ASPECTOS FILOSÓFICOS DEL ANÁLISIS DE DATOS EN SISTEMÁTICA MOLECULAR

JULI CAUJAPE-CASTELLS, JOAN PEDROLA-MONFORT Y NURIA MEMBRIVES

Estació Internacional de Biologia Mediterrània - Jardí Botànic Marimurtra. Ap. de correus 112. 17300 Blanes (Girona) España.

Recibido: enero 1999

Palabras clave: Sistemática Molecular, Epistemología, Parsimonia, Máxima verosimilitud, Árbol filogenético, Refutacionismo, Verificacionismo.

Key words: Molecular Systematics, Epistemology, Parsimony, Maximum Likelihood, Phylogenetic tree, Refutationism, Verificationism.

RESUMEN

Incidimos en algunos aspectos epistemológicos básicos de las dos metodologías de análisis con datos de secuenciación en Sistemática Molecular: parsimonia y máxima verosimilitud. El hecho de que el árbol filogenético verdadero sea desconocido e irreconocible hace que la posición representada por la parsimonia sea actualmente la única herramienta epistemológicamente válida para obtener una topología de relaciones refutable. Aunque las aproximaciones de máxima verosimilitud no parecen aplicables a la selección de un árbol filogenético, sí pueden proporcionarnos hipótesis referentes a los datos utilizados para construir esas topologías de relaciones. Sería deseable un marco híbrido de análisis sistemático molecular que permitiera usar el poder predictivo de las inferencias probabilísticas basadas en los datos para dirigir el criterio de decisión de la parsimonia.

ABSTRACT

We highlight some basic epistemological aspects of the two analytical methodologies for sequence data in Molecular Systematics: parsimony and maximum likelihood. The fact that the true phylogenetic tree is unknown and non-recognizable brings out parsimony as the only current epistemologically valid tool to choose a refutable topology of relationships. Although maximum likelihood approaches do not seem applicable to selecting a phylogenetic tree, they can provide us with hypotheses referring to the data used to build those topologies of relationships. A hybrid analytical framework for molecular systematics that used the predictive power of probabilistic inference based on data to tailor the decision criterion of parsimony would be desirable.

ISSN 0211-7150

INTRODUCCIÓN

La Sistemática Molecular trata de comprender el origen y diversificación de la vida a partir de la información contenida en la molécula de DNA. Este objetivo se construye en torno a dos postulados básicos: que la vida en la Tierra tiene un sólo origen y que los organismos contienen caracteres heredables potencialmente informativos de su historia evolutiva. La Sistemática Molecular se basa, por tanto, en la premisa de que el DNA puede suministrarnos información evolutiva. Dicho de otra forma, los organismos contienen un código evolutivo además de un código genético.

No es tarea sencilla trabajar con el código evolutivo. La expresión e interpretación del código genético son «casi» universales, pero las del código evolutivo varían dependiendo de las contingencias históricas y biológicas que han sufrido las entidades taxonómicas después de su origen común. Existen diferentes códigos evolutivos en función del grado de relación entre los organismos (o entidades taxonómicas) comparados. Podría, de hecho, establecerse un paralelismo entre los "códigos evolutivos" y los idiomas mediante los cuales se comunican diferentes grupos humanos. Por ello, mientras los caracteres del código genético son informativos al nivel de cualquier organismo individual, los del código evolutivo no siempre tienen sentido fuera del conjunto de entidades taxonómicas comparadas. Acceder al código evolutivo supone muestrear una representación necesariamente insuficiente del problema sistemático a analizar, ya que normalmente no podemos trabajar con organismos ya extintos. Descifrarlo implica comparar los organismos muestreados en términos de datos incompletos: es imposible disponer de la secuencia entera del genoma o conocer todos los eventos históricos que han afectado a los organismos bajo consideración.

Uno de los más recientes avances de la Sistemática Molecular ha sido el desarrollo de métodos que nos permiten casar ideas de cambio evolutivo con procesos estadísticos. De esta manera, se hace posible testar formalmente aspectos de la evolución de los organismos teniendo en cuenta las inevitables carencias de nuestro muestreo. Pero el alcance de estas posibilidades se restringe si una buena formación estadística no va acompañada de la base epistemológica necesaria. Para los Sistemáticos Moleculares, la epistemología adquiere mayor relevancia que para cualquier otro Biólogo, porque intentamos entender un fenómeno único (la Evolución), que no podemos observar ni reproducir.

Cuando no tenemos la capacidad de replicar el fenómeno que estudiamos, nuestros errores se hacen más difíciles de detectar y de corregir; tienden a amplificarse. Por esta razón, debemos ser capaces de calibrar el alcance de nuestras conclusiones mediante la evaluación crítica de los métodos que usamos para obtenerlas. En este trabajo examinamos los aspectos epistemológicos más relevantes de las dos metodologías de análisis filogenético de secuencias de DNA: (parsimonia y máxima verosimilitud) utilizando el marco teórico desarrollado por POPPER (1968, 1992).

OBSERVACIONES

La sistemática molecular como disciplina científica

El concepto de Ciencia que suscribiremos aquí es el más universalmente aceptado hoy en día y emana de la solución de Popper al problema de la demarcación; es decir, el de hallar un criterio para establecer el carácter científico de las teorías. Históricamente, esta cuestión tiene su punto de partida en el descubrimiento de Hume de que es imposible justificar una ley de la naturaleza mediante la observación o el experimento, ya que trasciende la experiencia. Lo que esto significa es que nunca podemos garantizar la universalidad de una teoría aunque esté basada en un sinfín de observaciones que la corroboren, porque el número de observaciones posibles es infinito. No se puede saber si el futuro será igual que el pasado. El problema con la Ciencia surge al intentar conciliar éste hecho con la idea según la cual solamente la observación y el experimento pueden determinar la aceptación de leyes y teorías. La aparente incompatibilidad de los dos enunciados llevó a Hume a formular el llamado "problema de la inducción" en unos términos parecidos a los siguientes:

¿Cómo es posible que la Ciencia confirme sus teorías utilizando la inducción si al mismo tiempo ninguna regla puede garantizar la verdad de una generalización inferida a partir de observaciones verdaderas, por repetidas que estas sean?

