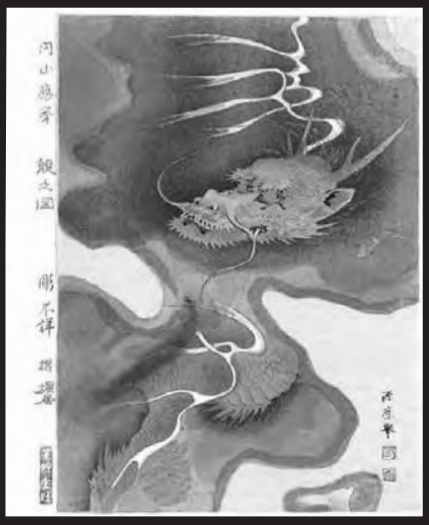
Alineación e interferencia filogenética

Manuel Feria Ortiz



Desde hace varias décadas se ha observado un interés creciente por realizar estudios filogenéticos moleculares. En un principio se utilizaron con mayor frecuencia datos de secuencias de proteínas o de sitios de restricción, pero desde la implementación de los métodos de secuenciación de ADN a finales de los setentas y en particular desde que se comenzó a utilizar la reacción en cadena de la polimerasa ha habido un aumento en el uso de datos de secuencias de ácidos nucleicos (nucleares o de organelos celulares). Actualmente, los datos de secuencias representan una herramienta rutinaria dentro de muchos campos de la biología y su uso en la inferencia filogenética ha tenido un impacto considerable en la sistemática de una gran diversidad de organismos.

La preponderancia en el uso de secuencias de ADN (o ARN) en estudios filogenéticos se debe a varios factores; uno de ellos consiste en el gran poder resolutivo de los datos de secuencias. Los genomas nucleares y extranucleares (de mitocondrias y cloroplastos) ofrecen una enorme gama de caracteres. Además, diferentes segmentos del genoma pueden poseer propiedades diferentes y ser útiles en la resolución de problemas distintos; por ejemplo, mientras que algunas secuencias permiten investigar relaciones filogenéticas a nivel poblacional o de especies cercanamente emparentadas (como ADN de mitocondrias), otras permiten valorar las relaciones de grupos tan distantes como filos o reinos (genes nucleares que codifican para ribosomas). Otro factor importante para el predominio en el uso de secuencias de ADN consiste en la gran variedad de métodos que se han propuesto para inferir y evaluar árboles filogenéticos a partir de datos de estas secuencias. También es necesario señalar

el desarrollo en la tecnología de cómputo, pues paralelamente a la obtención y acumulación de datos de secuencias ha mejorado enormemente la capacidad de procesamiento de los equipos de cómputo y se ha implementado una variedad de programas que permiten obtener árboles a partir de los métodos de inferencia propuestos.

Elección de secuencias

La aplicación de algún método de inferencia, por efectivo que sea, no garantiza por sí mismo la obtención de resultados confiables. La eficiencia de cualquiera de los métodos disponibles depende no sólo de sus propias cualidades sino también de la calidad, elección y obtención de los datos. Si los datos no son adecuados para el problema que se estudia, entonces los resultados serán espurios. En el caso de datos de secuencias la elección consiste en cuál segmento de ADN o ARN ha de utilizarse, el de un gen o el de una secuencia particular, por lo que es uno de los aspectos más importantes en cualquier estudio filogenético molecular.

Uno de los aspectos más importantes a considerar cuando se elige un gen es su tasa de evolución, que ésta sea adecuada o que no dependa del grupo particular bajo análisis. Si la tasa de evolución de un gen es relativamente alta y los tiempos de divergencia de los taxones involucrados son muy extensos (como en el caso de taxones lejanamente emparentados) se espera que la variación dentro de un sitio nucleotídico particular no refleje adecuadamente el cambio evolutivo ocurrido en el mismo. Esto es así debido a que si la tasa de sustitución es alta y el tiempo transcurrido es relativamente largo hay oportunidad para que en un mismo sitio (de un mismo gen) hayan ocurrido varias sustituciones. Este hecho puede conducir a inferencias erróneas; por ejemplo, si dos secuencias tuvieran el nucleótido citosina en una determinada posición podría asumirse erróneamente que no ha habido cambio evolutivo en dicho sitio cuando realmente han ocurrido varias sustituciones, pero éstas han ocurrido de tal manera que ahora las dos secuencias poseen el mismo nucleótido, (esto es, que han convergido hacia el mismo nucleótido).

