QÜESTIIÓ, Vol 20, 2 pp. 327-346, 1996

SOBRE LA SIMULACIÓN DE PROCESOS DE EVOLUCIÓN MOLECULAR: CONSIDERACIONES SOBRE LA DERIVACIÓN Y CONTRASTACIÓN DE UN ESTIMADOR DE LA VARIANZA DE LA DIVERGENCIA NUCLEOTÍDICA A PARTIR DE FRAGMENTOS DE RESTRICCIÓN

SANTIAGO F. ELENA, ANDRÉS MOYA y
FERNANDO GONZÁLEZ CANDELAS

Una de las áreas de la Biología que ha experimentado mayor expansión en los últimos años es la investigación de procesos evolutivos mediante la aplicación de técnicas de la Biología Molecular. La puesta en marcha de programas de obtención de información cartográfica de genes y de secuencias de los mismos no ha hecho más que aumentar esta tendencia. La investigación de procesos evolutivos lleva emparejada el contraste de hipótesis alternativas, tales como el orden de división de los linajes en una reconstrucción filogenética o las estimas de distancias entre los nodos de un árbol filogenético. Existen modelos evolutivos que, bajo supuestos más o menos restrictivos, han permitido la construcción de tests paramétricos de contraste de hipótesis. Una de las dificultades que se encuentra en el desarrollo de estos tests es la derivación de las propiedades de los estadísticos correspondientes. En tales situaciones es muy frecuente recurrir a la simulación de procesos de evolución para, así, contrastar la bondad de los estadísticos. En este trabajo exponemos la derivación de un estimador de la varianza de la divergencia nucleotídica a partir de la comparación de los fragmentos producidos por la digestión con endonucleasas de restricción del DNA de especies descendientes de un ancestro común y los resultados de las simulaciones realizadas para comprobar su bondad, prestando especial atención al efecto que tienen sobre los resultados de la simulación distintos parámetros inicialmente considerados no relevantes y la importancia de establecer controles rigurosos e independientes sobre las simulaciones.

Santiago F. Elena, Andrés Moya y Fernando González Candelas. Departament de Genètica i Servei de Bioinformàtica. Facultat de Biología. Universitat de València «Estudi General». Dr. Moliner 50, 46100 Burjassot, València.

- Article rebut el setembre de 1995.
- Acceptat el setembre de 1996.

On the simulation of molecular evolutionary processes: considerations on the derivation and confirmation of an estimator for the variance of nucleotide divergence estimated from restriction fragments

Keywords: Simulation, nucleotide divergence, restriction fragment data, parameter estimation, delta method.

1. INTRODUCCIÓN

A lo largo del proceso evolutivo todo par de especies descendientes de un ancestro común acumulan cambios en sus moléculas de DNA que son, aproximadamente, proporcionales al tiempo transcurrido desde su divergencia. El estudio de estos cambios entre pares de secuencias contemporáneas es la base de una nueva metodología para la reconstrucción de procesos filogenéticos, disciplina conocida como Sistemática Molecular. Además, en cualquier población de individuos contemporáneos de cualquier especie se producen de forma constante nuevas mutaciones, material básico para la evolución, que quedan igualmente plasmadas como cambios en las moléculas de DNA de los distintos individuos. El estudio de las relaciones entre las distintas poblaciones de una especie, los patrones de intercambio genético entre ellas, la difusión de aquellas variantes especialmente favorables bajo circunstancias ambientales particulares, etc., estudios que habitualmente corresponden al ámbito de la Genética de Poblaciones, también pueden basarse en la comparación de los patrones de variación entre moléculas de DNA al nivel o niveles (individual, poblacional, regional, etc.) adecuados. De esta forma, la relación entre la Genética de Poblaciones y la Teoría de la Evolución se amplía desde la fundamentación teórica y formal que representa la primera para la segunda, hasta compartir una misma metodología experimental sobre la que contrastar y basar sus avances. Una excelente introducción a las distintas ramas de la Biología, tanto académicas como aplicadas, que utilizan estas técnicas de estudio de la variación genética a nivel molecular se halla en Avise (1994).

Para analizar estos cambios en las secuencias de DNA, el investigador dispone de diversas técnicas experimentales, de complejidad y coste crecientes aproximadamente a medida que aumenta la calidad de la información que proporcionan. Así, la secuenciación, es decir, la determinación de la secuencia de nucleótidos de parte de la molécula de DNA, proporciona el mayor grado de información posible, pero a costa de un mayor esfuerzo metodológico y económico y de una reducción muy drástica en la proporción de genoma estudiado. Con el fin de reducir los costes, así como para aumentar la fracción de genoma analizado aún a costa de perder parte de la información, se disponen de técnicas alternativas basadas, muchas de ellas, en la digestión con enzimas de restricción.

Una enzima de restricción es una proteína que reconoce una secuencia específica de DNA, normalmente de pequeña longitud (4 ó 6 nucleótidos) provocando un corte en la molécula de DNA nativa. Los fragmentos generados por este corte pueden ser separados mediante electroforesis en función de su tamaño y ser comparados con los producidos mediante el mismo procedimiento en otros individuos de la misma o distintas especies. Con la combinación adecuada de enzimas de restricción en una misma reacción de digestión es posible, además, establecer los puntos específicos de corte de cada uno de ellos a lo largo del genoma analizado. La primera de las técnicas se conoce como análisis del polimorfismo en la longitud de los fragmentos de restricción (RFLPs) mientras que la segunda se conoce como análisis de los sitios de restricción.

Otras técnicas de análisis de la variación en las secuencias de DNA que son en algunos aspectos asimilables a las anteriores utilizan secuencias más o menos largas de cebadores para amplificar fragmentos de la molécula original mediante la reacción en cadena de la polimerasa (PCR). La separación de estos fragmentos permite, de nuevo, la comparación entre distintas moléculas de DNA. Una de estas técnicas, que emplea cebadores relativamente cortos (10 pares de bases) que sirven para amplificar fragmentos aleatorios del DNA, es conocida como RAPDs-PCR (Hadrys, Balick y Schierwater, 1992). Esta técnica es una de las de utilización más frecuente en los últimos años.