Según Popper, la solución al problema de la inducción consiste en aceptar que las teorías científicas nunca se infieren en base a la acumulación de observaciones verdaderas. Popper arguye que las teorías tan sólo son conjeturas audaces (a menudo basadas en muy pocas observaciones) que sometemos a los más severos tests. Si la teoría supera nuestros tests, se la acepta provisionalmente y se produce el progreso científico. Por el contrario, la teoría debe rechazarse si no resiste nuestros tests.

La originalidad del razonamiento Popperiano consiste en desmantelar la creencia (difundidísima) de que la Ciencia procede de la observación a la teoría. Es el destino de una teoría, el avance científico, lo que se decide aplicando tests severos basados en la observación. Lo esencial de éste postulado fue bellamente plasmado por WEYL en la frase "la Naturaleza responde a nuestros experimentos con un no decisivo o con un sí inaudible". De manera más prosaica, podríamos sintetizar esta idea diciendo que aunque los datos empíricos nunca pueden verificar una teoría, son indispensables para su posible rechazo (o **refutación**). Renunciar al riesgo de la refutación excluye a nuestras ideas del ámbito de la Ciencia. La Ciencia debe ser arriesgada y refutable mediante tests. Vale decir que la razón de que este concepto de Ciencia sea el más universalmente aceptado es que aún no ha sido refutado.

A la luz de esta conclusión, afirmar que la Sistemática es una Ciencia implica ser capaz de conciliar lo singular de la historia de la vida con la pluralidad necesaria para sostener el principio de testabilidad. Para ello, debemos examinar en qué condiciones puede ser testada la descripción de eventos singulares.

En estos términos, la Sistemática es una Ciencia porque podemos efectuar retrodicciones en el sentido de POPPER (1992): predicciones testables derivadas de

la descripción de eventos únicos. Una hipótesis sistemática basada en datos moleculares del tipo "la dispersión del género *Androcymbium* en el norte de África posiblemente discurrió a partir de finales del mioceno" (CAUJAPE-CASTELLS *et al*., 2001) adquiere valor científico (a pesar de referirse a un evento único) porque:

A) es susceptible de refutación mediante la adición de nuevos datos referentes a la dispersión del género *Androcymbium*, y porque
B) su forma generalizada "la dispersión de muchos géneros de plantas en el norte de África discurrió a partir de finales del mioceno" puede convertirse en una (arriesgada) predicción a testar.

La Biología comparada en el ámbito molecular

La Biología comparada indaga en el proceso de cambio evolutivo a través del estudio de los cambios en diversos caracteres heredables. En la vertiente de la Sistemática Molecular que estamos considerando, estos caracteres son cada uno de los nucleótidos constituyentes de la secuencia de ciertas regiones de la molécula de DNA. Aunque describir cómo seleccionamos el fragmento de DNA que vamos a utilizar dista mucho del objetivo de este artículo, sí es conveniente reseñar unos pocos principios fundamentales.

A diferencia de los datos morfológicos y de otros datos moleculares, las secuencias de DNA nos permiten acumular caracteres sin tener que reinterpretar las adiciones en función de los datos ya existentes. Por esta economía en el esfuerzo y por el progresivo abaratamiento de los precios en los servicios de secuenciación, no parece arriesgado predecir un auge de este tipo de datos mayor que el que ha tenido lugar en los últimos cinco años. Pero hay que advertir desde el principio que la sencillez del uso de nucleótidos como caracteres filogenéticos es sólo aparente y, por tanto, engañosa. Además de requerirse el uso de moléculas **ortólogas** (derivadas de una molécula ancestral común a través de eventos de especiación), el análisis sistemático de datos de secuencia requiere **homología posicional**: los nucleótidos observados en una posición determinada de las entidades taxonómicas muestreadas han de derivar de la misma posición en un ancestro común de esas entidades (SWOFFORD *et al*., 1996). Escoger moléculas cuya evolución no pudiera explicarse por una continuidad hereditaria redundaría en una filogenia probablemente correcta para las moléculas usadas que diferiría marcadamente de la de los organismos a partir de los cuales las secuencias fueron muestreadas. Por este motivo, no son útiles para el análisis filogenético secuencias derivadas de una duplicación génica (denominadas **parálogas**) o las transmitidas mediante elementos transponibles o a través de otros organismos como retrovirus (secuencias **xenólogas**).

Debemos también tener en cuenta que el cambio a escala molecular se produce a diferente ritmo dependiendo de la ubicación intracelular del DNA (nuclear, mitocondrial o cloroplástico) y, para una molécula dada, en función de la región seleccionada. Nuestra primera faena ha de consistir pues en escoger la molécula de DNA y la región convenientes, de manera que su ritmo de cambio se adecue al marco temporal del problema que intentamos clarificar. Si estudiamos un nivel alto de la jerarquía biológica (esto es, organismos que divergieron hace mucho tiempo,

como familias, órdenes e incluso géneros), debemos cerciorarnos de que la secuencia de DNA que utilicemos no cambie de forma muy rápida, o nos exponemos a detectar excesiva variación para establecer una hipótesis de relaciones coherente. Las secuencias pertenecientes a genes que eventualmente dan lugar a proteínas funcionales (como por ejemplo *rbcL*) o los espaciadores intragénicos transcritos (ITS) parecerían una elección adecuada a este nivel de estudio. Si, en cambio, el problema sistemático a investigar afecta a un nivel bajo de la jerarquía biológica (organismos que divergieron más recientemente, como especies o subespecies), es aconsejable escoger regiones variables; de lo contrario, nos exponemos a no detectar variación. Las secuencias no funcionales, como por ejemplo los espaciadores intergénicos (IGS) o los microsatélites serían una primera elección obvia.

