Rev. peru. biol. 18(2): 265 - 267 (Agosto 2011)

ISSN 1561-0837

COMENTARIO

Métodos de inferencia filogenética

Carlos Peña

Editor asociado, Laboratory of Genetics, Department of Biology, University of Turku, 20014 Turku, Finland. Email: carlos.pena@utu.fi

Presentado: 07/07/2011
Aceptado: 19/07/2011
Publicado online: 25/08/2011

Existen tres métodos comúnmente utilizados en estudios de sistemática filogenética: i) la Cladística, usando el principio de Máxima Parsimonia (MP); ii) Maximum Likelihood (ML) y la iii) Inferencia Bayesiana (IB). En recientes años se viene utilizando con frecuencia el método Neighbor-Joining (NJ), el cual no es un método filogenético; en las siguientes líneas explicare brevemente, el porqué no debe ser utilizado para inferir relaciones filogenéticas.

Fundamentos filogenéticos

De los tres métodos válidos para inferencia filogenética, MP, ML, e IB, algunos sistemáticos utilizan solo uno de ellos para sus manuscritos, o combinaciones de los tres. Por lo general, estas preferencias se basan en fundamentos filosóficos y/o razones prácticas.

La finalidad de los análisis filogenéticos es estimar una filogenia (árbol filogenético) que muestre la historia evolutiva del grupo taxonómico de estudio. Es decir, el objetivo final es un árbol filogenético que sea reflejo del proceso de evolución donde las entidades biológicas son el resultado de "descendencia con modificación" (Darwin 1859) entre especies ancestrales y descendientes. Una manera de cumplir este objetivo es mediante la búsqueda de indicios de descendencia con modificación en las especies, ya sea caracteres morfológicos o moleculares. Por ejemplo, se puede utilizar el carácter morfológico "notocorda" que apareció en la especie ancestral de todos los cordados y ha sido heredada y modificada múltiples veces (estados de carácter) a lo largo de la historia evolutiva del filo Chordata. La aparición de esqueleto rodeando la notocorda (o vestigios de esta) se utiliza para agrupar al grupo "vertebrados". En el contexto filogenético, el grupo que incluye la especie ancestral, donde apareció este estado de carácter, y todas sus especies descendientes se denomina clado o grupo monofilético (en este caso, clado Vertebrata).

Es por esto que en la sistemática filogenética es vital encontrar caracteres útiles en el rastreo de casos de descendencia con modificación para indentificar grupos monofiléticos. Obviamente los caracteres que serán de mayor utilidad son los caracteres homólogos, mientras que los caracteres homoplásicos podrán ser causantes de errores en los árboles filogenéticos al inferir relaciones de parentesco erróneas. Además, parte inherente de la práctica filogenética es la relación ancestro–descendientes que se trata de estimar a partir de los caracteres estudiados, y cuya relación se trata de representar en los árboles filogenéticos.

Neighbor-Joining (NJ)

El método NJ es frecuentemente usado en estudios de "código de barras molecular" (*DNA barcoding*) (Hebert et al. 2004, 2010), el cual consiste en utilizar un segmento de 648 bases nitrogenadas del gen mitocondrial citocromo *c* oxidasa I (COI) como identificador único para cada especie del reino animal (Hebert et al. 2003). El objetivo fundamental de este método es poder identificar individuos cuando se desconoce la especie a la que pertenecen. Además, existe la posibilidad de descubrir especies crípticas en grupos de morfología compleja o inaccesible (Hebert et al. 2003, Silva-Brandão et al. 2009). Esto es factible ya que la variabilidad intra-específica de esta secuencia es menor que la variación existente entre especies diferentes. Luego de secuenciar el fragmento del COI para grupos de individuos, especies de un mismo género, o incluso especies pertenecientes a diferentes filos (Hebert et al., 2003), dependiendo del nivel taxonómico que se quiera estudiar, es necesario analizar las secuencias. El algoritmo de preferencia para estudios de código de barras molecular es el Neighbor-Joining (Hebert et al. 2003, 2004, 2010).

El algoritmo NJ fue creado por Saitou & Nei (1987) y consiste en generar un único "árbol filogenético" final, el cual, según los autores, no necesariamente será el "árbol verdadero". En el paso inicial, se unen los dos *neighbors* (secuencias) que tengan la menor distancia genética. Luego, este par inicial se considera como una sola entidad, y se busca el siguiente terminal que tenga la menor distancia genética con este. El procedimiento continúa hasta unir todos los terminales al "árbol filogenético" (Saitou & Nei 1987). En el caso de secuencias de ADN, la distancia genética entre dos terminales se calcula en base al número total de sustituciones de bases nitrogenadas (Saitou & Nei 1987), es decir en el número de bases nitrogenadas que son diferentes en estas dos secuencias. Para seleccionar los terminales que tengan menor distancia genética, lo que en la práctica es escoger las secuencias más parecidas, es necesario construir una matriz estándar de distancias entre todas las posibles combinaciones de terminales. Entonces se construye el "árbol filogenético" agregando terminales tomando como información los valores de la matriz de distancias (Saitou & Nei 1987).

