El panel a muestra la topologia obtenida usando plac, mietras que los paneles b, c y d muestran las filogenias derivadas de *P1*, *P2* y *P3* respectivamente.

cromosomal) mientras que las distorsiones entre *P1*, *P2* y *P3* fueron de 42, 38 y 44. Esto apoya la conclusión de Young y Wexler (1988) de que debe haber existido una cierta cantidad de transferencia horizontal de genes entre estos aislados.

Para estimar el número de eventos que se fijaron en estos aislados se hizo un bootstrapping (Efron 1982) con los datos, muestreando 8 aislados con reemplazo 50 veces.

El bootstrap.

La idea básica del bootstrap consiste en inferir la variabilidad en una distribución desconocida de la cual han sido extraídos ciertos datos. Supóngase que se tienen los puntos $x_1, x_2, \ldots x_n$ que fueron extraidos de manera independiente de la mismá distrubución. A partir de éstos, aplicando algún método T de estimación estadística se obtiene un estimador

$$t = T(x_1, x_2, \ldots x_n)$$

de algún parámetro en el que estamos interesados.

El procedimeinto de bootstrap es muy útil cuando o bien no se conoce la distribución de las x_i o cuando T es tan compleja que su error standard es muy difícil de calcular. Lo que se hace entonces es remuestrear los datos y construir una serie de conjuntos ficticios de datos. Cada uno de éstos se construye muestrando con reemplazo m puntos ($m \leq n$) de las x_i. Cada uno de los conjuntos de datos ficticios consiste de m puntos $x_1^*, \ldots x_m^*$ donde cada punto x_i^* se tomo al azar de los n datos originales. Para cada conjunto ficticio de datos se calcula el estimador

$$t^* = T(x_1^*, x_2^*, \ldots x_m^*)$$

El proceso de remuestreo se repite r veces calculando cada vez t^*. La idea esencial es que el conjunto de estimadores t^* tiene una distribución que se aproxima a la distrubución real de t.

Las filogenias de las muestras obtenidas con el bootrap fueron reconstruidas usando los datos de los tres marcadores plasmídicos y el marcador cromosomal, y los índices de distorsión para las 6 comparaciones necesarias se calcularon. Es decir, para cada combinación de 8 cepas de obtuvieron cuatro árboles filogenéticos que fueron comparados entre sí. Una muestra sesgada de las réplicas ilustra el caso en el que aparentemente no ha ocurrido transferencia de genes entre los aislados utilizados (figura 16a), el caso más común en el que hubo transferencia de plásmidos Sym (figura 16b) y un caso más que sugiere recombinación entre plásmidos mediada por transferencia de genes (figura 16c). Un análisis de varianza mostró que las diferencias entre las medias de la tabla 4 son significativas ($F = 80.03$; $p<0.001$) y una prueba de Bartlett de

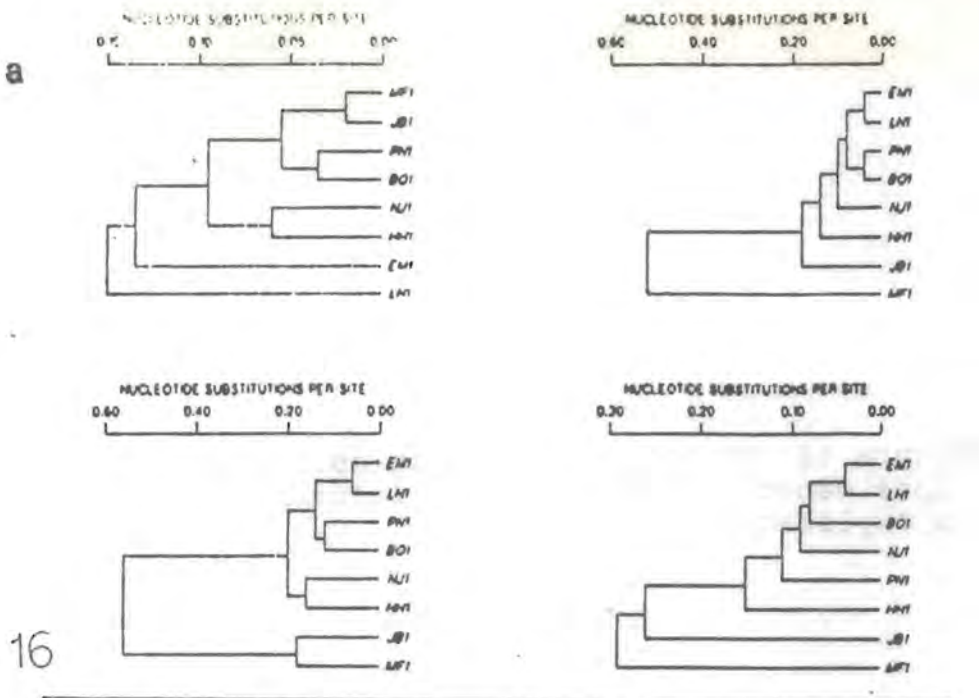
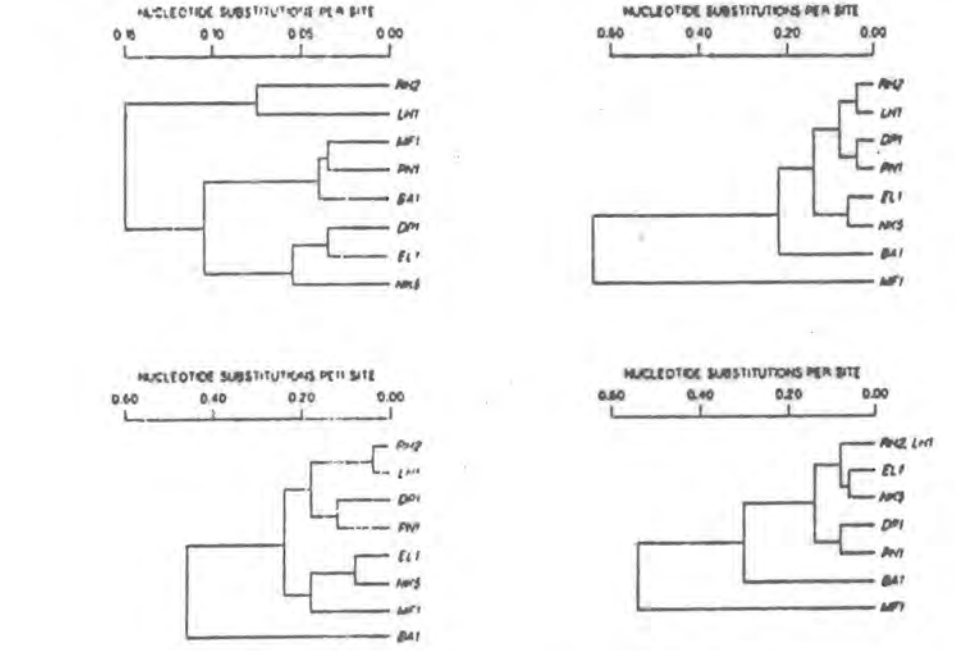