Otro requisito que la región escogida ha de cumplir es no variar apenas dentro del nivel estudiado. Lo que esto significa es que si nuestro objeto de estudio son las relaciones a nivel interespecífico, el marcador de DNA escogido ha de ser suficientemente variable para garantizar la diferenciación entre las especies pero, a la vez, no ha de presentar variación dentro de las especies. De otra forma, es probable que nuestras hipótesis relacionales variaran considerablemente dependiendo de cuántos y cuales individuos incluyéramos en nuestro estudio. Como los niveles de variabilidad para una secuencia dada acostumbran a variar dependiendo de los organismos estudiados, asegurar éste aspecto siempre ha de implicar un examen preliminar de variabilidad. El mismo caso se aplica al nivel intergenérico o interfamiliar. ¿Existe variabilidad para la región de DNA escogida dentro del mínimo nivel de la jerarquía que estudiamos? Esta es la pregunta que debemos responder ne-

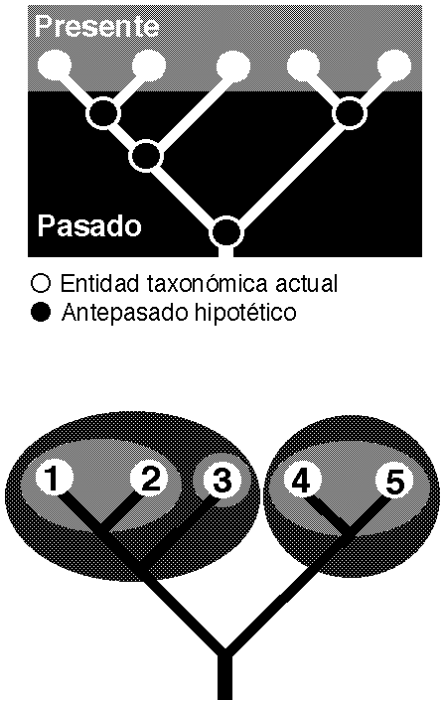

Figura 1.-
Un árbol filogenético es una jerarquía de relaciones en la cual se infiere una hipótesis sobre el pasado común de un grupo de entidades taxonómicas a través de los cambios detectados en caracteres estudiados en el presente. Un grupo de entidades taxonómicas que contiene a todos los descendientes de un mismo antepasado se denomina monofilético. Dependiendo del antepasado considerado, un grupo monofilético estará constituido por un diferente número de organismos. En la figura de la derecha se simbolizan las diferentes agrupaciones monofiléticas (simbolizadas por tonalidades de gris) que resultan en base a los diferentes antepasados hipotéticos. Todos los organismos en el árbol no forman un grupo monofilético porque esta agrupación se apoyaría exclusivamente en simplesiomorfías (caracteres compartidos no derivados).

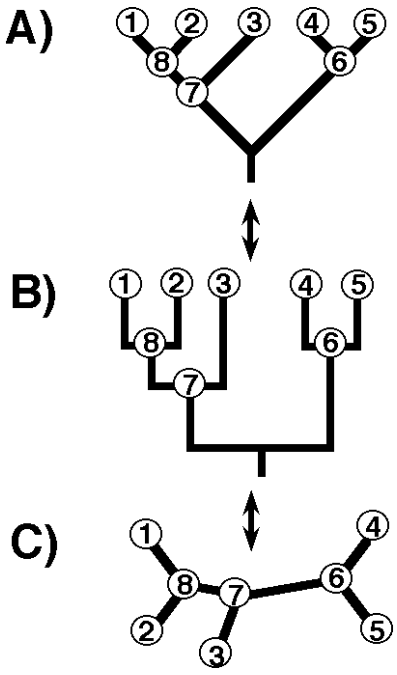

Figura 2.-
Cualquier árbol filogenético puede representarse gráficamente de varias maneras topológicamente equivalentes. En la figura se muestran las tres más usadas: cladograma (A), cladograma rectangular (B) y árbol no enraizado (C). Los cinco nodos terminales (numerados de 1 a 5) representan las entidades taxonómicas. Los tres nodos internos 6, 7 y 8 simbolizan los diferentes antepasados hipotéticos de los clados correspondientes. Este árbol contiene dos clados monofiléticos representados por los organismos (1, 2, 3) y (4, 5) respectivamente y puede notarse también en forma parentética como ((1, 2, 3),(4, 5))

gativamente para tener alguna certeza de no errar en nuestra elección.

La reconstrucción de la historia evolutiva en Sistemática se plasma en el árbol filogenético: una estructura jerárquica de ramificaciones que permite representar hipótesis relacionales a partir de los cambios detectados en caracteres heredables (Figs. 1 y 2).

Este procedimiento tiene su origen en el conjunto de reglas formuladas por Willi Hennig. HENNIG (1966) fue el primero en discriminar entre caracteres derivados compartidos (**sinapomorfías**), caracteres derivados únicos de una entidad taxonómica (**autapomorfías**) y caracteres primitivos compartidos (**simplesiomorfías**), es decir, caracteres presentes en todas las entidades taxonómicas consideradas. El procedimiento Hennigiano usa solamente las sinapomorfías para construir árboles filogenéticos y asume que conocemos los estados ancestrales para cada carácter. Además, considera que la evolución de los estados de carácter es irreversible y que cada carácter puede cambiar solamente una vez en el árbol filogenético verdadero. En el contexto de estas asunciones, cada carácter define un **grupo monofilético**: una agrupación de organismos que contiene todos los descendientes de un mismo antepasado (Fig. 1).