Las diferencias genéticas obtenidas mediante RAPDs-PCR no son susceptibles del análisis cuantitativo necesario para obtener medidas de distancias filogenéticas entre taxones con divergencias medias (al menos intergenéricas) debido a diversos problemas ampliamente discutidos por Clark y Lanigan (1993) y por Lynch y Milligan (1994). No obstante, es posible realizar comparaciones filogenéticas entre poblaciones de una misma especie o entre especies muy próximas entre sí, en el ámbito de lo que Avise (1994) define como Filogeografía. En lo fundamental, la estimación de la divergencia nucleotídica a partir del análisis de bandas de RAPDs compartidas entre dos genomas es formalmente similar a la propuesta por Nei y Li (1979) para estimar ese mismo valor mediante el análisis de fragmentos de restricción. Esta similitud fue puesta de manifiesto por Clark y Lanigan (1993).

Dada la incertidumbre asociada al proceso de estimación de la divergencia nucleotídica, se hace necesario disponer de una buena estimación de la varianza del estimador de la divergencia nucleotídica empleado. De los diversos estimadores propuestos para datos obtenidos mediante análisis de los fragmentos de restricción, por comparación sólo del tamaño de los mismos y no de los sitios donde se producen los cortes, el más frecuentemente empleado es el desarrollado por Nei y Li (1979). Nei y Miller (1990) propusieron un método de remuestreo mediante «bootstrap» para obtener una estimación de la varianza de ese estimador, mientras que González-Candelas, Elena y Moya (1995) han propuesto el uso de un estimador analítico.

2. EL MODELO DE CAMBIO EVOLUTIVO EN LOS SITIOS DE RESTRICCIÓN

Nei y Li (1979) desarrollaron un método para estimar el número, d, de sustituciones nucleotídicas por sitio entre dos secuencias de DNA cuando los datos de que se dispone son los fragmentos de restricción. Para ello se basaron en el siguiente modelo de cambio evolutivo de los sitios de restricción.

Sea $n(t)$ el número de sitios de restricción en la molécula (o porción de ella) de DNA considerada en el instante t y sea $n(0) = n_0$. Se supone

(1) que el contenido esperado G+C permanece constante y

(2) que la sustitución de nucleótidos se produce de forma aleatoria según un proceso Poisson con tasa de sustitución λ por unidad de tiempo (año o generación).

A medida que transcurre la evolución, algunos de los sitios de restricción originales desaparecerán mientras que aparecerán otros nuevos. Denotemos por $n_1 t$ y $n_2 t$ respectivamente esos valores. En ese caso podemos escribir $n(t) = n_1(t) + n_2(t)$. La probabilidad de que un sitio original permanezca inalterado al cabo de un tiempo t viene dada por $P = e^{-r\lambda t}$, por lo que la esperanza de $n_1(t)$ es $n_0 \cdot e^{-r\lambda t}$.

Para obtener la esperanza de $n_2(t)$ pensemos que el enzima de restricción considerado reconoce una secuencia de r nucleótidos (habitualmente r = 4 ó 6). La probabilidad de que esta secuencia haya sufrido uno o más cambios en el tiempo t es $1 - P$ y sea a la probabilidad de que la nueva secuencia generada sea un sitio de restricción. El valor de a está relacionado con la distribución de frecuencias en el equilibrio de los cuatro nucleótidos en la molécula de DNA correspondiente y en la proporción con que aparecen los mismos en la secuencia diana de la enzima de restricción. Si en la molécula de DNA considerada existen m_T secuencias posibles de longitud r, entonces la esperanza de $n_2(t)$ es $m_T a(1 - P)$. En ese caso, el valor esperado, $\mathrm{E}[n]$, de $n(t)$ es

$$\mathrm{E}[n] = n_0 P + m_T a(1 - P). \tag{1}$$

Su varianza puede obtenerse teniendo en cuenta que n_1 se distribuye binomialmente y que n_2 sigue una distribución Poisson, siendo ambos valores independientes:

$$\mathrm{Var}[n] = n_0 P(1 - P) + m_T a(1 - P). \tag{2}$$

Consideremos ahora la divergencia entre dos linajes evolutivos o poblaciones X e Y. Asumimos que sus DNAs se derivan de una secuencia ancestral común a partir del instante 0. Sean n_{X1} y n_{X2} el número de sitios de restricción ancestrales y el de

nuevos sitios, respectivamente, en el linaje X, con $n_X = n_{X1} + n_{X2}$, y sean n_{Y1}, n_{Y2} y n_Y los valores correspondientes en el linaje Y. Sea n_{XY} el número de sitios idénticos compartidos por los dos linajes. Dada la baja probabilidad de que se produzcan *de novo* y de forma independiente dos nuevos sitios de restricción idénticos en los linajes X e Y, consideraremos que todos los sitios compartidos se derivan de sitios presentes ya en la secuencia ancestral común. Bajo este supuesto n_{XY} sigue una distribución binomial cuya media y varianza vienen dadas por n_0P^2 y $n_0P^2(1-P^2)$, respectivamente.

Consideremos ahora la relación entre los fragmentos de restricción compartidos entre las dos especies. Para que un fragmento de DNA se conserve a lo largo de t generaciones se precisan dos condiciones:

(1) los dos sitios flanqueantes al fragmento deben permanecer inalterados y

(2) no debe aparecer ningún nuevo sitio de restricción en su interior.

La probabilidad del primer suceso es obviamente P^2, y la del segundo es $(1-b)^{(m-r+1)}$, donde $b = a(1-P)$ y m es el número de nucleótidos en ese fragmento. Teniendo en cuenta que para que el fragmento siga siendo compartido las anteriores condiciones deben cumplirse en los dos linajes considerados, la proposición de fragmentos compartidos por ambos será

$$F = \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} p^4 (1-b)^{2(m_i - r+1)} \tag{3}$$

Ahora bien, en la práctica no puede aplicarse esta fórmula porque se desconocen tanto n_0 como m_i. No obstante, bajo ciertos supuestos simplificadores adicionales (Nei y Li, 1979), se puede derivar la siguiente aproximación

$$F \approx \frac{P^4}{3-2P} \tag{4}$$

Usando $P = e^{-r\lambda t}$ y $d = 2\lambda t$, se establece una relación entre F y d. El número d representa el número esperado de sustituciones nucleotídicas entre los dos linajes al cabo de un tiempo t.