a
NUCLEOTIDE SUBSTITUTIONS PER SITE
0.60 0.40 0.20 0.00
0.30 0.20 0.10 0.00

Figura 16


b
NUCLEOTIDE SUBSTITUTIONS PER SITE
0.60 0.40 0.20 0.00

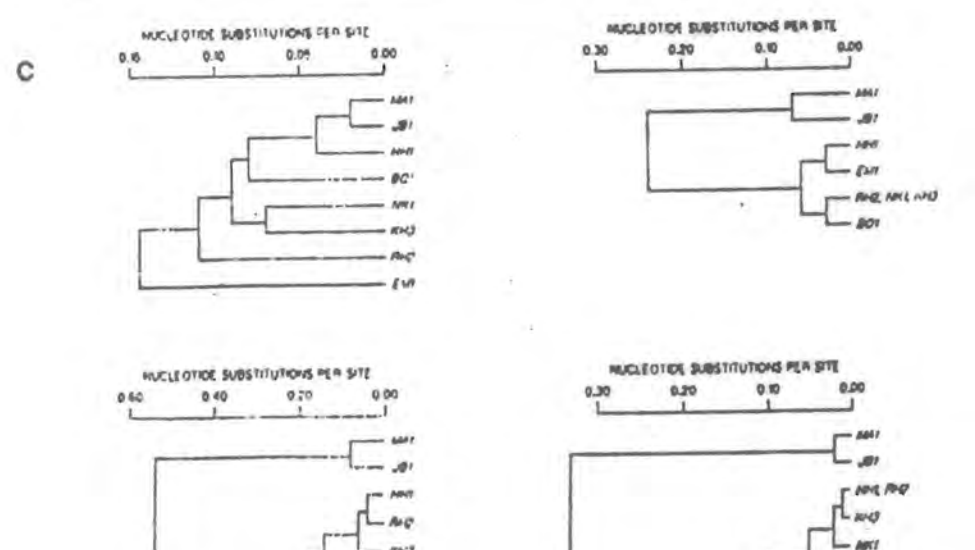

c
NUCLEOTIDE SUBSTITUTIONS PER SITE
0.30 0.20 0.10 0.00
0.60 0.40 0.20 0.00

Figura 16. Ejemplos de topologías reconstruidas (UPGMA) para ocho tipos genéticos de *Rhizobium leguminosarum* biovar *viceae* (véase la explicación en el texto).

homoscedasticidad mostró que no hay heterogeneidad significativa entre las varianzas (Xi cuadrada 2.06; p= 0.067) aunque una de las varianzas es particularmente pequeña (*plac* y *P3*; tabla 4). Al mismo tiempo, una prueba de Student-Newman-Keuls formó tres grupos estadisticamente significativos de índices de distorsión. El primero de estos es la distorsión entre *P1* y *plac*, que tiene el valor más alto. El segundo grupo corresponde a la distorsión entre *plac* y los otros dos probes del Sym. Finalmente, el tercer grupo está compuesto por las distorsiones entre *P1*, *P2* y *P3*. Con esta información se puede utilizar la relación logística obtenida de las simulaciones para estimar el número de eventos de transferencia que han ocurrido en los aislados estudiados por Young y Wexler (1988).

