TEMA 2.2. Puntos clave en el diseño y elaboración de una filogenia molecular

Contacto: Virginia Valcárcel (virginia.valcarcel@uam.es)

Las filogenias moleculares son estimas, por lo que para un mismo conjunto de datos puede haber múltiples soluciones igualmente correctas. El punto clave en el diseño de un estudio de filogenia molecular es maximizar las probabilidades de encontrar el árbol 'correcto' según el método seleccionado. Estas probabilidades están afectadas por características del propio conjunto de datos y del árbol; el balance del árbol (la longitud relativa de las ramas internas y los terminales), así como por la propia topología del árbol (pectinada o dicotómica; Smith 1994).

Si bien el gran avance en sistemática de los últimos 30 años se fundamenta sobre una robusta base metodológica, aún existen numerosas fuentes de error. Algunas de estas fuentes de error están asociadas a cuestiones prácticas y tecnológicas como falta de marcadores moleculares adecuados (Hughes et al. 2006, Shaw et al. 2007) o falta de potencia de cálculo, entre otros (véase a continuación). Además de estas cuestiones "prácticas", el desconocimiento de los procesos evolutivos subyacentes puede incurrir en errores de interpretación o en la recuperación de reconstrucciones filogenéticas "falsas". De cara a la interpretación de una reconstrucción filogenética habría que considerar: (1) la distinción entre árboles de genes y árboles de especies (Doyle 1992, Maddison 1997), que puede conducir a errores de interpretación por ejemplo debido al análisis de genes parálogos (Pamilo & Nei 1988); (2) la incidencia de procesos como lineage sorting, ya que el mantenimiento de polimorfismos ancestrales y la pérdida azarosa de linajes implica estimas filogenéticas incorrectas; y (3) el grado y tasa de homogenización inter e intragenómica después de hibridación debida a la evolución concertada en genes de copia múltiple, entre otros.

Procedencia, número y selección de marcadores moleculares y regiones de ADN. Las células animales y vegetales presentan dos (nuclear y mitocondrial) y tres tipos de genomas (nuclear, plastidial y mitocondrial), respectivamente. Estos tipos de genomas presentan diferentes modos de herencia y distintas tasas de cambio. Un estudio en sistemática de plantas debería prospectar regiones moleculares tanto del genoma nuclear como del plastidial y/o mitocondrial (Judd et al. 2002). Del mismo modo, un estudio en sistemática de animales debería prospectar regiones tanto del genoma nuclear como del mitrocondrial. Esta recomendación se convierte en requisito fundamental cuando existen indicios de hibridación.

Un mayor número de regiones analizadas aumenta la resolución y fiabilidad de los datos (Hillis et al. 1996, aunque véase Philippe et al. 2005) y el apoyo, siempre que las regiones prospectadas sean congruentes y pueda realizarse un análisis combinado (Hughes et al. 2006).

La selección de las regiones analizadas viene determinada por el rango taxonómico al que se vaya realizar el estudio (Hillis et al. 1996). En la búsqueda de marcadores moleculares adecuados se persigue un equilibrio entre variabilidad (tasa de mutación) e información (superposición de cambios). En el genoma nuclear se ha impuesto el uso de la región espaciadora del ADN ribosómico ITS (Baldwin et al. 1995, Feliner & Roselló 2007) y en menor medida algunos genes con bajo número de copias (Álvarez & Wendel 2003, Hughes et al. 2006). La dificultad en obtener secuencias de genes de copia simple, que están exentos de muchos de los problemas evolutivos de los de copia múltiple, ha reducido su uso en filogenias (Álvarez & Wendel 2003,