El estimador máximo verosímil de F es

$$\hat{F} = \frac{2m_{XY}}{m_X + m_Y} \tag{5}$$

donde m_X y m_Y representan el total de fragmentos de restricción observados en las secuencias X e Y, respectivamente, y m_{XY} representa el número de tales fragmentos que son compartidos por ambas secuencias.

Para poder realizar inferencias sobre la igualdad o no de dos estimaciones de d, es necesario disponer de una estimación de sus errores respectivos. En el presente trabajo, en primer lugar, describimos el desarrollo de una expresión analítica aproximada del estimador de $\mathrm{Var}(\hat{d})$, a partir de un estimador ya conocido de la varianza de $\hat{F}$ (Nei y Tajima, 1981) y, en segundo lugar, comprobamos la validez de esta expresión mediante una simulación numérica, comparando los resultados con los obtenidos para otros estimadores.

Nei y Tajima (1981) derivaron la siguiente expresión para la varianza muestral de $\hat{F}$:

$$\widehat{\mathrm{Var}}(\hat{F}) = \frac{1}{\overline{m}}\left\{\hat{F}\left(1-\hat{F}\right) - \hat{F}^2\left(1-\sqrt{\hat{F}}\right)\left[1+\frac{1}{2}\left(1-\sqrt{\hat{F}}\right)\right]\right\} \tag{6}$$

donde $\overline{m} = \dfrac{m_X + m_Y}{2}$ es el número promedio de fragmentos de restricción observados en las dos secuencias analizadas. Este valor es estimador del número de fragmentos de restricción presentes originalmente en la secuencia ancestral.

3. DESARROLLO DEL ESTIMADOR DE VARIANZA DE $\hat{d}$

Asumimos que se ha obtenido una estimación empírica de F a partir de la digestión con enzimas de restricción de dos secuencias X e Y y que los fragmentos obtenidos son separados y comparados por tamaños (lo que nos permite obtener los valores m_X, m_Y y m_{XY} antes indicados). A partir de esa estimación, utilizando la ecuación (4) se obtiene por iteración una estimación de P, con lo que finalmente se puede estimar la divergencia nucleotídica, $\hat{d}$, entre ambas secuencias.

Los detalles de la derivación del estimador de la varianza del estimador de d pueden encontrarse en González-Candelas, Elena y Moya (1995). No obstante, delineamos aquí las ideas generales empleadas. Aplicamos en varias ocasiones la fórmula aproximada de Fisher para la obtención de la varianza de un parámetro o método delta (Fisher, 1925), reteniendo hasta el momento de tercer orden en la expansión de Taylor de (1) para aumentar la precisión de la estimación. Utilizamos la expresión (4) antes mencionada para obtener una estimación de la varianza del estimador de F a partir de los datos empíricos antes mencionados. Además, tenemos en cuenta que $\mathrm{E}(m_X) = \mathrm{E}(m_Y) = \overline{m}P$, que $\mathrm{E}(m_{XY}) = \mathrm{E}(m_{XY}\, m_X) = \mathrm{E}(m_{XY}\, m_Y) = \overline{M}P^2$, que $\mathrm{Cov}(m_{XY}\, m_X) = \mathrm{Cov}(m_{XY}\, m_Y) = \overline{m}P^2(1-P)$, y que $\mathrm{Cov}(m_X, m_Y) = 0$, siendo P la probabilidad de que un fragmento dado aparezca compartido por un par de secuencias y $\overline{m}$ el número de fragmentos de restricción en la secuencia ancestral (Nei y Li, 1979).

Con todo ello se obtiene la siguiente expresión para el estimador de la varianza del estimador de d:

$$\widehat{\mathrm{Var}}(\hat{d}) = \frac{(3-2\hat{P})^4}{9r^2\hat{P}^8(2-\hat{P})^2}\left[\mathrm{Var}(\hat{F}) - \frac{72-117\hat{P}+64\hat{P}^2-12\hat{P}^3}{6\hat{P}^4(2-\hat{P}^2)}\mu_3(\hat{F})\right] \tag{7}$$

donde

$$\mu_3(\hat{F}) = \frac{\hat{F}(1-\hat{F})(1-2\hat{F})}{\overline{m}^2} - \frac{\hat{F}^3\left\{\overline{m}\sqrt{\hat{F}}\left(1-\sqrt{\hat{F}}\right)\left(1-2\sqrt{\hat{F}}\right)+\sqrt[4]{\overline{m}\left(1-\sqrt{\hat{F}}\right)}\right\}}{4\overline{m}^3} -$$

$$- \frac{3\hat{F}^2\left(1-\overline{m}\sqrt{\hat{F}}-2\overline{m}\hat{F}+2\overline{m}^2\sqrt{\hat{F}^3}\right)}{4\overline{m}^4} \tag{8}$$

Para ciertas aplicaciones nos bastará con tener una estimación de la varianza del estimador de d, pero en otras ocasiones estamos interesados en realizar un constraste de hipótesis directamente sobre el valor estimado. Para la construcción del test es imprescindible conocer la distribución del estimador, y no sólo su varianza, lo que es imposible. No obstante, podemos suponer que las estimaciones de d se comportan de forma asintóticamente normal, por ser transformaciones suaves de estimadores de máxima verosimilitud. Supongamos que se han calculado las estimaciones $\hat{d}_a$ y $\hat{d}_b$ y sus varianzas respectivas a partir de $m_a (m_a = m_X + m_Y - m_{XY}$, si consideramos las secuencias X e Y según lo indicado anteriormente) y m_b fragmentos, respectivamente, entre pares de secuencias independientes. En ese caso, el estadístico

$$t'_s = \frac{\hat{d}_a - \hat{d}_b}{\sqrt{\dfrac{\widehat{\mathrm{Var}}(\hat{d}_a)}{m_b} + \dfrac{\widehat{\mathrm{Var}}(\hat{d}_b)}{m_a}}} \tag{9}$$

se distribuye aproximadamente como una t de Student con $m_a + m_b - 2$ grados de libertad.