Dado que el índice de distorsión promedio entre los árboles filogenéticos del plásmido es diferente de cero (5.12 en promedio), resulta lógico preguntarse si estas diferencias se deben a errores de reconstrucción filogenética o más bien reflejan pequeñas cantidades de recombinación entre diferentes porciones del plásmido Sym. Por esta razón se estimaron los árboles filogenéticos más parsimoniosos para 10 conjunto datos de 8 OTUs cada uno de los probes *P1* y *P2*. Para ello se usó la versión 2.4 para PC de PAUP escrita por David Swofford de la Universidad de Illinois. El algoritmo asumía que el ancestro hipotético no contenía ninguno de los caractéres y los árboles más parsimoniosos fueron derivados usando un algoritmo de branch-and-bound. Los árboles así obtenidos fueron comparados con los árboles estimados por UPGMA y se obtuvo un valor promedio de d_T de 3.894. Suponiendo que 4.00 sea el error de reconstrucción de fondo y sustituyendo c en la ecuación 2 por este valor el número de eventos obtenidos son los que se muestran en paréntesis en la diagonal inferior de la tabla 4. Los estimados del número de eventos de transferencia de genes a partir de comparaciones entre plásmido y cromosoma no son muy diferentes de los obtenidos suponiendo que no hay error de reconstrucción. Sin embargo, los estimados de número de eventos de transferencia para comparaciones entre plásmidos disminuyen considerablemente. A pesar de esto último, no se puede concluir de manera definitiva que las diferencias topológicas entre las filogenias de varias porciones del plásmido se deban exclusivamente a errores de reconstrucción por las siguientes razones: a) Los métodos de máxima parsimonia dan varios árboles, y aún si el error de reconstrucción fuera cero usando UPGMA algunos de los árboles m¿as parsimoniosos deben de ser incorrectos. b) Es posible, aunque no muy probable, que toda la diferencia topológica se deba a la existencia de recombinación intraplásmido, y que la diferencia topológica entre los árboles obtenidos por los dos métodos se deba a errores de reconstrucción sólo en el método de máxima parsimonia. c) Más importante aún, el error de reconstrucción no explica toda la diferencia topológica. Por lo tanto no se puede excluir la posibilidad de recombinación entre plásmidos.

Nuestras simulaciones estan diseñadas para estimar el número de eventos de transferencia que han ocurrido en la totalidad de la población de tipos genéticos. Para poder estimar este valor en los datos de Young y Wexler (1988) se muestrearon al azar sin reemplazo

Tabla 4.

Indice de distorsión promedio y número esmtimado de eventos de transferencia de genes a partir de los datos de Young y Wexler (1988).

	plac	P1	P2	P3
plac	-	11.62 ± 2.46 [a (1,2,3)]	10.26 ± 2.28 [b]	10.00 ± 1.85 [b]
P1	55 (36) [(4)]	-	5.36 ± 2.93 [c]	4.96 ± 2.48 [c]
P2	29 (18)	7 (2)	-	5.06 ± 2.42 [c]
P3	27 (16)	6 (1)	6 (1)	-

(1) Diagonal superior: índice de distorsión derivado de comparaciones entre ár boles filogenéticos.

(2) Las desviaciones standard se muestran al lado de cada valor de distorsión.

(3) Las medias seguidas por la misma letra no son significativamente diferentes al nivel del 5%.

(4) Diagonal inferior: número estimado de eventos de transferencia de genes sin tomar en cuenta el error de reconstucción o con un error $d_T = 4.0$ (números entre paretesis).

los 85 aislados y se construyó una gráfica con el número acumulado de tipos genéticos (figura 17). Estos datos se ajustaron a una función exponencial negativa usando el procedimento de Marquardt (Hewlett Packard 98820A; Statistical Library Series 9000) con el siguiente modelo:

$$g = m_0[1 - exp(-m_1)(i)]; \qquad (3)$$

donde g es el número acumulado de tipos genéticos, i es el número acumulado de aislados y m_0 y m_1 son parámetros de la regresión no lineal. Este modelo fue propuésto por Miller y Wiegaert (1989) como un modelo robusto para ajustar datos a fin de estimar el número total de especies de plantas vasculares raras en los Apalaches.