Tal como los enunció Hennig, los criterios para la reconstrucción de la historia evolutiva son muy estrictos para poder ser aplicados; nunca podemos conocer con seguridad los estados de carácter ancestrales, o construir un árbol filogenético con todos los caracteres si estos pueden cambiar una sola vez. La parsimonia y la máxima verosimilitud son las dos estrategias más comúnmente utilizadas para establecer hipótesis filogenéticas relajando las asunciones de Hennig.

Ambas poseen dos puntos en común. En primer lugar, utilizan una o varias entidades taxonómicas como grupos externos ('**outgroups**') para estimar los estados de carácter ancestrales. Con ello se asume que los caracteres que se hallan en los 'outgroups' son más parecidos a los del desconocido antepasado común del grupo a analizar ('**ingroup**') (así se denomina al conjunto de entidades taxonómicas cuyas relaciones pretendemos averiguar) por haberse producido la divergencia entre 'outgroup' e 'ingroup' en un punto temporal más cercano a dicho antepasado.

En segundo lugar, tanto parsimonia como máxima verosimilitud permiten que los caracteres evolucionen más de una vez. Con ello, surge el problema de que todos los árboles posibles van a encajar con nuestros datos. Para percibir la magnitud del obstáculo que esto puede significar, basta ver que el número total de árboles no enraizados estrictamente bifurcados para 'T' ramas terminales (entidades taxonómicas) viene dado, según FELSENSTEIN (1978a), por el producto:

$$B(T) = \prod_{i=3}^{T} (2i - 5)$$

Para las 5 entidades taxonómicas representadas en la Fig. 1, existen 15 árboles posibles. Para 50 entidades taxonómicas, el número de árboles posibles supera el de átomos en el universo.

DISCUSIÓN

¿Lo más probable o lo 'más sencillo'?

Los cladistas adoptan el criterio de **parsimonia** para elegir entre las numerosísimas posibilidades que se nos pueden presentar en cualquier estudio Sistemático Molecular. Este criterio consiste en utilizar solamente las sinapomorfías para postular hipótesis de relaciones y escoger el árbol con el menor número de pasos, interpretados como transiciones entre estados de carácter (Figs. 3 y 4). El árbol filogenético obtenido de esta manera se denomina también **cladograma**, y puede ser refutado si la hipótesis de relaciones que propugna es rechazada por la adición de nuevas sinapomorfías al problema sistemático (o por la reconsideración de las ya utilizadas). Antes de proseguir, creemos conveniente llamar la atención sobre la identificación entre 'mayor simplicidad' y 'menor número de pasos' que se da en los artículos y libros especializados sobre técnicas filogenéticas. Advirtamos que, en la práctica sistemática del cladista, 'simplicidad' equivale a 'economía de pasos'.

Los sistemáticos cladistas seleccionan las hipótesis máximo parsimoniosas que no han sido aún refutadas, o bien aquellas que se hayan refutado un menor número de veces. En consecuencia, los sistemáticos cladistas adoptan una filosofía **refutacionista**.

Según la aproximación de la máxima verosimilitud adoptada por los probabilistas, la reconstrucción filogenética es un problema eminentemente estadístico cuya solución consiste en encontrar el árbol con más alta probabilidad de haber dado

lugar a los datos observados (Figs. 3 y 4). Para ello, se utilizan tanto las sinapomorfías como las autapomorfías, ya que mientras las primeras nos permiten establecer hipótesis de relaciones, las segundas nos informan de la magnitud de la diferencia de una entidad taxonómica respecto de las demás. La estimación filogenética mediante máxima verosimilitud debe basarse en un modelo evolutivo por el cual regular los cambios de estados de carácter. Para ello, se asigna a los datos un valor de verosimilitud (en el sentido de la teoría de la probabilidad) dependiendo de su bondad de ajuste a las asunciones contenidas en un modelo evolutivo (SWOFFORD et al., 1996). Un cambio en las asunciones biológicas cambia el modelo probabilístico y, consecuentemente, el método para seleccionar el mejor estimador de la filogenia. No hay árboles imposibles, sólo árboles más probables que otros; por tanto escogeremos el árbol de máxima probabilidad. La verosimilitud se contempla como una función del árbol, y buscamos el árbol que maximiza la verosimilitud (FELSENSTEIN, 1984). Los sistemáticos probabilistas son calificados como **verificacionistas** por escoger la hipótesis más verosímil.

Aunque estas consideraciones escasamente cubren los conceptos básicos de cada una de estas dos escuelas, sirven para ubicar el punto central de nuestra