Alineación

Otro paso crítico en un análisis filogenético consiste en manejar los datos disponibles de tal modo que se obtenga una matriz de datos confiable y que pueda ser leída por los paquetes de cómputo disponibles. El que sea confiable implica que los caracteres involucrados sean legítimamente comparables, lo que en sistemática implica que sean potencialmente homólogos. En el caso de caracteres morfológicos el manejo de los datos incluye principalmente la delimitación, codificación y ordenación de los estados de carácter, así como la ponderación de caracteres y estados de caracteres. En el caso de datos de secuencias, los caracteres que se van a utilizar consisten en los sitios nucleotídicos presentes a lo largo de la misma (o los sitios correspondientes a los aminoácidos en el caso de proteínas), y los nucleótidos o aminoácidos presentes en cada sitio particular. En consecuencia, en el caso de datos de secuencias, el manejo preliminar de los datos consiste en obtener una matriz en la cual las filas sean las secuencias correspondientes a los taxones bajo análisis y las columnas los caracteres (sitios nucleotídicos) involucrados en el análisis. Esto implica acomodar las secuencias de tal modo que todos los nucleótidos que queden en una columna particular sean homólogos entre sí. A este proceso se le denomina alineación e involucra varias hipótesis de homología, una por cada posición de nucleótido.

La alineación de una serie de secuencias no es un problema trivial. Si bien en algunos casos puede ser sencillo en otros llega a convertirse en un problema analítico del mismo orden de complejidad que la inferencia filogenética. Por esta razón, como ha ocurrido en el caso de los métodos de inferencia filogenética, se ha propuesto una variedad de métodos de alineación de secuencias que poseen diferentes ventajas y desventajas; asimismo, debido a que en muchos casos el proceso de alineamiento se realiza an-

tes de la aplicación de un método de inferencia, las relaciones filogenéticas que finalmente se obtengan dependerá en gran medida del método de alineación que se utilice. De hecho, muchos autores consideran que la filogenia resultante depende más del método de alineamiento que del método de inferencia que se utilice.

La necesidad de alinear las secuencias de nucleótidos se debe a la ocurrencia de mutaciones puntuales (sustituciones) y de eventos de inserción o deleción llamados "indeles" (contracción de estas palabras). Si no hubiera indeles, la longitud de un gen o segmento de nucleótido particular (o de aminoácidos en el caso de proteínas) sería la misma en todos los organismos que lo tuvieran. En consecuencia, todas las secuencias tendrían las mismas posiciones y todos los nucleótidos presentes en una posición o columna particular serían homólogos entre sí; en tal caso, evidentemente cada posición representaría un carácter para el análisis filogenético y no habría motivo para intentar realizar una alineación. Por otro lado, si sólo ocurrieran eventos de inserción o deleción el alineamiento de las secuencias sería un proceso muy sencillo, únicamente tendrían que acomodarse las secuencias de tal modo que en cada posición se encontraran siempre los mismos nucleótidos, una situación en la que la única fuente de información filogenética serían los indeles (huecos en las alineaciones) y la reconstrucción filogenética, si fuera posible, requeriría secuencias extremadamente largas con el fin de que se incluyera una cantidad suficiente de indeles filogenéticamente informativos.

La ocurrencia de diferentes tipos de mutaciones provoca que se obtengan secuencias de distinta longitud, que los nucleótidos homólogos no sean los mismos en cada secuencia considerada y que éstos se encuentren desfasados con respecto de la posición homóloga a la cual pertenecen. En consecuencia, la alineación consiste básicamente en detectar en qué posiciones debieron haber ocurrido eventos de inserción o deleción y ubicar espacios en dichos sitios a fin de que cada posición incluya únicamente nucleótidos homólogos. Realmente no es posible asegurar la homología de los nucleótidos presentes en un sitio determinado; esto implicaría asegurar que los nucleótidos presentes en el sitio derivan de un mismo nucleótido presente en una secuencia antecesora, lo cual es imposible de saber. La alineación, por lo tanto, consiste en el establecimiento de hipótesis de

homología de estados de carácter (una para cada posición de la secuencia), y es un paso fundamental en el análisis filogenético de las secuencias de nucleótidos. Como se señaló antes, los errores en esta fase del análisis se verán reflejados en todos los análisis subsecuentes (obtención de árboles, valoración de los mismos, etcétera), al margen del rigor con el que se realicen.