4. COMPROBACIÓN POR SIMULACIÓN DE LA BONDAD DEL ESTIMADOR DE $\widehat{\mathrm{Var}}(\hat{d})$

Para comprobar la bondad del procedimiento de derivación y para comprobar las estimaciones de la varianza del estimador de divergencia nucleotídica obtenidas

por los análisis de fragmentos de restricción, se hicieron simulaciones siguiendo el procedimiento descrito por Li (1981). Junto a la evaluación del estimador desarrollado, hemos procedido a comprobar la bondad tanto del estimador de divergencia nucleotídica basado en sitios de restricción (Nei y Li, 1979) como del calculado directamente sobre la secuencia nucleotídica (Jukes y Cantor, 1969). Al realizar estas comparaciones pretendíamos establecer controles adicionales sobre el proceso que se estaba simulando, dada la conocida validez de estos estimadores.

Se simuló la evolución de una secuencia de DNA que da lugar a dos secuencias derivadas para distintos valores de la tasa de divergencia. En este modelo solamente se ha considerado la aparición de mutaciones por sustitución nucleotídica y no por deleciones o inserciones. En la simulación se consideró una única tasa de sustitución nucleotídica, la misma para cambios transicionales que para transvesiones, por lo que se sigue el modelo evolutivo propuesto por Jukes y Cantor (1969) para considerar la superposición de sustituciones en una misma posición. Este modelo, sólo es aceptable para valores bajos de divergencia ($d < 0.1$), lo que corresponde con las condiciones para las que es adecuado estimar la divergencia a partir de fragmentos de restricción (Nei y Li, 1979; Li, 1981). Tras la simulación del proceso evolutivo, se procedió a simular la digestión con enzimas de restricción y a comparar los fragmentos resultantes de las mismas.

Los análisis de simulación se basaron en tres secuencias nucleotídicas aleatorias de longitudes 1000, 10000 y 100000 pares de bases respectivamente, de composición nucleotídica equiprobable. Las tasas de evolución, $2\lambda t$, variaron en el intervalo 0.002 y 0.10 sustituciones por nucleótido y unidad de tiempo. Para cada valor de la tasa de evolución y longitud del genoma, se realizaron 2000 réplicas. Todo el proceso de simulación fue implementado en un programa escrito en Pascal estándar que se ejecutó en una estación de trabajo DEC-AXP 3000-400 en entorno Open VMS.

Se simularon dos procesos de digestión diferentes, uno empleando un enzima que reconoce dianas de 4 nucleótidos ($r = 4$) y otro con diez enzimas que reconocen dianas de 6 nucleótidos ($r = 6$). Las secuencias diana de las enzimas de restricción empleadas se generaron también aleatoriamente al principio de cada simulación. El mismo procedimiento de simulación se realizó también manteniendo constantes las enzimas de restricción empleadas en todos los casos, obteniéndose resultados semejantes a los anteriores y que no son detallados aquí. La divergencia entre cada par de secuencias se estimó por tres vías:

(i) a partir de la longitud de los fragmentos obtenidos en ambas restricciones empleando las ecuaciones descritas en el apartado 2,

(ii) a partir de datos de sitios de restricción usando la ecuación 5.42 de Nei (1987, página 101) y

(iii) a partir de las secuencias nucleotídicas empleando el estimador de Jukes y Cantor (1969). Las varianzas para cada estimación de divergencia entre pares de secuencias se obtuvieron de acuerdo con la ecuación (7) para datos de fragmentos de restricción, según la ecuación 5.45 de Nei (1987, página 101) para datos de sitios de restricción y según la ecuación 5.4 de Nei (1987, página 66) para los datos directos de secuencia nucleotídica.

5. RESULTADOS DE LAS SIMULACIONES

Los valores de las divergencias promedio estimados directamente a partir de los datos de secuencia, con la corrección de Jukes-Cantor, y las estimaciones de sus varianzas se encuentran en la figura 1. Los datos estimados mediante el uso de sitios de restricción se muestran en las figuras 2 (para un enzima con $r = 4$) y 3 (para 10 enzimas con $r = 6$), mientras que los valores estimados a partir de los fragmentos de restricción se muestran en las figuras 4 y 5, respectivamente. En los dos últimos casos se muestra en cada figura tanto el valor promedio de las 2000 varianzas proporcionadas por las estimaciones en cada réplica, como la varianza de las 2000 estimaciones de divergencia nucleotídica calculadas usando sitios y fragmentos de restricción para cada valor de la tasa de evolución.

La fiabilidad del proceso evolutivo simulado para todas las tasas y longitudes de secuencia puede ser apreciada a partir de las estimaciones de divergencia nucleotídica obtenidas mediante el estimador de Jukes y Cantor (figura 1). En la figura se pone de manifiesto la utilidad del método de Monte-Carlo empleado para comprobar la validez de la expresión 5.4 dada por Nei (1987, página 66) para estimar la varianza del estimador de divergencia nucleotídica a partir de las secuencias. Teniendo en cuenta la validez del proceso evolutivo simulado y del método de contraste empleado, las estimaciones de divergencia evolutiva a partir de sitios y fragmentos de restricción muestran las siguientes características.

Para datos de sitios de restricción, se obtuvieron mejores estimaciones usando un sólo enzima de $r = 4$ que usando 10 enzimas de $r = 6$. Esta observación es válida para todas las longitudes de secuencia empleadas en la simulación, aunque generalmente se obtienen mejores estimaciones empleando secuencias más largas. Empleando un enzima de $r = 4$, y para $L = 1000$ y $L = 10000$ con $2\lambda t \leq 0.07$, generalmnte se obtiene una ligera sobreestimación de $\mathrm{Var}(\hat{d})$ empleando la ecuación 5.45 de Nei (1987). Para $L = 100000$ y para cualquier tasa de evolución, así como para $L = 10000$ y tasas $2\lambda t \geq 0.07$, es habitual encontrar subestimaciones. Cuando se usan 10 enzimas de $r = 6$, las estimaciones de $\hat{d}$ son aceptables únicamente cuando

$L = 100000$, obteniéndose una clara subestimación para los demás valores de L. Estas subestimas no pueden ser atribuidas a un error en la simulación, como se demostró en el párrafo anterior al coincidir los valores estimados mediante secuencia con la corrección de Jukes y Cantor. No existe un patrón claro para las correspondientes varianzas, donde se pueden observar pequeñas desviaciones entre las estimaciones y los valores obtenidos en las simulaciones para todas las longitudes y tasas de evolución empleadas.