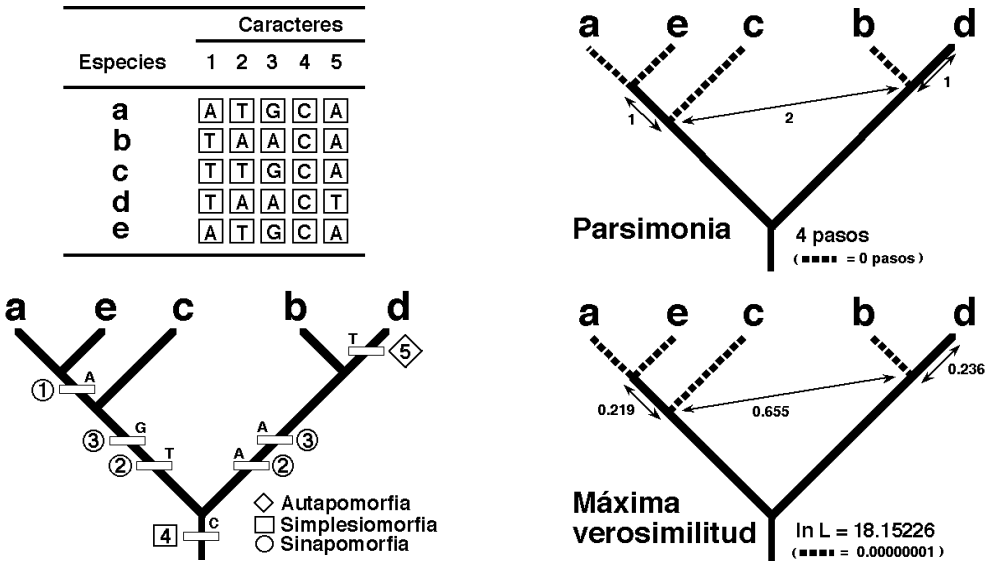

Figura 3.- A) Único árbol filogenético obtenido mediante el algoritmo de parsimonia a partir de la secuencia de cinco nucleótidos en cinco organismos hipotéticos a, b, c, d, e. La ubicación de autapomorfías, simplesiomorfías y sinapomorfías (ver texto para definiciones) se indica en el árbol con un símbolo diferente encuadrando el número de carácter. En cada caso, el nucleótido implicado está escrito encima de la barra vertical que simboliza la presencia de los caracteres.
B) En este caso, tanto parsimonia como máxima verosimilitud llegan a hipótesis únicas topológicamente idénticas. El único árbol máximo parsimonioso consta de cuatro pasos. El árbol máximo verosímil tiene un valor de verosimilitud de -ln L = 18.15226, bajo la asunción de que la proporción de transiciones y transversiones es la estimada por el programa PAUP 3.1.1. (SWOFFORD, 1993). Estos árboles sustentan la hipótesis de que los organismos 'a' y 'e' comparten un antepasado común más cercano que el que cualquiera de los dos comparte con el organismo 'c', y que 'b' y 'd' están más relacionados entre sí que respecto a 'a', 'c' y 'e'.

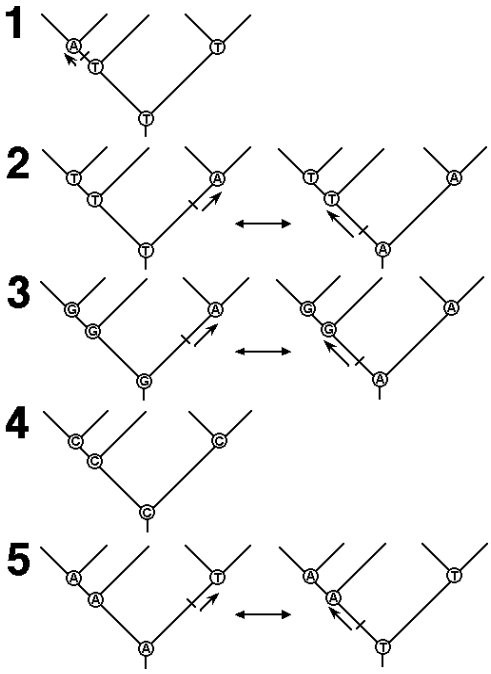

Figura 4.-
Cambios requeridos para cada uno de los caracteres en cada uno de los 3 nodos internos del árbol de la Figura 3. Los cambios se simbolizan mediante una barra perpendicular en un lugar arbitrario de la rama afectada acompañado de una flecha en el sentido del cambio. Los nodos internos contienen el nucleótido asignado al antepasado hipotético correspondiente. Cuando hay dos posibilidades de cambio (caracteres 2, 3 y 5), se representan ambas (aunque para evaluar el número de pasos, solo se tiene en cuenta una de ellas).

discusión. La parsimonia puede no ser la opción más probable, pero lo improbable es posible y lo posible puede ocurrir. ¿Es más lógico proponer las hipótesis filogenéticas basándonos en la explicación más sencilla o a partir de la estimación más verosímil? He ahí el dilema.

Demarcación entre verificacionismo y refutacionismo

Dado que tanto verificacionismo como refutacionismo se refieren a una disciplina científica, examinarlas a la luz del concepto lógico de testabilidad puede proporcionarnos un **criterio de demarcación**.

El poder de testar hipótesis es una función de la interacción entre la evidencia disponible (E), la hipótesis (H) y la base conceptual (B). Dados estos parámetros, y siempre que E sea posible de acuerdo con B, el grado de corroboración se define mediante la expresión (POPPER, 1968):

$$C(H,E,B) = \frac{P(E,HB) - P(E,B)}{P(E,HB) - P(EH,B) + P(E,B)} \qquad (1)$$

donde P(E, HB): probabilidad de E dados H y B; P(E, B): probabilidad de E dada solamente B; P(EH, B): probabilidad de E según H, dada B.

Aún sin entrar en detalles acerca de esta fórmula, podemos ver que cuanto más bajo sea P(E, B), mayor será el nivel de corroboración de H. Por tanto, E debe de ser improbable dada B para que H reciba corroboración de E. La conclusión

lógica de éste examen de la testabilidad es que hemos de preferir las evidencias más improbables dentro de las posibles, puesto que son más corroborables.