En la práctica, el alineamiento de secuencias correspondientes a taxones cercanamente emparentados o de secuencias de genes que codifican proteínas es normalmente sencillo; estas secuencias son comúnmente muy similares entre sí en virtud del poco tiempo transcurrido desde su divergencia. La similitud en las secuencias facilita su alineación. En el caso de secuencias de genes que codifican proteínas, las inserciones o deleciones son raras y generalmente afectan a tres —o a algún múltiplo de tres— sitios nucleotídicos (de otro modo se alteraría drásticamente el marco de lectura del gen bajo alineación). Por lo tanto, a menos que la divergencia entre las secuencias sea muy alta, la alineación de genes que codifican proteínas generalmente no presenta problemas.

Contrariamente, los segmentos de ADN no codificadores, tales como los intrones o aquellos segmentos que se encuentran entre genes, son particularmente problemáticos de alinear. En tales casos, sobre todo en secuencias muy divergentes, es típicamente común la ocurrencia de los tres tipos de mutaciones: inserciones, deleciones y sustituciones. En consecuencia, las secuencias involucradas comúnmente poseen longitudes muy diferentes y existen muy pocos sitios conservados, esto es, muy pocos sitios en donde todas las secuencias poseen el mismo nucleótido; esto impide enormemente el alineamiento debido en particular a la dificultad de determinar los sitios exactos en los que ocurrieron los indeles.

Métodos de alineación de secuencias

Algunos autores han alineado las secuencias "a mano". Es frecuente que el investigador se base en el conocimiento de la estructura de la molécula codificada por la secuencia en cuestión para alinear sus secuencias. En el caso de secuencias que codifican proteínas, la alineación debe mantener

el marco de lectura que produce la proteína en cuestión; por lo tanto, una manera de verificar la alineación consiste en convertir las secuencias de nucleótidos en secuencias de aminoácidos y corregir las inconsistencias, es decir, los cambios en el marco de lectura por la introducción de huecos únicos. Sin embargo, la alineación manual sólo es posible en el caso de que haya relativamente pocos indeles y las secuencias no sean muy divergentes. Comúnmente, aun en el caso de secuencias de genes que codifican para proteínas, la ocurrencia de sustituciones y algunos indeles complica el proceso, de modo que se vuelve impráctica la alineación manual. Por esta razón, en prácticamente todos los casos la alternativa más viable es la utilización de algún algoritmo matemático. Aun así, con el fin de obtener una matriz de datos más confiable, es frecuente que un investigador utilice algún algoritmo y posteriormente afine las secuencias obtenidas "a mano". Con frecuencia tales ajustes también se realizan con base en el conocimiento de la estructura secundaria de la molécula producida por las secuencias en cuestión.

Se han propuesto diversos algoritmos para alinear secuencias de nucleótidos y muchos de éstos están diseñados para comparar y alinear parejas de secuencias. Sin embargo, si el propósito es realizar estimaciones filogenéticas, necesariamente se tienen que alinear más de dos secuencias. La mayoría de los algoritmos diseñados para alinear secuencias múltiples se basa en el algoritmo de alineación descrito en 1970 por Needleman y Wunsch, en el cual se asignan valores positivos a las coincidencias positivas (el mismo nucleótido en ambas secuencias), y cero o valores negativos a las coincidencias negativas (los nucleótidos involucrados son diferentes). Debido a la facilidad con la que pueden insertarse huecos de tal modo que se obtengan solamente coincidencias positivas, se hace necesario penalizar la introducción de los mismos en el proceso de alineamiento. En consecuencia, a los huecos se les asigna comúnmente un valor negativo mayor que el asignado a las coincidencias positivas; no obstante, los huecos de más de una posición no se penalizan normalmente en proporción directa a su tamaño. La razón es que es más probable que dos o más nucleótidos se inserten o eliminen simultáneamente a que ocurran dos o más indeles independientes en sitios contiguos.

El procedimiento usual en muchos algoritmos de alineación múltiple consiste en alinear progresivamente las secuencias: primero se alinea un par de secuencias y después se van agregando y alineando una a una las demás hasta obtener todas las secuencias alineadas. Sin embargo, dado que la alineación final depende del orden en el que se agreguen las secuencias, una estrategia consiste en realizar alineaciones para pares de secuencias y, con base en sus similitudes, obtener un árbol que sirva de guía para decidir el orden con el que se alinearán las secuencias.