El uso de datos de longitud de fragmentos de restricción para estimar divergencias nucleotídicas presenta algunos problemas. Primero, las divergencias nucleotídicas pueden ser estimadas correctamente a partir de la combinación adecuada de longitudes de secuencia y número de fragmentos. En nuestro caso, esto se consigue usando o bien $L = 10000$ y un solo enzima de $r = 4$ o bien $L = 100000$ y diez enzimas de $r = 6$. Todas las demás combinaciones de longitud de secuencia y enzimas originan series subestimaciones de la divergencia simulada. Para secuencias de longitud pequeña o para un número de fragmentos generados también pequeño, esto puede ser debido al pequeño número de fragmentos compartidos que aparecen. Para el caso de secuencias largas, la explicación reside en la redundancia de fragmentos no homólogos con igual longitud generados cuando $2\lambda t \geq 0,02$. Para los dos casos en los que las estimas de divergencia fueron aceptables ($L = 1000, r = 4$ y $L = 100000, r = 6$), las estimas de varianza obtenidas usando la ecuación (7) siempre subestiman las varianzas obtenidas mediante las réplicas de Monte-Carlo (figuras 4 y 5). Esto produce un aumento en la probabilidad de error de tipo I si se acepta que las varianzas obtenidas a partir de las 2000 estimas de d son una buena aproximación al valor real de la varianza.

En resumen, para las tasas de divergencia analizadas y el modelo evolutivo empleado, la estimación de la divergencia mediante el estimador de Jukes-Cantor para datos de secuencia es prácticamente exacta. Para datos de sitios de restricción, las estimaciones son adecuadas cuando se emplea un enzima de $r = 4$ y no lo son cuando se usan 10 enzimas con $r = 6$ excepto para $L = 100000$. El mismo patrón se observa para las estimaciones de divergencia cuando se emplean las longitudes de los fragmentos originados en las digestiones, si bien con una ligera pérdida de precisión respecto a la estimación con sitios. Para las estimaciones de las varianzas de los correspondientes estimadores se observa un buen ajuste en el caso de sitios de restricción pero una ligera sobrestima cuando se emplea el estimador desarrollado por González-Candelas, Elena y Moya (1995) para el caso de fragmentos de restricción.

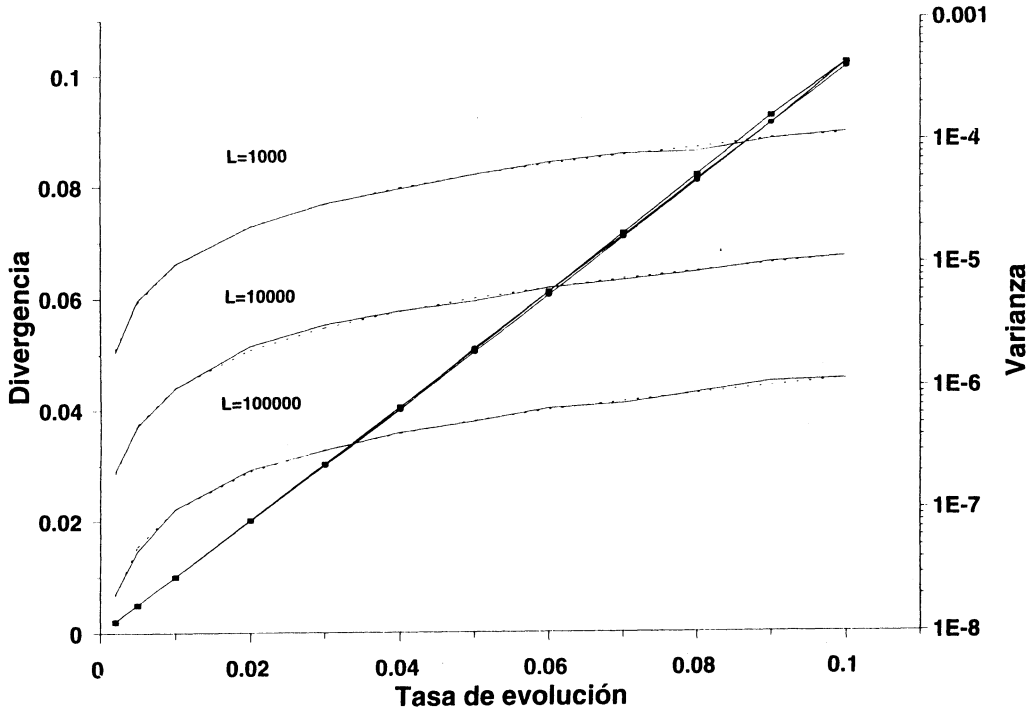

Figura 1

Resultados de la simulación para el análisis directo de la secuencia nucleotídica utilizando el método de Jukes-Cantor.

(a) En el eje de ordenadas principal se muestran las divergencias estimadas para los tres valores de longitud de secuencia considerados (L = 1000, L = 10000 y L = 100000). En este caso no se muestra el valor esperado de la divergencia por ser prácticamente coincidente con el obtenido en las simulaciones.

(b) En el eje de ordenadas secundario (escala logarítmica) se muestran las estimaciones de las varianzas del estimador de divergencia nucleotídica según la ecuación 5.4 dada por Nei (1987). En todos los casos se dan los resultados de la simulación para las tres longitudes de secuencia estudiadas (líneas continuas) y junto a ellas los valores de varianza muestral obtenidos en las correspondientes simulaciones (líneas punteadas).