A nivel del análisis lógico de la corroboración, la demarcación entre refutacionismo y verificacionismo reside en el diferente riesgo asumido al proponer una hipótesis filogenética. Adoptamos una postura refutacionista a medida que p(E, B) disminuye; entramos en los dominios verificacionistas cuando p(E, B) aumenta (KLUGE, 1997). La postura refutacionista es más arriesgada que la verificacionista. No sólo eso, sino que se pone de manifiesto que la interpretación lógica de la teoría de la probabilidad es incompatible con la interpretación frecuentista (la utilizada por los verificacionistas), de acuerdo con la cual habríamos de preferir siempre las evidencias más probables.

Críticas del refutacionista al verificacionista

Una de las críticas de más calado a los verificacionistas es la que cuestiona la asignación de probabilidades a una hipótesis acerca de relaciones genealógicas entre organismos que, por definición, no son independientes y para las que sólo existe un árbol verdadero posible.

Que todos los organismos tienen un origen común es inconstatable, pero no está en tela de juicio. El apoyo más fuerte de que disponemos para sostener esta hipótesis es la (casi) universalidad del código genético. Si el origen de todos los seres vivos no fuera común, resultaría imposible entender que la síntesis de proteínas en organismos tan diferentes como la cabra cimarrona y la alcachofa proceda basándose en el mismo código genético.

La cuestión a la que nos enfrentamos es: ¿cómo decidir si un árbol filogenético es más probable que otro si nunca podemos disponer de un muestreo aleatorio de una población de fenómenos evolutivos que afecten al mismo grupo de organismos? Desde una perspectiva puramente lógica, si la cuestión de la probabilidad de una hipótesis filogenética pudiera interpretarse como una probabilidad de eventos, deberíamos poder asimilarla a una fórmula del tipo (POPPER, 1968)

$$P(\textit{hipótesis}) = \frac{\text{contrastaciones superadas}}{\text{contrastaciones posibles}} \qquad (2)$$

Evidentemente, no es posible estimar el denominador de esta expresión de modo preciso y, aún en ese caso, el resultado de tal 'probabilidad' sería siempre cero por ser infinito el número de contrastaciones posibles.

Podríamos no rendirnos ante esta refutación y sugerir la alternativa

$$P(\textit{hipótesis}) = \frac{\text{contrastaciones favorables}}{\text{contrastaciones indiferentes}} \qquad (3)$$

Pero esto devaluaría el concepto de probabilidad de hipótesis al equipararlo a algo totalmente subjetivo, más dependiente de los conocimientos y habilidad del experimentador que de resultados tangibles. Agotadas las opciones, debemos concluir forzosamente que no tiene sentido utilizar las cuestiones basadas en lógica probabilística para abordar el concepto de probabilidad de una hipótesis de

relaciones, ya sea esta filogenética o de cualquier otro tipo. Bajo ninguna circunstancia podemos traducir un enunciado sobre la probabilidad de una hipótesis por otro acerca de la probabilidad de eventos (POPPER, 1968).

Puesto que solamente existe una solución al fenómeno de la evolución (esto es, sólo hay un árbol verdadero posible), cualquier hipótesis de relaciones filogenéticas que propongamos puede ser verdadera o falsa, pero no más o menos probable. Conviene, pues, dejar claro que la probabilidad esgrimida por los partidarios de la máxima verosimilitud no es “la del árbol dados los datos”, sino “la de los datos dado el árbol” (KLUGE, 1997). Y aunque ésta es una precisión muy importante, entenderla no basta para contrarrestar el hecho de que, en la práctica Sistemática, en ningún caso nos es dado el árbol verdadero.

La aproximación verificacionista parece además criticable al menos desde dos puntos de vista a la luz del desarrollo Popperiano de la corroboración, según la cual aquella maximiza la probabilidad de que la evidencia E sea compatible con la hipótesis H incrementando la base conceptual B. En primer lugar, tal incremento de la base conceptual mediante la adición de parámetros al modelo se hace al coste de disminuir el valor de la corroboración. Por ello, esta línea de acción conlleva el riesgo de reducir nuestra hipótesis a una tautología; a una afirmación que no dice nada nuevo.

En segundo lugar, la transferencia subyacente de hipótesis a la base conceptual es lógicamente errónea, al hacerse ora a expensas de una corroboración inexistente (si las hipótesis transferidas no han sido testadas), ora partiendo de una identificación entre corroboración y verificación (en caso de que las hipótesis transferidas hayan sido testadas y corroboradas).

A pesar de estas críticas, los verificacionistas sostienen que estimar la filogenia mediante máxima verosimilitud posee varias propiedades deseables (HUELSENBECK, 1998). Las dos más importantes son la **consistencia** (convergencia al valor verdadero a medida que se añaden más datos al problema) y la **eficiencia** (poca variación respecto al valor verdadero). Pero de nuevo tropezamos con la imposibilidad de disponer del árbol verdadero y debemos reconocer que tales propiedades, aunque deseables, pierden mucho de su significado en la práctica Sistemática.

Los verificacionistas deben aún explicar cómo soslayan la asunción de independencia (indispensable para aplicar razonamientos probabilísticos) si, al mismo tiempo, han de aceptar como parte de su base conceptual que todos los organismos tienen un origen común.

Críticas del verificacionista al refutacionista

En ciertas circunstancias, la estimación de una hipótesis filogenética puede converger a una solución incorrecta a medida que se agregan caracteres al problema. Se dice entonces que tal estimación es inconsistente. La región del espacio paramétrico donde un método de estimación filogenética resulta inconsistente se denomina 'zona Felsenstein', ya que fue éste autor quien llamó la atención de la comunidad sistemática sobre el fenómeno (FELSENSTEIN, 1978b). Los verificacionistas aseveran que la parsimonia es un método que puede dar lugar a estimaciones inconsistentes de la filogenia (Fig. 5).