Otra estrategia consiste en realizar simultáneamente la alineación y el análisis filogenético, y una más se basa en la idea de que el alineamiento y la inferencia filogenética tienen una meta común y por lo tanto no deben de ser procedimientos independientes. La implicación es que el alineamiento y la inferencia filogenética deben ser procedimientos ligados y el primero no debe preceder al segundo.

Actualmente existen muchos programas de cómputo que permiten realizar alineamientos de datos de secuencias (de nucleótidos o aminoácidos), los cuales ejecutan algún algoritmo como los señalados previamente y una matriz de secuencias alineada. Ciertos programas, como ClustalX o W, pueden bajarse gratuitamente de la red. Otros, sin embargo, tales como MEGA, tiene un costo y deben pagarse antes de poder ser utilizados plenamente.

Alineaciones problemáticas

En muchos casos es difícil estar seguro de que se ha obtenido una alineación confiable, particularmente cuando la divergencia entre las secuencias es grande y su alineación requiere la incorporación de muchos huecos. La razón es que cuando ha habido mucha divergencia entre las secuencias, diferentes combinaciones de parámetros (costos de sustitución y penaltis de huecos y extensión de huecos) pueden dar alineamientos distintos y es difícil precisar cuál de todos ellos es el más confiable. Ante esta situación es necesario verificar la alineación y en su caso modificarla con el fin de obtener una alineación razonable. Como ya se señaló, dicha verificación puede realizarse "a ojo" o con base en la estructura secundaria de la molécula codificada. No obstante, deben justificarse claramente todas las modificaciones que se realicen.

Por otro lado, una situación común en muchos estudios filogenéticos es que las secuencias contengan regiones en las cuales no sea posible identificar adecuadamente las columnas de nucleótidos homólogos. La mayor parte de los investigadores simplemente borran estas regiones debido a que sus caracteres o posiciones pueden ser engañosos o contener muy poca información filogenética. No obstante, aun las regiones más problemáticas pueden contener información filogenéticamente útil y resulta difícil delimitar objetivamente qué secciones son ambiguas y cuáles confiables.

Se han propuesto diferentes procedimientos para tomar en cuenta las regiones ambiguas. Uno consiste simplemente en producir todos los alineamientos posibles para un intervalo de parámetros de alineamiento particular y, posteriormente, se obtiene un árbol para cada alineamiento y se aceptan únicamente los clados que estén presentes en todos los árboles. Otro método es el de elisión, que consiste en obtener dos o más alineamientos y concatenarlos en una matriz única más grande. De este modo, automáticamente se le da más peso a las posiciones no ambiguas (que resultan similares en todas las alineaciones) y menos a las más ambiguas que podrían aparecer, por ejemplo, en sólo una de las alineaciones.

En otro método, denominado *fixed caracter state* o *fragment-level alignment*, las regiones ambiguamente alineadas se tratan como caracteres multiestado; los estados de cada carácter consisten en las variantes de secuencia distintas encontradas. Posteriormente se construye una matriz de pasos para asignarle un costo a la trasformación de un estado en otro (esto es, el cambio de una variante de secuencia en otra) y dicho costo se calcula con base en las sustituciones y huecos que se requieren para transformar un estado en otro. De este modo, las transformaciones entre estados muy divergentes tienen costos más altos.

Si bien todos estos métodos presentan ventajas y desventajas representan una alternativa viable a la simple eliminación de las regiones problemáticas.

Manuel Feria Ortiz
Museo de Zoología, Facultad de Estudios Superiores-Zaragoza,
Universidad Nacional Autónoma de México.

Referencias bibliográficas

Doyle, Jeff J. y Jerrold I. Davis. 1998. "Homology in molecular phylogenetics: a parsimony perspective", en *Molecular Systematics of Plantas II. DNA sequencing*, Soltis, Douglas, Pamela Sotis y Jeff J. Doyle (eds.). Kluwer Academic Publishers, Londres. Pp. 101-131.

Felsenstein, Joseph. 2004. *Inferring phylogenies*. Macmillan Education, Massachusetts.

Lemey, Philippe, Marco Salemi y Anne-Mieke Vandamme. 2009. *The phylogenetic handbook. A practical approach to phylogenetic. Analysis and hypothesis testing*. Cambridge University Press, Cambridge.

Lutzoni, François *et al.* 2000. "Integrating ambiguously aligned regions of dna sequences in phylogenetic analysis without violating positional homology", en *Syst. Biol.*, vol. 49, núm. 4, pp. 628–651.

Morrison, D. A. y J. T. Ellis. 1997. "Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S *rDNAS* of Apicomplexa", en *Molecular Biology Evolution*, vol. 14, núm. 4, pp. 428–441.