Hughes et al. 2006). Sin embargo, y teniendo en cuenta las limitaciones, las regiones de copia múltiple proporcionan reconstrucciones fiables una vez consideradas sus limitaciones (Feliner & Roselló 2007). En cualquier caso, se están realizando avances en la criba de marcadores moleculares de cara a poder realizar filogenias con diferentes marcadores nucleares (Hughes et al. 2006). Desde Taberlet et al. (1991) se ha profundizado en el estudio y detección de regiones del ADN plastidial (Small et al. 1998, Shaw et al. 2005, 2007). Por este motivo, las posibilidades son mucho mayores y la selección de la región pasa a ser un punto primordial en el diseño del experimento (Small et al. 1998, Shaw et al. 2005, 2007). El gran tamaño del genoma mitocondrial en plantas, junto con la existencia de secuencias de ADN de otros organelos (plastos), presencia efímera de grandes regiones duplicadas, inestabilidad estructural, transferencia de genes al núcleo y baja tasa de cambio ha limitado el uso del genoma mitocondrial en filogenias moleculares en plantas (Palmer 1992, Soltis & Soltis 1998).

Tamaño muestral. Una de las fuentes de error más frecuentes en las filogenias se deriva del *taxon sampling effect*, sobre todo a nivel específico (Hughes et al. 2006). No muestrear todas las especies puede incurrir en errores produciendo sesgos al eliminar los eventos más recientes de especiación o los que afectan a especies 'raras' (Nee et al. 1994). Ni qué decir tiene que todo estudio parte del error de no poder tratar las especies extintas. Aumentar el número de muestras aumenta la probabilidad de obtener el árbol correcto (Wheeler 1992), se rompen la atracción de ramas largas lo que hace que la homoplasia se disperse facilitando el reconocimiento de la señal filogenética (Hillis et al. 1996), identificándose los caracteres con homoplasia global pero localmente informativos (Wenzel & Siddall 1999).

Selección del grupo externo (*outgroup*). El reducido número de cambios posibles en los datos moleculares incrementa sensiblemente la probabilidad de homoplasia (convergencia y reversión). El grupo externo debe ser seleccionado en concordancia con las tasas relativas de evolución y los tiempos relativos de divergencia para minimizar problemas de homoplasia y aumentar la probabilidad de obtener el árbol correcto. Grupos externos lejanos del grupo interno alteran la longitud relativa de las ramas lo que puede generar topologías desequilibradas y alterar la topología interna del grupo de estudio debido al fenómeno de atracción de ramas largas (*long branched attraction*, Felsenstein 1978; Wheeler 1990) y por lo tanto disminuir la probabilidad de alcanzar el árbol correcto. El *outgroup* debe estar compuesto por diferentes táxones de los distintos grupos taxonómicos cercanos para evitar la aparición de falsas sinapomorfías en el grupo de estudio (Smith 1994, pero véase Nixon & Carpenter 1993). Es deseable también, contar con una representación relativamente exhaustiva dentro del grupo externo ya que minimiza la atracción de ramas largas. Por este motivo es preferible aumentar el muestreo del grupo externo mediante la adición de secuencias al grupo hermano (*sister-group*) mejor que aumentar el número de grupos distantes. En cualquier caso, la inclusión del todas las muestras de categoría taxonómica por encima de la del grupo interno es recomendada de cara a esclarecer las relaciones de grupo hermano y obtener una filogenia de confianza.

Alineamiento (véase tema 3.2). El reconocimiento de regiones 'homólogas' (con mismo origen y disposición) es de vital importancia pues establece la hipótesis de homología primaria de modo que tiene un gran impacto en las reconstrucciones filogenéticas resultantes (Simmons et al. 2001). El reconocimiento de estas regiones es importante dado que sólo por azar dos secuencias de ADN puede presentar hasta un 25% de identidad (Simmons & Freudenstein 2003). Los errores de alineamiento que afectan a la homología y a la superposición de cambios (*multiple hits*) en una rama, suponen graves errores en la resolución de los mismos en la topología de los árboles y en los apoyos de las ramas (Simmons & Freudenstein 2003). Existen algunos programas