Existen dos argumentos objetivos que debilitan considerablemente esta crítica. En primer lugar, recientes investigaciones han descubierto que la máxima verosimilitud también es susceptible de incurrir en esta indeseable situación, y en condiciones incluso menos restrictivas que las planteadas por Felsenstein en su artículo

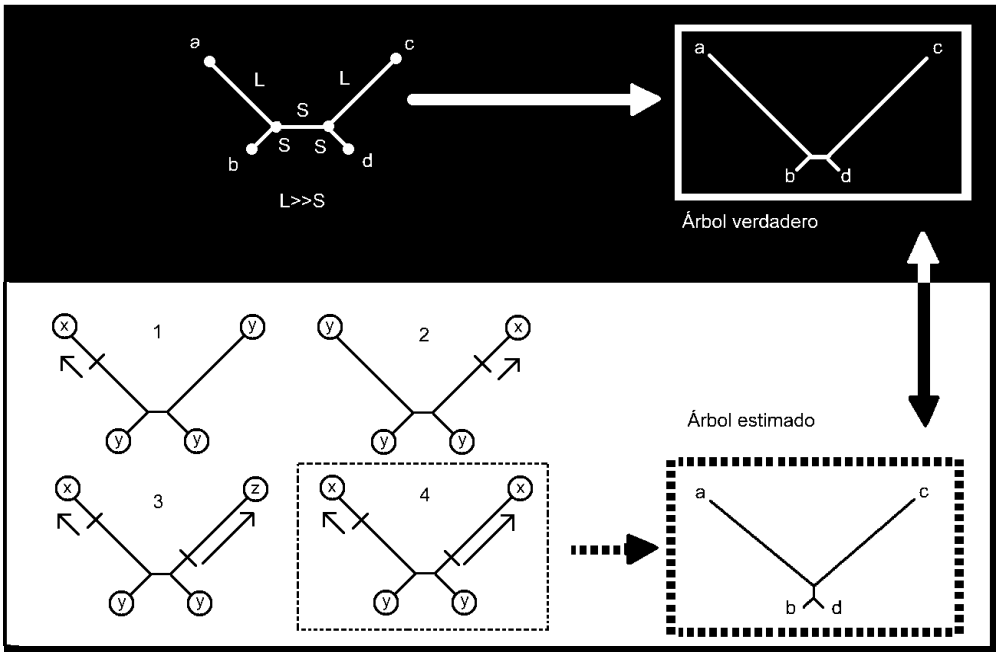

Figura 5.- Ejemplificación de la problemática de la 'zona Felsenstein' en parsimonia mediante un árbol no enraizado de cuatro nodos terminales (a, b, c, d) con sólo dos longitudes de ramas posibles: ramas cortas (S) y ramas largas (L), con L >>S. Cuando dos ramas periféricas opuestas son muy largas y las dos restantes muy cortas, el método de parsimonia puede converger a una solución que conecte incorrectamente las dos ramas más largas.
La probabilidad de que un cambio ocurra a lo largo de las ramas más cortas (S) es muy pequeña comparada con la probabilidad de que tenga lugar en una de las ramas más largas (L). Cuando S es extremadamente pequeño, podemos ignorar la posibilidad de que se produzca un cambio en las ramas cortas, en cuyo caso observaremos el mismo nucleótido en los nodos 'b' y 'd'. En las ramas largas, existen tres patrones posibles: a) que no haya cambios, en cuyo caso observaremos 'yyyy' (donde 'y' representa el nucleótido asignado y puede ser A, C, G o T) en los nodos terminales n1, n2, n3, n4 respectivamente; b) que solo tenga lugar un cambio en las ramas largas, en cuyo caso los patrones posibles son (1) 'xyyy' o (2) 'yyxy' (donde 'y' representa un nucleótido diferente de 'x'); c) que ocurra un cambio en ambas ramas largas. Entonces el patrón sería 'xyzy' (3) si el cambio se produce hacia diferentes nucleótidos, o 'xyxy' (4) si el cambio es hacia el mismo nucleótido.
Para el método de parsimonia, de entre todos los patrones de cambio posibles, sólo 'xxyy', 'xyxy' o 'xyyx' son eventualmente informativos y pueden distinguir entre diferentes árboles, ya que el resto de cambios posibles no dan lugar a sinapomorfías. Los patrones de cambio cuando L>>S incluyen 'xxxx', 'xyyy', 'xxyx', 'xyzy' y 'xyxy'; solamente 'xyxy' es informativo para la parsimonia, y lo es en la dirección del árbol incorrecto (mostrado en el recuadro con línea discontínua). Para una mejor percepción, el fondo del espacio 'verdad' se ha coloreado en negro, mientras que el del espacio 'estimación' se ha coloreado en blanco. Nótese que, a diferencia de cualquier situación real, esta demostración de inconsistencia asume que conocemos el árbol verdadero.

pionero. En máxima verosimilitud, las ramas largas no solamente son problemáticas cuando ocurren en partes opuestas del árbol verdadero, sino también cuando son adyacentes. En este último caso, se hace particularmente difícil distinguir entre una estimación de la filogenia que presente las dos ramas largas separadas y una donde las dos ramas largas estén juntas (HUELSENBECK, 1998). El problema de la inconsistencia tiene pues una fuerte componente histórica: como Felsenstein demostró la inconsistencia de la parsimonia, se ha criticado a la parsimonia por inconsistente.

En segundo lugar, la existencia de la temida 'zona Felsenstein' aún no se ha encontrado en la práctica, por lo cual se desconoce si tan sólo se trata de una emanación teórica.