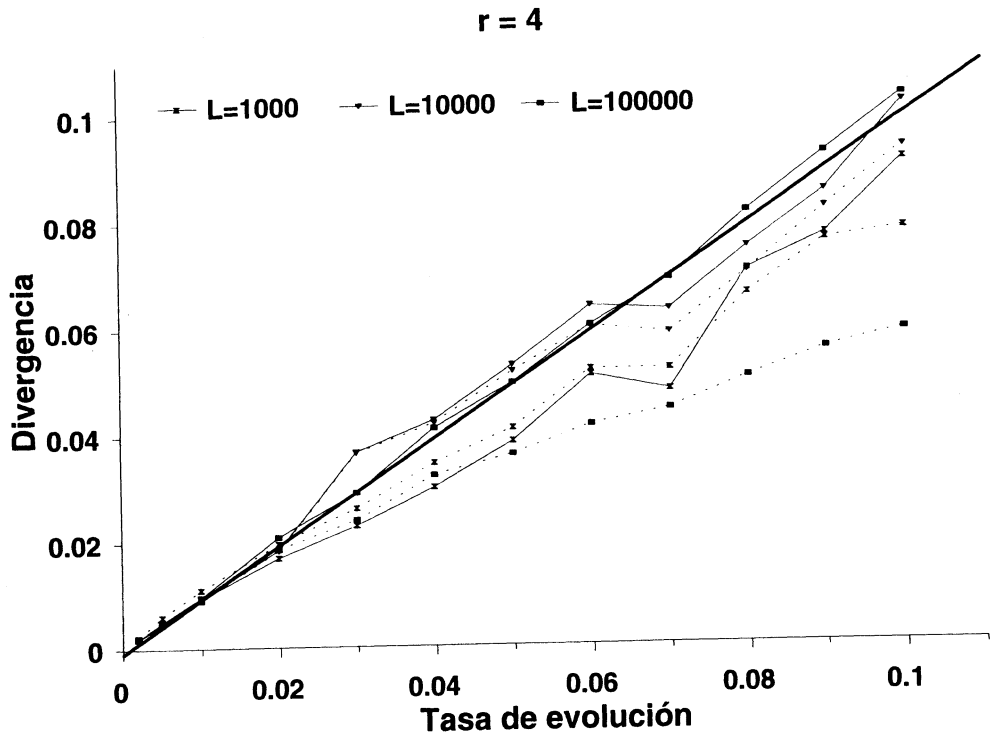

Figura 2

Resultados de la simulación para el análisis de sitios de restricción empleando 1 solo enzima que reconoce dianas de 4 nucleótidos.

(a) Líneas continuas: divergencias estimadas mediante el análisis de sitios de restricción.

(b) Líneas punteadas: divergencias estimadas mediante el análisis de fragmentos de restricción. En todos los casos se muestran los resultados para las tres longituds genómicas estudiadas. La línea gruesa bisectriz indica el valor esperado de la divergencia nucleotídica.

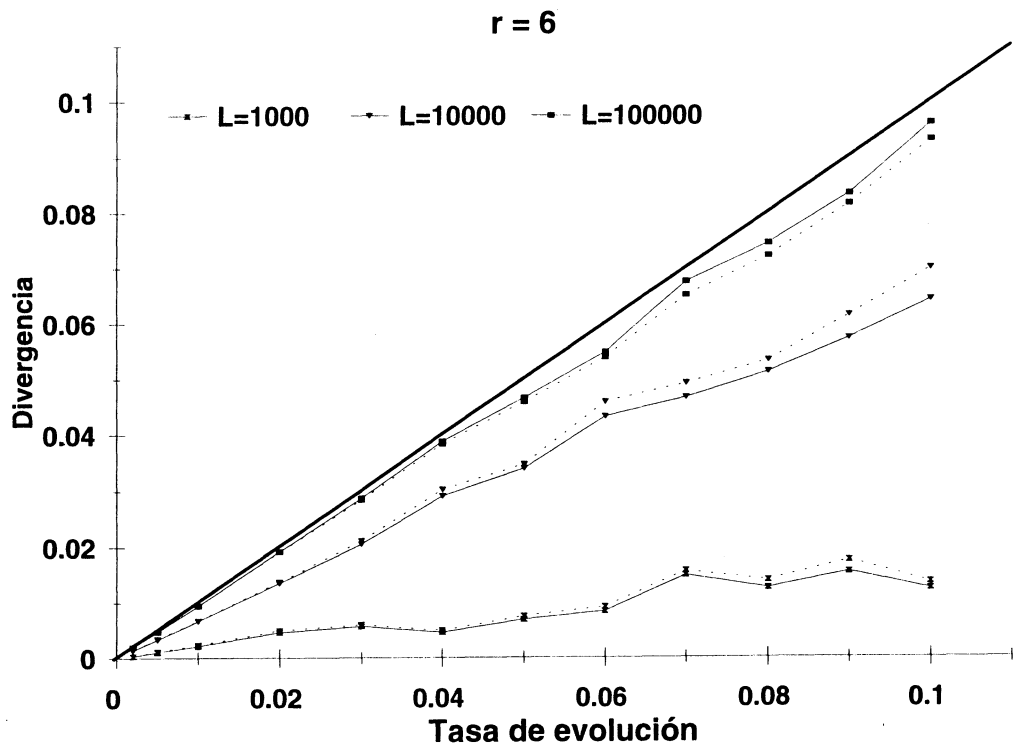

Figura 3

Igual que la figura 2, pero empleando 10 enzimas que reconocen una diana de 6 nucleótidos de longitud.