(MUSCLE, Edgar 2004; MALIGN, Wheeler & Gladstein 2000; POY, Gladstein & Wheeler 1996; DIALIGN, Morgenstern et al. 1998; Clustal W, Thompson et al. 1994) que implementan distintos algoritmos para el alineamiento automático de las secuencias. Sin embargo, los algoritmos desarrollados no resuelven satisfactoriamente los alineamientos múltiples (Lee 2001, Soltis & Soltis 2003). Por lo que todo alineamiento automático debería ser posteriormente revisado manualmente (Doyle & Gaut 2000, Simmons & Ochoterena 2000). El alineamiento de secuencias implica la incorporación de datos inciertos (*missing data*) y caracteres generados como producto del alineamiento (*gaps*). El tratamiento de estos datos producto del establecimiento de homologías posicionales es también muy importante de cara a las reconstrucciones filogenéticas (Simmons & Ochoterena 2000).

Selección de los métodos de análisis. La selección del método de análisis viene determinada por el tipo de datos y sobre todo por la pregunta planteada. En cualquier caso, es común y deseable evaluar distintos métodos (Doyle & Gaut 2000).

Medidas de apoyo y confianza. Existen distintas medidas de apoyo para las ramas de los árboles obtenidas mediante técnicas de remuestreo (*bootstrap* y *jacknife*, para MP, ML y NJ; *Bremer support* o índice de decay (Bremer 1994), para MP; o *Posterior Probabilities*, para BI). Estas medidas no pueden ser tomadas como estrictamente estadísticas, pero sí como una estimación de la robustez de las ramas. Existen estudios experimentales que recuperan valores bootstrap de 70% para clados reales (Hillis & Bull 1993); sin embargo, suelen tomarse valores iguales o superiores a 90% bootstrap como señales fuertes de apoyo. Las probabilidades a posteriori proporcionadas por la inferencia bayesiana superiores al 0.95 (95%) son igualmente las más fiables (Murphy et al. 2001, aunque véase Suzuki et al. 2002). Aunque se han detectado falsos positivos en PP cuando se usa un modelo evolutivo sencillo (Cummings et al. 2003) no detectados en apoyos *bootstrap* de ML y sobreestimas en PP cuando existe una estrecha relación entre las secuencias (Suzuki et al. 2002).

Selección del árbol consenso (véase tema 3.3). Los tres métodos de consenso más usados son el estricto (*Strict consensus tree*; Nelson 1979), semiestricto (*Semistrict consensus tree*; Bremer 1990), recomendado cuando los terminales son muy parecidos o cuando se utilizan distintas fuentes de datos y mayoritario (*Majority Rule consensus tree*; Margush & McMorris 1981). La combinación del consenso estricto con el mayoritario puede aportar información sobre señales filogenéticas débiles.

Selección del modelo evolutivo (véase tema 3.4). En la actualidad hay numerosos modelos de substitución nucleotídica (modelos evolutivos; Hillis et al. 1996); muchos de los cuales son submodelos de unos principales. Estos modelos evolutivos se utilizan para describir los cambios de las secuencias generalmente a través de la estimación de parámetros (frecuencia de bases, intercambio de bases y tasa de heterogeneidad). De esta forma, se consiguen los modelos evolutivos que mejor se ajustan al tipo de datos manejado. En los métodos de inferencia filogenética que asumen un modelo evolutivo de cambio (distancias, ML y BI) la selección del modelo evolutivo tiene un gran impacto en los árboles recuperados (Sullivan & Swofford 1997; Whelan et al. 2001). Modelos simples pueden subestimar las longitudes de las ramas (Yang 1994, Whelan et al. 2001). Los tres criterios más empleados para seleccionar el modelo evolutivo son hierarchical likelihood ratio test (hLRT), Akaike information criterion (AIC) y bayesian information criterion (BIC) (Posada 2001, Posada and Buckley 2004).