Nei, Masatoshi y Sudhir Kumar. 2000. *Molecular evolution and phylogenetics*. Oxford University Press, Nueva York.

Rosenberg, S. M. 2009. "Sequence alignment. Concepts and history", en *Sequence alignment. Methods, models, concepts, and strategies*, Rosenberg, Michael S. (ed.). University of California Press, Londres. Pp. 1–22.

Scotland, Robert y R. Toby Pennington (eds.). 2000. *Homology and systematics*. Taylor and Francis, Londres.

Swofford, David L. *et al.* 1996. "Phylogenetic inference", en *Molecular Systematic*, Hillis, David M., Craig Moritz y Barbara K. Mable (eds.). Sinauer Asociates, Sunderland. Pp. 407–514

Wheeler, Ward C. 1995. "Sequence alignment, parameter sensivity, and phylogenetic analysis of molecular data", en *Systematic Biology*, vol. 44, núm. 3, pp. 321–331.

Imágenes

P. 142: Ogata Gekkō, *Ryū Shō Ten*, 1897. P. 143: Maruyama Ōky, *Dragón*, s. XVIII. P. 144: Horiyoshi III, *Toryumon gate*, *ca.* 2000. Utagawa Hiroshige: p. 145: *Gato cruzando un camino para comer*, 1830-1834; p. 146: Utagawa Hiroshige, *Dragon entre nubes*, *ca.* 1835. P. 145: Utagawa Kunisada, *Captura de un pez gato con una calabaza*, 1857. Katsushika Hokusai: p. 145: *Ascenso de dragón y Fuji*, s. XIX; p. 148: *Dragon*, s. XIX. P. 146: *Concurso de poesía de los doce animales del zodiaco*. P. 147: Itō Jakuchū, *Dragón de lluvia*, 1760. P. 148: Yamada Hôgyoku, *Tanuki y el conejo*, 1820-1840; Utagawa Kuniyoshi, *Peces dorados caminando y cantando*, s. XIX.

Alignment and phylogenetic inference

Palabras clave. Inferencia filogenética, secuencias de nucleótidos, alineación, homología.
Key words. Phylogenetic inference, nucleotide sequences, alignment, homology.

Resumen. La sistemática molecular usa la información contenida en moléculas, básicamente proteínas y ADN, para inferir relaciones de parentesco entre organismos. Un procedimiento común consiste en obtener las secuencias de nucleótidos de un segmento particular de ADN a partir de una serie de organismos para posteriormente utilizar estas secuencias para inferir su historia de descendencia. Se han propuesto muchos métodos para inferir relaciones de parentesco con base en datos de secuencias. Sin embargo, antes de poder utilizar alguno de ellos es necesario asegurarse de que las secuencias en mano, y en particular de cada uno de los nucleótidos que las integran, sean realmente comparables entre sí. Si no son homólogos los métodos de inferencia, por eficientes que sean, producirán resultados erróneos. De hecho, muchos investigadores han señalado que la reconstrucción filogenética depende más del método de alineamiento que del método de inferencia en sí.

Abstract. Molecular systematics uses the information contained in molecules, basically proteins and DNA, to infer relationships of kinship among organisms. A common procedure is to obtain nucleotide sequences of a particular DNA segment from a series of organisms to then use those sequences to infer their history of descendancy. Many methods have been proposed to infer relationships of kinship based on data from sequences. However, before any of them can be used, we need to ensure that the sequences in hand, and in particular those of each of their constituent nucleotides, are truly comparable. If methods of inference are not uniform, however efficient they may be, they will produce erroneous results. In fact, many researchers have remarked that phylogenetic reconstruction depends more on the method of alignment than on the method of inference per se.

Manuel Feria Ortiz se recibió como biólogo en la Facultad de Estudios Superiores Zaragoza, UNAM en 1986. Realizó sus estudios de maestría y doctorado en la Facultad de Ciencias, también en la UNAM. Actualmente es Profesor de Carrera Titular "A" de tiempo completo y desde 1990 es responsable de la colección de anfibios y reptiles de la FES-Zaragoza. A partir de su ingreso a la Facultad ha impartido las materias de evolución y taxonomía que forman parte del plan de estudios de la carrera de Biología. Como investigador se ha dedicado principalmente al estudio de anfibios y reptiles.

Recibido el 22 de septiembre de 2011; aceptado el 4 de diciembre de 2013.