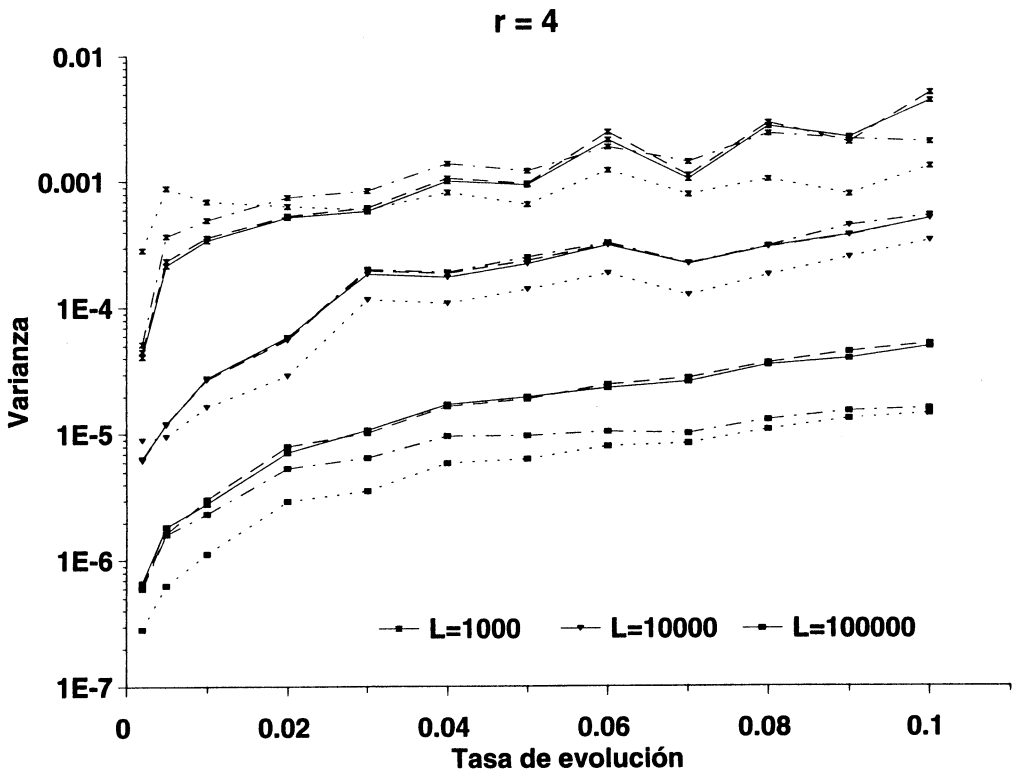

Figura 4

Estimaciones de la varianza de los estimadores de divergencia para sitios y fragmentos de restricción según la ecuación 5.45 de Nei (1987) y la ecuación (7), respectivamente tras la simulación por el método de Monte-Carlo. Para cada longitud de genoma considerado, los puntos unidos por una línea continua representan las estimaciones de la varianza para el estimador de divergencia a partir de sitios de restricción obtenidos a partir de las simulaciones y la línea discontinua el promedio de las 2000 estimaciones correspondientes según la ecuación 5.45 de Nei (1987). De igual forma, las líneas que alternan puntos y rayas corresponden a la varianza entre las 2000 réplicas del estimador de divergencia nucleotídica a partir de fragmentos de restricción y las líneas punteadas corresponden al promedio de las 2000 estimaciones de la varianza según la ecuación (7).

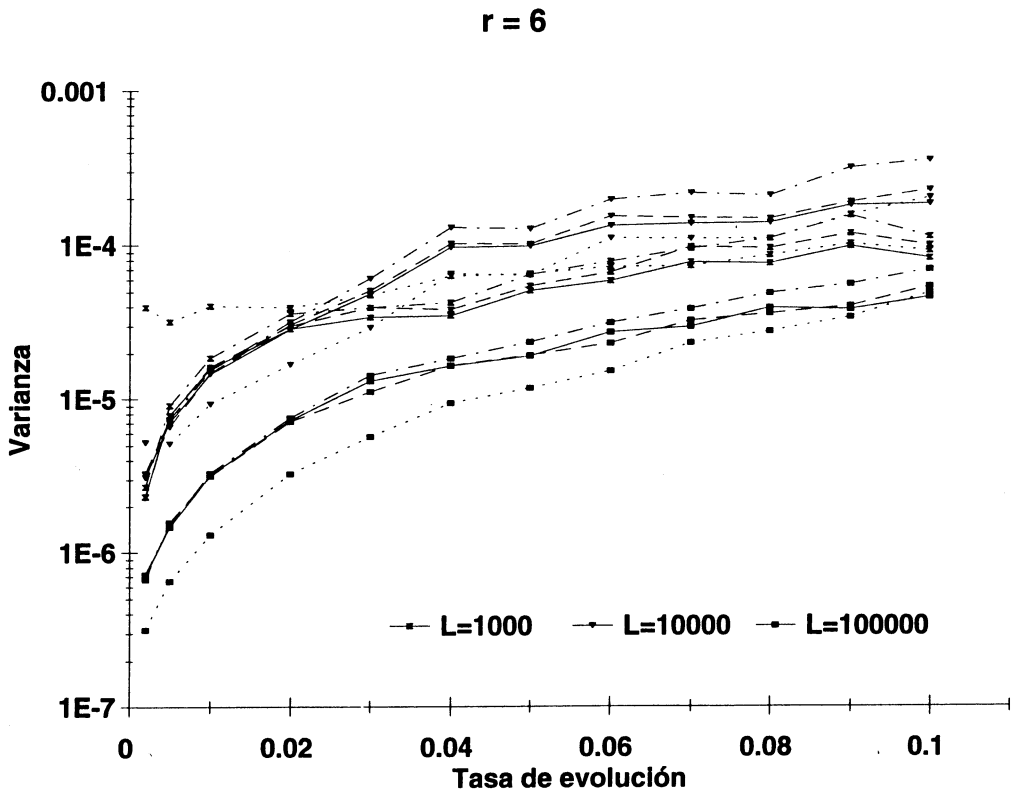

Figura 5

Igual que la figura 4, pero empleando 10 enzimas que reconocen una diana de 6 nucleótidos de longitud.

6. DISCUSIÓN

A lo largo de este trabajo hemos podido comprobar el efecto que tienen parámetros que aparentemente no intervienen en las expresiones a comprobar, como la longitud de la secuencia analizada, sobre los resultados de la simulación. A medida que aumenta la longitud de la secuencia de DNA, por ejemplo, disminuye el ajuste entre las divergencias observadas y esperadas tanto con datos de sitios como de fragmentos de restricción. Esta observación no es predecible a partir de las expresiones analíticas correspondientes y sin un análisis amplio de los posibles valores de L. Por otra parte, hemos podido comprobar la fiabilidad de las simulaciones gracias a la introducción de un control adicional, representado en este caso por la estimación de la divergencia a partir de la secuencia nucleotídica directa usando el método de Jukes-Cantor. Estas consideraciones nos llevan a destacar la necesidad de, por una parte, incorporar controles exhaustivos a lo largo de todo el proceso de simulación y, por otra, de explorar de la forma más amplia posible todos aquellos parámetros que intervienen en la simulación con independencia de que directa o indirectamente puedan estar influyendo en las expresiones estudiadas.