Usando este enfoque se obtuvo un estimado de 89.09 tipos genéticos presentes en los dos campos agrícolas estudiados por Young y Wexler (1988) según revelaron las sondas usadas por ellos. Este estimado nos permite calcular el número promedio de eventos de transferencia por tipo genético usando los datos de la tabla 4. Si se toman las comparaciones entre el cromosoma y las sondas de plásmido despreciando un efecto debido a errores de reconstrucción, se tiene que entre el 30 y el 60 % de los tipos genéticos han participado en eventos de transferencia desde que empezaron a divergir, mientras que si se considera el error de reconstrucción este valor cae a entre el 20 y el 40 % de los tipos genéticos. Por otra parte, las comparaciones entre las distintas sondas del plásmido dan un valor superior de entre 7 y 8% de tipos genéticos involucrados en eventos de transferencia (despreciando el error de reconstrucción) y un valor inferior de 1.5% (tomando en cuenta el error de reconstrucción).

Figura 17

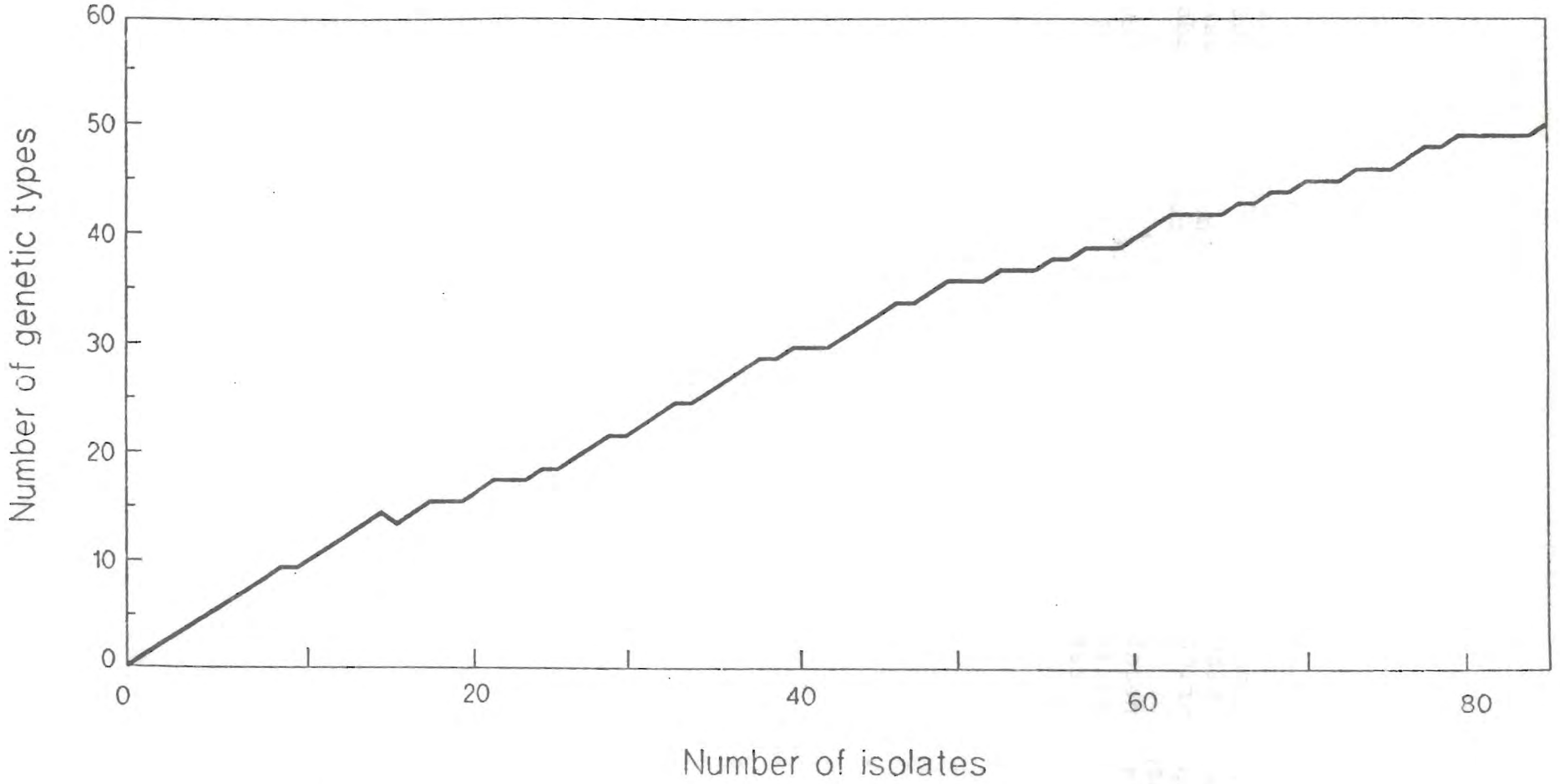

Number of genetic types
60
50
40
30
20
10
0
0
20
40
60
80
Number of isolates

Figura 17. Número acumulado de tipos genéticos con respecto al número de aislados muestreados al azar en las poblaciones británicas agrícolas de *Rhizobium leguminosarum* biovar *viceae* analilzadas por Young y Wexler (1988). La curva ajustada es una exponencial negativa de la forma descrita en el texto (ecuación 3). Los parámetros estimados y sus respetivos límites a los intervalos de confianza al 95% fueron m_0 = 89.09 (85.08,93.10) y m_1 = 0.0098 (0.0092, 0.0104).