Si bien Farris et al. (1996) señala algunos problemas en el método NJ, es importante notar que el "árbol filogenético" final solamente refleja el grado de similaridad de los terminales. Por esto, NJ es útil en estudios de código de barras molecular ya que los individuos pertenecientes a una misma especie formarán grupos aislados debido a su alto nivel de semejanza (o semejanza de sus secuencias COI). Sin embargo, el método NJ no toma en cuenta la relación ancestro–descendientes. Tampoco considera el principio de descendencia con modificación (Darwin 1859). Por lo tanto, el árbol que se obtiene usando NJ no refleja relaciones evolutivas, en realidad es una representación del grado de similaridad de los terminales, lo cual corresponde al campo de la

fenética (Crisci & López, 1983), no de la sistemática filogenética. Por estos motivos, el árbol inferido con NJ no debe ser considerado como árbol filogenético, y sería incorrecto suponer que los agrupamientos resultantes consisten en una especie ancestral y sus especies descendientes. En realidad son agrupaciones de miembros muy similares que no pueden ser considerados clados ni grupos monofiléticos.

Métodos filogenéticos: Máxima Parsimonia, *Maximum Likelihood* e Inferencia Bayesiana

En MP se usa un mínimo de asunciones *a priori* sobre los caracteres utilizados como fuente información —se asume que cualquier carácter heredable es una homología potencial (Grandcolas et al. 2001). Entonces, todos los caracteres son tratados de igual manera, con el "mismo peso" o misma influencia, al momento de inferir los árboles filogenéticos debido a que no se puede (o no se quiere) identificar homoplasias *a priori* (Hennig 1968). En MP, el árbol filogenético que se prefiere es el que implica la mínima cantidad de cambios evolutivos ("pasos" evolutivos) que se requieren para explicar una determinada matriz de caracteres (Farris 1970, Swofford et al. 1996).

Maximum Likelihood y la Inferencia Bayesiana son métodos estadísticos basados en modelos de evolución molecular, donde se toma en cuenta conocimiento *a priori* acerca de los caracteres, especialmente cuando son caracteres moleculares (frecuentemente secuencias de nucleótidos de ADN). El método ML estima la probabilidad de qué tan bien la matriz de caracteres es explicada por los árboles filogenéticos (Felsenstein 2004), mientras que IB estima la probabilidad de qué tan bien los árboles filogenéticos son explicados por los datos (la matriz de caracteres) (Huelsenbeck et al. 2001, Brooks et al. 2007). Maximum Likelihood necesita calcular cada árbol posible que pueda ser derivado de los datos según el modelo de evolución seleccionado. Además, debe calcular la longitud de ramas para cada árbol diferente (Huelsenbeck & Rannala 1997). Algunos prefieren usar IB sobre ML debido a que el primer método utiliza "atajos" para los cálculos al emplear el algoritmo conocido como *Markov Chain Monte Carlo* (MCMC), el cual permite realizar búsquedas a través de un número menor de árboles según sus valores de probabilidades posteriores (Huelsenbeck et al. 2001). Esto permite que la IB demande menos poder computacional y sea más rápida que ML.

Si bien estos tres métodos son ampliamente utilizados, no están libres de críticas. Máxima Parsimonia es afectado por el fenómeno conocido como atracción de ramas largas, *long branch attraction* (Felsenstein 1978), lo que causa que los árboles reflejen relaciones filogenéticas espurias cuando la cantidad de caracteres homoplásicos abruma los caracteres homólogos (Bergsten 2005). Maximum Likelihood es afectado por la "repulsión" de grupos hermanos cuando estos se ubican en ramas largas de los árboles (Siddall 1998). Más aún, ML y la IB son poco confiables cuando las tasas de evolución de ADN no son homogéneas en el tiempo ni entre linajes (Kolaczkowski & Thornton 2004).

Algunos sistemáticos, tal vez simpatizantes de cada uno de estos métodos, han sido muy explícitos al defender su método preferido, señalando los defectos de los otros (Swofford et al. 1996, Siddall 1998, Farris 1999, Ebach et al. 2008). Aunque es bien conocido que estos métodos no son efectivos en todas las circunstancias, estos desacuerdos reflejan la carencia de consenso en la comunidad científica en lo que respecta a la metodología filogenética. Es en esta situación algunos investigadores estarán indecisos en escoger un método para analizar sus datos. Algunos prefieren utilizar un procedimiento conciliatorio al emplear los tres métodos. Si las topologías de los árboles obtenidos, usando los diferentes métodos, son concordantes, la hipótesis filogenética resultante es considerada robusta (ej. Martin et al. 2002). Si este no es el caso, se deberá tener cautela al discutir las relaciones filogenéticas de los nodos discordantes (Pol 2001, Kolaczkowski & Thornton 2004).