El estimador de la varianza del estimador de divergencia nucleotídica a partir de fragmentos de restricción aquí analizado resulta en una ligera sobrestima de la varianza obtenida mediante la simulación por Monte-Carlo, por lo que se reduce la probabilidad de cometer errores de tipo II en los contrastes realizados utilizando esas estimaciones. La utilización del análisis de la divergencia nucleotídica mediante fragmentos de restricción debe quedar restringida a aquellos casos en que los valores de divergencia sean bastante bajos, debiendo realizarse sobre secuencias lo más largas posibles ($>$ 100 kilobases). Los problemas detectados de coincidencia de tamaño entre fragmentos no homólogos no pueden sino acentuarse en un experimento real, en el que la capacidad de resolución a la hora de visualizar y codificar la presencia o ausencia de fragmentos sobre el gel de electroforesis está muy alejada de la de un nucleótido con que se trabaja en las simulaciones. De esto se desprende, también, la conveniencia de elegir la combinación adecuada entre longitud de secuencia analizada y probabilidad de corte con el enzima o enzimas de restricción empleados.

La extensión de las expresiones aquí desarrolladas para fragmentos de restricción a los estimadores de la divergencia nucleotídica a partir de RAPDs es inmediata. Dado que estos análisis se suelen restringir al ámbito intraespecífico, en el que se cumple la condición de escasa divergencia nucleotídica, y que suele analizarse una longitud considerable de genoma, si se cumplen las demás condiciones indicadas por Lynch y Milligan (1993) y Clark y Lanigan (1993), podemos considerar adecuadas las estimaciones de divergencia nucleotídica obtenidas por este procedimiento. En esas circunstancia es donde pensamos que el estimador de la varianza desarrollado encontrará mayor aplicación.

7. AGRADECIMIENTOS

Estamos en deuda con los Drs. J. Ferrándiz y M. Sendra por sus valiosos comentarios y sugerencias, con el Dr. W.-H. Li por habernos incitado con sus sugerencias a profundizar en nuestro análisis y con dos revisores de este artículo por sus indicaciones que, en todo caso, han contribuido a mejorarlo. S.F.E. ha sido becario predoctoral de la Consellería d'Educació i Ciència de la Generalitat Valenciana. Este trabajo ha sido subvencionado por los proyectos PB93-0690 y PB93-0350 de la DGICYT.

8. BIBLIOGRAFÍA

[1] **Avise, J.C.** (1994). *Molecular Markers, Natural History and Evolution.* Chapman & Hall, New York.

[2] **Clark, A.** y **Lanigan, C.M.S.** (1993). «Prospects for estimating nucleotide divergence with RAPDs». *Mol. Biol. Evol.*, **10**, 1096–1111.

[3] **Fisher, R.A.** (1925). *Statistical methods for research workers.* Edición 13ª. Hafner, New York.

[4] **González Candelas, F., Elena, S.F.** y **Moya, A.** (1995). «Approximate variance of nucleotide divergence between two sequences estimated from restriction fragment data». *Genetics*, **140**, 1443–1446.

[5] **Hadrys, H., Balick, M.** y **Schierwater, B.** (1992). «Applications of random amplified polymorphic DNA (RAPD) in molecular ecology». *Mol. Ecol.*, **1**, 55–63.

[6] **Jukes, T.H.** y **Cantor, C.R.** (1969). «Evolution of protein molecules». En *Evolution of genes and proteins*, 191–207. Sunderland MA, Sinauer Associates.

[7] **Li, W.-H.** (1981). «A simulation study on Nei and Li's model for estimating DNA divergencie from restriction enzyme maps». *J. Mol. Evol.*, **17**, 251–255.

[8] **Lynch, M.** y **Milligan, B.G.** (1994). «Analyzing population genetic structure with RAPD markers». *Mol. Ecol.*, **3**, 91–100.

[9] **Nei, M.** (1987). *Molecular Evolutionary Genetics.* Columbia University Press, New York.

[10] **Nei, M.** y **Li, W.-H.** (1979). «Mathematical model for studying genetic variation in terms of restriction endonucleases». *Proc. Natl. Acad. Sci. USA*, **76**, 5269–5273.

[11] **Nei, M.** y **Miller, J.C.** (1990). «A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data». *Genetics,* **125**, 873–879.

[12] **Nei, M.** y **Tajima, F.** (1981). «DNA polymorphism detectable by restriction endonucleases». *Genetics,* **97**, 145–163.

ENGLISH SUMMARY:

ON THE SIMULATION OF MOLECULAR EVOLUTIONARY PROCESSES: CONSIDERATIONS ON THE DERIVATION AND CONFIRMATION OF AN ESTIMATOR FOR THE VARIANCE OF NUCLEOTIDE DIVERGENCE ESTIMATED FROM RESTRICTION FRAGMENTS

Santiago F. Elena, Andrés Moya and Fernando González-Candelas

1. INTRODUCTION

In the last few years an increasing number of studies that make use of DNA fragments to characterize genetic divergences between individuals, populations, and species are being employed. This is the case, for instance, of RFLPs or RAPDs-PCR and related methods. In order to make statistical inferences about the equality or not of two divergence estimates, it is necessary to know its variance. Nei and Li (1979) developed an statistical method to obtain the divergence between two DNA sequences using this kind of information. However, they did not develop an estimate for the variance of their divergence estimate. In the present work, we have developed such an estimator.

We have started from the expression for the divergence derived by Nei and Li (1979). Applying Fisher's delta method (Fisher, 1925) over their expressions and retaining up to third order moments in the Taylor's expansion we have derived an approximate estimator for the variance of the nucleotide divergence (equation 7).

In order to test the accuracy of our expression, a computer simulation following the procedure described by Li (1981) has been carried out. The simulation analyses were based on three random DNA sequences of different lengths. These sequences were made to evolve, allowing only for nucleotide substitutions, giving rise 2000 pairs of derived sequences for each evolutionary rate. The resulting pairs were compared by three different methods:

(i) Using the Jukes-Cantor correction for estimating nucleotide divergence,

(ii) using Nei's model for divergence estimation with restriction sites data (Nei, 1987) and

(iii) using only restriction fragment length data. The first method gave us an idea of the reliability of the simulated process: we found a good agreement between expected and predicted divergence values and also for their variances.

For restriction site data, better estimates are obtained using one single four cutter than ten six-cutter enzymes. This is independent of the length of the sequence used in the simulation process. The use of fragment lengths of the sequence used in the simulation process. The use of fragments lengths presents several problems. For instance, nucleotide divergences can only be estimated reliably from a combination of the right sequence length and number and type of enzymes, although underestimates have usually been found.

The need for exhaustive controls and exploration of all parameters, independently of their incorporation into the final expressions, during simulations aimed at checking analytical derivations is discussed.