DISCUSION

Nuestros resultados demuestran que cuando hay transferencia de genes, la magnitud de este proceso se refleja en el índice de distorsión entre árboles filogenéticos (derivados de diferentes genes) de manera cuantitativa. Bajo diferentes condiciones, el comportamiento de d_T fue esencialmente el mismo con excepción del parámetro d_{Tmax}. Como ya se ha discutido, hay razones teóricas para esto. Se propone que para analizar un conjunto de datos donde se observa incongruencia filogenética pero sin nodos terminales etiquetados por más de una OTU, se utilice $2n - 3$ como cota superior para d_{Tmax} y $2n - 6$ como cota inferior.

Sin embargo, las incongruencias filogenéticas pueden deberse a otros factores aparte de la transferencia de genes, algunos de los cuales son inherentes al proceso evolutivo (Holmquist et al. 1988). Entre éstos se encuentra la escasez relativa de sitios filogenéticamente informativos entre especies relacionadas de manera cercana. Esto no parece ser un problema en bacterias, donde los estimados que se tienen de diversidad genética son muy altos (Selander et al. 1987; Piñero et al. 1988). Más aún, Lawrence y colaboradores (1989) analizaron varias secuencias de inserción para 23 cepas muy relacionadas de *Escherichia coli*. Sus resultados apoyan la idea de que el número y la localización de los elementos de inserción en el genoma bacteriano evolucionan tan rápido que es posible inferir las relaciones filogeneticas entre cepas muy relacionadas. Otro factor evolutivo que puede dar lugar a incongruencias filogenéticas es la existencia de polimorfismo en la población ancestral. Pamilo y Nei (1988) han desarrollado un algoritmo para estimar la probabilidad que el árbol derivado de un gen en particular corresponda topologicamente con el árbol de especies o poblaciones. Aunque se requieren estimados de el tamaño efectivo de la población y los tiempos de divergencia, es posible en principio saber con qué probabilidad dos árboles filogenéticos derivados de diferentes genes son incongruentes debido a transferencia de genes o a polimorfismo ancestral.

Por otra parte, la presencia de secuencias repetidas debe ser tomada en cuenta al interpretar datos filogenéticos. El uso de secuencias que pueden estar repetidas ya sea en el cromosoma, el plásmido o ambos y la interpretación de dichos datos como si correspondieran a secuencias ortólogas puede llevar a conclusiones erroneas si los genes son en realidad parálogos. De hecho, se ha reportado la existencia de genes repetidos en *Rhizobium leguminosarum* biovar *phaseoli* (Flores et al. 1987) y por lo tanto no puede despreciarse la posibilidad de que un cierto gen este repetido en el genoma bacteriano. Más aún, se ha encontrado que las secuencias repetidas algunas veces se hallan en diferentes replicones (V. Gonzalez, comunicación personal) lo que puede hacer que el análisis de dichos datos sea aún más complicado.

Los factores metodológicos (ver Holmquist et al. [1988] para una

revisión) pueden también afectar el resultado si el error involucrado es muy grande. No obstante, como se ha demostrado, la topologia verdadera no es necesaria siempre y cuando sea posible estimar el error de reconstrucción asociado. Puesto que no existe trabajo teórico sobre la función de densidad de probabilidad de d_T para árboles que no comparten todos los nodos terminales, no es posible por el momento derivar una varianza para d_T ni para el número de eventos de transferencia (x) para una sola comparación de árboles. Sin embargo, el enfoque de bootstrap como el empleado en este trabajo puede ser muy útil para obtener una aproximación del error asociado con la estimación de d_T.

Además de los factores evolutivos y metodológicos, los factores biológicos, al afectar las tasas de sustitución nucleotidica, pueden también dificultar la reconstrucción filogenética. Las tasas de sustitución variables, sin embargo, no afectan la estimación del número de eventos de transferencia si el error de reconstrucción asociado no es muy grande, como lo indican los resultados usando el árbol modelo E (figura 7). Más aún, el error de reconstrucción puede ser compensado utilizando la ecuación (2) en vez de la ecuación (1) y el estimado del error de reconstrucción como valor para el parámetro c.

Loa datos en este trabajo apoyan la idea que la magnitud de transferencia horizontal de genes de hecho puede ser estimada basándose en el principio de congruencia filogenética (Wilson et al. 1977). Se pudo estimar el número de eventos de transferencia de genes capaz de explicar las diferencias topológicas entre árboles reconstruidos usando marcadores del plásmido Sym y del cromosoma de *Rhizobium leguminosarum* biovar *viceae* (Young y Wexler 1988). Este análisis no excluye la posibilidad de que exista cierto grado de recombinación entre varias porciones del plásmido mediada por transferencia de genes. Aunque este fenómeno no obscurece las conclusiones principales sí señala nuevas lineas de investigación relacionadas con el análisis de más secuencias del plásmido Sym que representen diferentes funciones o localizaciones genéticas. Esto puede ser distinto en otras bacterias que presenten medios más dinámicos de transferencia de genes como podría ser el caso de *Pseudomonas* (Chakrabarty et al. 1978) en donde se ha observado que los plásmidos pueden operar como factores de mobilización para gran cantidad de genes.