Afortunadamente, existen estrategias para evitar algunos de los artefactos que sufren los métodos —por ejemplo la extracción de ramas largas en MP (Siddall & Whiting 1999); el uso de complejos modelos mixtos de evolución molecular en ML (Kolaczkowski & Thornton 2004). Aunque se culpa a la imperfección de los métodos como la causa de las discordancias, es menos frecuente que la culpa se atribuya a la complejidad de las matrices de caracteres, probablemente porque la mayoría de estos estudios se basan en datos simulados (ejemplo: Felsenstein 1978, Siddall 1998, Kolaczkowski & Thornton 2004) de entidades biológicamente inverosímiles (ver Steel 2005). A pesar que se ha recalcado que aumentar el número de taxones y caracteres en las matrices pueda ayudar a resolver relaciones filogenéticas ambiguas y nodos débilmente soportados (Wiens 1998, Zwickl & Hillis 2002), esto ha comenzado a ser explorado recientemente mediante análisis de datos empíricos (Hallström & Janke 2008, Wahlberg & Wheat 2008, Peña et al. 2011).

Agradecimientos

A Niklas Wahlberg por sus comentarios acerca de este tema.

Literatura citada

Bergsten J. 2005. A review of long-branch attraction. Cladistics, 21, 163–193.

Brooks D.R., J. Bilewitch, C. Condy, et al. 2007. Quantitative Phylogenetic Analysis in the 21st Century. Revista Mexicana de Biodiversidad, 78, 225–252.

Crisci J.V. & M. López. 1983. Introducción a la teorìa y práctica de la taxonomìa numérica. Secretarìa General de la Organización de los Estados Americanos. 132 pp.

Darwin C. 1859. On the Origin of Species by Means of Natural selection, or, the Preservation of Favoured Races in the Struggle for Life. London.

Ebach M.C., D.M. Williams & A.C. Gill. 2008. O Cladistics, Where Art Thou? Cladistics, 24, 851–852.

Farris J.S. 1970. Methods for computing Wagner trees. Systematic Zoology, 19, 83–92.

Farris J.S. 1999. Likelihood and inconsistency. Cladistics, 15, 199–204.

Farris J.S., V.A. Albert, M. Källersjö, et al. 1996. Parsimony jackknifing outperforms Neighbor-Joining. Cladistics, 12, 99–124.

Felsenstein J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Systematic Zoology, 27, 401–410.

Felsenstein J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Grandcolas P., P. Deleporte, L. Desutter-Grandcolas & C. Daugeron. 2001. Phylogenetics and Ecology: as many characters as possible should be included in the cladistic analysis. Cladistics, 17, 104–110.

Hallström B.M. & A. Janke 2008. Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations. BMC

Evolutionary Biology, 8, 162.
Hebert P.D.N., A. Cywinska, S.L Ball. & J.R. DeWaard. 2003. Biological identifications through DNA barcodes. Proceedings. Biological sciences / The Royal Society, 270, 313–21.
Hebert P.D.N., J.R. Dewaard & J.F. Landry. 2010. DNA barcodes for 1/1000 of the animal kingdom. Biology letters, 6, 359–62.
Hebert P.D.N., E.H. Penton, J.M. Burns, et al. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. PNAS, 101, 14812–14817.
Hennig W. 1968. Elementos de una sistemática filogenética. Eudeba, Buenos Aires.
Huelsenbeck J.P. & B. Rannala. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science, 276, 227–232.
Huelsenbeck J.P., F. Ronquist, R. Nielsen & J.P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science, 294, 2310–4.
Kolaczkowski B. & J.W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature, 431, 980–984.
Martin J.F., A. Gilles, M. Lörtscher & H. Descimon. 2002. Phylogenetics and differentiation among western taxa of the Erebia tyndarus group (Lepidoptera: Nymphalidae). Biological Journal of the Linnean Society, 75, 319–332.
Peña C., S. Nylin & N. Wahlberg 2011. The radiation of Satyrini butterflies (Nymphalidae: Satyrinae): a challenge for phylogenetic methods. Zoological Journal of the Linnean Society, 161, 64–87.
Pol D. 2001. Biases in Maximum Likelihood and Parsimony: A Simulation Approach to a 10-Taxon Case. Cladistics, 17, 266–281.
Saitou N. & M. Nei. 1987. The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4, 406–425.
Siddall M.E. 1998. Success of Parsimony in the Four-Taxon Case: Long-Branch Repulsion by Likelihood in the Farris Zone. Cladistics, 14, 209–220.
Siddall M.E. & M.F. Whiting. 1999. Long-branch abstractions. Cladistics, 15, 9–24.
Silva-Brandão K.L., M.L. Lyra & A.V.L. Freitas. 2009. Barcoding Lepidoptera: current situation and perspectives on the usefulness of a contentious technique. Neotropical entomology, 38, 441–51.
Steel M. 2005. Should phylogenetic models be trying to 'fit an elephant'? Trends in Genetics, 21, 307–309.
Swofford D.L., G.J. Olsen, P.J. Waddell & D.M. Hillis. 1996. Phylogenetic inference, Sinauer Associates, Sunderland, MA, pp. 407–514.
Wahlberg N. & C.W. Wheat. 2008. Genomic outposts serve the phylogenomic pioneers: designing novel nuclear markers for genomic DNA extractions of Lepidoptera. Systematic Biology, 57, 231–242.
Wiens J.J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Systematic Biology, 47, 625–640.
Zwickl D.J. & D.M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology, 51, 588–598.