En muchos casos, no es posible evaluar qué tan representativo es el muestreo de una población natural de bacterias. Al aplicar un método utlizado en estudios de flora regional (Miller y Wiegert 1989) a los datos de Young y Wexler (1988) se hace evidente que al menos la mitad de los tipos genéticos de *Rhizobium leguminosarum* biovar *viceae* fueron muestreados. Para los propósitos de este trabajo este dato es crítico a fin de obtener las tasas de transferencia por tipo genético. Se obtuvieron estimados de tasa de transferencia que van desde 1/89 hasta 55/89 eventos por tipo desde que los linajes en cuestión empezaron a divergir. Desgraciadamente no se tiene ningún estimado confiable de tiempo de divergencia para

esta especie. No obstante, la investigación de la evolución molecular de bacterias en el futuro puede hacer accesibles este tipo de datos.

Al considerar las consecuencias evolutivas de la transferencia horizontal de genes, dos aspectos separados son importantes: la magnitud de su efecto en genes cromosomales y en genes extracromosomales. Las tasas de evolución, desde luego, se veran mucho menos alteradas si la tasa de transferencia horizontal sólo afecta secuencias extracromosomales. Sin embargo, mucho de la adaptación bacteriana es a través de la expresión de genes acarreados en elementos accesorios y la recombinación entre estos elementos es ciertamente importante para la evolución de las bacterias (Levin 1988). Más aún, la pregunta sobre si la relación genética observada entre plásmidos se debe a que tienen un ancestro común y no a la rápida dispersión a través de las barreras de especie queda sin contestar (Campbell 1981).

Los resultados de estudios elecroforéticos de enzimas para varias especies bacterianas (Selander et al. 1985; Selander et al. 1987; Piñero et al. 1988) indican el mantenimiento de una estructura de población clonal y por lo tanto, bajos niveles de recombinación. Sin embargo, DuBose et al. (1988) han sugerido que si la recombinación sólo involucra secuencias cortas de DNA (menos de 500 bp) la estimación de desequilibrio por ligamiento se vería afectada únicamente por un evento de intercambio en el que uno de los marcadores genéticos estuviera incluido en la porción de material intercambiado. Sus resultados comparando secuencias indican que la recombinación intragénica ha desarrollado un papel importante en la historia evolutiva del gene *phoA* en nueve cepas diferentes de *Escherichia coli*. Puesto que la transferencia de genes se requiere para que haya intercambio cromosomal, sería interesante si se tuvieran datos de secuencias plasmídicas homólogas para las mismas nueve cepas, estimar los niveles de transferencia de plásmidos en estas cepas y compararlos a los niveles estimados de recombinación. De esta manera sería posible estimar la fracción de transferencia de genes que da lugar a intercambio cromósomico.

Las técnicas estadísticas desarrolladas recientemente para analizar secuencias de DNA han mostrado fuertes evidencias de recombinación intragénica en primates (Stephens 1985), *Drosophila melanogaster* (Hudson y Kaplan 1985) y en bacterias (DuBose et al. 1988; Sawyer 1989). El método propuesto en este trabajo no intenta medir eventos de recombinación sino intercambio genético. En bacterias, sin embargo, ambos fenómenos estan interrelacionados de manera cercana y las incongruencias filogenéticas podrían ser utilizadas para medir recombinacion. Las pruebas estadísticas para medir conversión génica antes mencionadas no requieren la obtención de toplogías y ambos tipos de enfoque filogenético, topológicos y no topológicos pueden resultar complementarios. Una ventaja de los enfoques topológicos es que los datos de secuencia no son indispensables. Como Avise (1989) ha señalado "los enfoques filogenéticos pueden dar lugar a la larga a un marco conceptual

satifactorio que ligue los tipos mecanisísticos de comprensión posibles en la biología molecular con los fenómenos de niveles superiores característicos tradicionalmente de la biología de poblaciones y la evolución".

LITERATURA CITADA

AVISE, J.C.1989. Gene trees and organismal histories : a phylogenetic approach to population biology. Evolution. 43:1192-1208.

BEINTEMA, J.J. and R.R. CAMPAGNE. 1987. Molecular evolution of rodent insulins. Mol. Biol. Evol. 4: 10-18.

BLANKEN, R.L., L.C. KLOTZ y A.G. HINNEBUSCH. 1982. Computer comparison of new and existing criteria for constructing evolutionary trees from sequence data. J. Mol. Evol. 19:9-91.

BOUMA, J.E. y R.E. LENSKI .1988. Evolution of a bacteria/plasmid association. Nature 335:351-352.

BRODA, P. 1979. Plasmids. Freeman, San Francisco.

CAMPBELL,A. 1981. Evolutionary significance of accessory DNA elements in bacteria. Ann. Rev.Microbiol. 35:55-83.

CAUGANT, D.A., B.R. LEVIN y R.K. SELANDER. 1981. Genetic diversity and temporal variation in the *E. coli* population of a human host. Genetics 98: 467-90.

CHAKRABARTY, A.M., D.A. FRIELLO y L.H. BOPP. 1978. Transposition of plasmid DNA segments specifying hydrocarbon degradation and their expression in various microorganisms. Proc. Natl. Acad. Sci. USA. 75:3109-3112.

DUBOSE, R.F., D. DYKHUIZEN y D.HARTL. 1988. Genetic exchange among natural isolates of bacteria: recombination within the *phoA* gene locus of *Escherichia coli*. Proc. Natl. Acad. Sci. USA 85:7036-7040.

EBERHARD, W.G. 1989. Why do bacterial plasmids carry some genes y not others? Plasmid 21 : 167-174.

EFRON, B. 1982. The jacknife, the bootstrap, and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia.

EVANS, R. 1986. Niche expansion in bacteria: can infectious gene exchange affect the rate evolution? Genetics 113: 775-795.

FELSENSTEIN, J. 1981. Evolutionary trees from gene frequencies and quantitative character: Finding maximum likelihood estimates. Evolution 35: 1229-1242.

FELSENSTEIN, J. 1988. Sex and the Evolution of Recombination. En MICHOD, R.E. y B.R. LEVIN (eds.): The evolution of sex. Sinauer Associates Inc. Publishers. Sunderland MA, pp. 74-86.

FLORES, M., V. GONZALEZ, S. BROM, E. MARTINEZ, D. PIÑERO, D. ROMERO, G. DAVILA y R. PALACIOS. 1987. Reiterated DNA sequences in *Rhizobium* and *Agrobacterium* spp. J. Bacteriol. 169: 5782-5788.

HARTL, D. y D. DYKHUIZEN. 1984. The population genetics of *Escherichia coli*. Ann. Rev. Genet. 18:31-68.

HENDY, M.D., C. LITTLE y D. PENNY. 1984. Comparing trees with pendant vertices labelled. SIAM J. Appl. Math. 44:1054-1065.

HOLMQUIST, R., M. MIYAMOTO y M. GOODMAN. 1988. Higher primate phylogeny- why can't we decide? Mol. Biol. Evol. 5:201-216.

HUDSON, R.R. y N.L. KAPLAN. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111: 147-164.

JOHNSTON, A.W.B., J.L. BEYNON, A.U. BUCHANAN-WOLLASTON, S.M. SETCHELL, P.R. HISCH y J.E. BERINGER. 1978. High frequency of transfer of nodulating ability between strains and species of *Rhizobium*. Nature 276 :634-636.

JUKES, T.H.,y C.H. CANTOR. 1969. Evolution of protein molecules In : MUNRNO, H.N.(ed) Mammalian protein metabolism. Academic Press. New York, pp. 21-123.

KIMURA, M. 1980. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA 78: 454-458.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press.

KURTEN, B. 1959. Rates of evolution of fossil mammals. Cold Spring Harbor Symp. Quant. Biol. 24:205-215.

LAKE, J.A. 1987. A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. Mol. Biol. Evol. 4: 167-191.

LAWRENCE, G.H., D. DYKHUIZEN, R. DUBOSE y D. HARTL. 1989. Phylogenetic analysis using insertion sequence fingerprinting in *Escherichia coli*. Mol. Biol. Evol. 6:1-14.

LEDERBERG, J. 1955. Recombination mechanisms in bacteria. J. Cell. Comp. Physiol. 45:Suppl. 2 pp. 75-107.

LEVIN, B.R. 1981. Periodic selection, infectious gene exchange and the genetic structure of *Escherichia coli* populations. Genetics 99:1-23.

LEVIN, B.R. 1988. The evolution of sex in bacteria. En MICHOD, R.E. y B.R. LEVIN (eds.): The evolution of sex. Sinauer Associates Inc. Publishers. Sunderland MA, pp 194-211.

LEVIN, B.R. y R.E. LENSKI. 1985. Bacteria and phage: a model system for the study of the ecology and co-evolution for hosts and parasites. In: Ecology and Genetics of Host-Parasite Interactions. The Linnean Society of London. pp. 227-242.

LI, W.-H., K.H. WOLFE, J.SOURDIS y P.M. SHARP. 1987. Reconstruction of phylogenetic trees and estimation of divergence times under nonconstant rates of evolution. Cold Spring Harbor Symposia on Quant. Biol.LII: 847-856.

LOW, K.B., y D.D. PORTER. 1978. Modes of gene transfer and recombination in bacteria. Ann. Rev. Genet. 12:249-287.

LYNCH, J.D. 1989. The gauge of speciation, on the frequencies of modes of speciation. In OTTE, D. y J.A. ENDLER (eds): Speciation and its consequences. Sinauer Associates Inc.. Sunderland, MA, pp. 527-556.

MAYNARD SMITH, J. 1989. Evolutionary Genetics. Oxford University Press. pp. 189-194.

MILKMAN, R. y I. CRAWFORD. 1983. Clustered third-base substitutions among wild strains of *Escherichia coli*. Science 221:178-180.

MILLER, R.I. y R.G. WIEGERT. 1989. Documenting completeness, species-area relations, and the species abundance distribution of a regional flora. Ecology 70:16-22.

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

NEI, M. y W.-H. LI. 1979. Mathematical model for studying

genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76: 5296-5273.
NELSON,D.R. y H.W. STROBEL.1987. Evolution of cytochrome P-450 proteins. Mol. Biol. Evol. 4: 572-593.
PAMILO, P. y M. NEI. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5: 568-583.
PIÑERO, D., E. MARTINEZ y R.K. SELANDER. 1988. Genetic diversity and relationship among isolates of *Rhizobium leguminosarum* biovar *phaseoli*. Appl. Environ. Microbiol. 54: 2825-2832.
REANNY, D. 1978. Coupled evolution: adaptive interactions among the genomes of plasmids, viruses, and cells. Intl. Rev. Cytol. Suppl. 8:1-68.
ROBINSON, D.F. y L.R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci.53:131-147.
ROHLF, F.J. y M.C. WOOTEN. 1988. Evaluation of the restricted maximum-likelihood method for estimating phylogenetic trees using simulated allele-frequency data. Evolution. 42:581-595.
ROWBURY, R. 1977. Bacterial plasmids with particular reference to their replication and transfer properties. Prog. Biophys. Mol. Biol. 31: 271-317.
SAITOU,N. y M.NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol.Evol. 4:406-425.
SAITOU, N. y M. NEI. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J. Mol. Evol. 24:189-204
SAWYER, S. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. 6:526-538.
SCHOFIELD, P.R., A.H.GIBSON, W.F. DUDMAN y J.M. WATSON. 1987. Evidence for genetic exchange and recombination of *Rhizobium* symbiotic plasmids in a soil population. Appl. Env. Microbiol. 53:2942-2947.
SELANDER, R.K. y B.R. LEVIN. 1980. Genetic diversity and structure in *Escherichia coli* populations. Science 210: 545-547.
SELANDER, R.K., R. McKINNEY, T. WHITTAM, W. BIBB, D.BRENNER, F. NOLTE y P. PATTISON. 1985. Genetic structure of populations of *Legionella pneumophila*. J. Bacteriol. 163: 1021-1037.
SELANDER, R.K., D.CAUGANT y T.WHITTAM. 1987. Genetic structure and variation in natural populations of *Escherichia coli*. In *Escherichia coli* and *Salmonella typhimurium*, Cellular and Molecular Biology. J.L. INGRAM et al. (eds), ASM Publications, Washington D.C..
SLATKIN, M. 1985. Gene flow in natural populations. Ann. Rev. Ecol. Syst. 16:393-430.
SLATKIN, M. 1987. Gene flow and the geographic structure of natural populations. Science 236: 787-792.
SLATKIN, M. y W.P. MADDISON. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics, En prensa.
SOKAL, R.R. y C.D. MICHENER. 1958. A statistical mehtod for

evaluating systematic relationships. University of Kansas Sci. Bull. 28: 1409-1438.
SOKAL, R.R. y F.J. ROHLF. 1981. Biometry. Freeman y Co., San Francisco.
SOURDIS,J. y C. KRIMBAS. 1987. Accuracy of phylogenetic trees from DNA sequence data. Mol. Biol. Evol. 4:159-166.
SOURDIS,J. y M. NEI. 1988. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic trees. Mol. Biol. Evol. 5: 298-311.
STEPHENS, J.C. 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. 2:539-556.
SYVANEN, M. 1987. Molecular clocks and evolutionary relationships: possible distortions due to horizontal gene flow. J. Mol. Evol. 26:16-23.
TATENO,Y., M. NEI y F. TAJIMA. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J.Mol. Evol. 18: 387-404.
TATENO,Y., y F.TAJIMA. 1986. Statistical properties of molecular tree construction methods under the neutral mutation model. J.Mol.Evol. 23:354-361.
WILSON, A.C.,S.CARLSON y T.WHITE. 1977. Biochemical evolution. Ann. Rev.Biochem. 46:573-639.
WOESE,C.R., J.GIBSON y G.E.FOX. 1980. Do genealogical patterns in purple photosynthetic bacteria reflect interspecific gene transfer? Nature 283: 212-214.
WOESE,C.R. 1987. Bacterial evolution. Microbiol.Rev. 51:221-271.
YOUNG, J.P.W. y M. WEXLER. 1988. Sym plasmid and chromosomal genotypes are correlated in field populations of *Rhizobium leguminosarum*. J. Gen. Microbiol. 134: 2731-2739.