A todas luces, la controvertida zona Felsenstein parece más bien un problema general de la estimación filogenética que una incoherencia inherente a la adopción de una metodología particular. El problema planteado es de detección, y radica en determinar a partir de qué diferencia de longitud de las ramas un método filogenético entra en la zona de inconsistencia. La solución parece difícil, ya que para ello deberíamos conocer la topología del único árbol filogenético verdadero

Quizás la crítica más justificada de los verificacionistas al razonamiento refutacionista sea la asunción de que la evolución procede parsimoniosamente. Los refutacionistas responden diciendo que llevan a cabo el objetivo de obtener máxima corroboración a partir únicamente de la inclusión de la herencia con modificación de los caracteres en la base conceptual (Kluge, 1997). En este sentido, escoger la hipótesis más sencilla (más parsimoniosa) que explique nuestros datos representa una estrategia para obtener máxima corroboración. La hipótesis más sencilla es también la más predictiva, y es precisamente en éste sentido que la predictividad de una hipótesis cladista puede ser maximizada. Por ello, según el programa refutacionista, el principio de parsimonia ha de interpretarse como una consecuencia lógica de la aplicación de la filosofía Popperiana a la Sistemática y no como un componente de su base conceptual. Hemos de preferir las hipótesis más parsimoniosas porque son más fáciles de refutar en caso de que sean falsas.

CONCLUSIONES

Si conociéramos (o pudiéramos reconocer) el árbol verdadero no existirían las discrepancias metodológicas entre parsimonia y máxima verosimilitud y quizás la Sistemática Molecular sería una rama de la estadística. Ante el amplio abanico de eventos históricos que pueden haber afectado a las diferentes entidades taxonómicas y dada la variabilidad de los atributos biológicos de estas entidades, no es posible diseñar una estrategia de análisis universalmente válida. Es lugar común que intentar someter la plasticidad de la historia de la vida a la burda rigidez estadística es un ejercicio comúnmente condenado al fracaso. En Sistemática, ello es especialmente cierto y puede convertirse además en una excelente excusa para reinterpretar nuestras conclusiones en relación al conjunto orgánico de íntimas propensiones que siempre acompaña nuestros esfuerzos científicos. Las ya numerosas herramientas metodológicas que la Sistemática pone a nuestro alcance han de servir para hacernos cada vez más difícil incurrir en éste peligroso sesgo sub-

jetivo. Tal es su potencial distorsionador que algún destacado sistemático molecular ha propuesto que se oculten los nombres de las entidades taxonómicas durante el proceso de decisión sobre el muestreo (HILLIS, 1998). Cabe dentro de lo posible que los programas de análisis filogenético incorporen en el futuro una opción que permita también éste tipo de estrategia durante el proceso de análisis de datos.

Cuando dos posturas como verificacionismo y refutacionismo mantienen sus diferencias de manera tan irreconciliable, ambas deben tener al menos una alícuota de razón. Los enunciados refutacionistas poseen mayor contenido informativo puesto que asumen un mayor riesgo (recordemos que la parsimonia puede no ser la opción más probable). Este es un punto importante porque todos nosotros pensamos según la interpretación frecuentista de la probabilidad, que nos hace intuir que siempre hemos de pronosticar lo más probable. En Sistemática Molecular, esta intuición es equívoca. Hemos de preferir la facilidad de refutación de los enunciados más arriesgados ante la imposibilidad de establecer la probabilidad de nuestras hipótesis sobre conjuntos de organismos. Aunque esta es una ventaja epistemológica muy destacable del refutacionismo, no ha de hacernos sobrevalorar sus logros. Esta metodología nos dota de un criterio necesario para seleccionar un árbol de relaciones solamente a cambio de impedirnos testar si los datos que utilizamos para generar el árbol se adecuan a ese criterio.

En contrapartida, los postulados verificacionistas, basados en la teoría de la probabilidad, no son aplicables a la selección del árbol filogenético. Lo que esto significa es que no podemos asimilar la evolución al paradigmático dado con que se acostumbra a ilustrar los razonamientos probabilísticos elementales. Un 'dado evolutivo' no es un buen símil porque nosotros no somos observadores pasivos, sino sujetos que influyen en los movimientos del dado a la par que giramos con (dentro de) él. Si, por familiaridad, queremos ver la evolución como un dado virtual, hemos de ser conscientes de que éste puede tener muchas caras pero un sólo resultado posible, que es invisible e irreconocible. Y de que es éste un dado muy energético que empezó a moverse antes del principio de la vida y que estará girando pertinazmente hasta el fin del tiempo; no podremos nunca, por tanto, anotar el resultado de la "tirada" y lanzarlo de nuevo. Tan radical conclusión solamente mengua la validez de la aproximación verificacionista en lo que se refiere a la selección de árboles filogenéticos. La capacidad para indagar en los procesos que han dado lugar a estos árboles queda intacta, porque los tests y razonamientos probabilísticos sí pueden ser aplicados a los datos. Volviendo a utilizar el símil del dado, no sabremos nunca el resultado de la "tirada" pero quizás sí podremos aplicar la teoría de la probabilidad a la dinámica de los movimientos pasados del dado en términos de sus efectos en los organismos, de manera que se nos haga posible descartar ciertas configuraciones relacionales improbables.

Los innegables avances conceptuales de verificacionismo y refutacionismo parecen fútiles si no se aprovechan para intentar evaluar críticamente cuales de sus múltiples insuficiencias son complementarias. La ventaja más evidente de una actitud conciliadora entre parsimonia y máxima verosimilitud es la posibilidad que se vislumbra de averiguar en qué condiciones los organismos analizados pueden haber evolucionado de manera parsimoniosa y en qué condiciones la asunción de parsimonia es injustificada. La definición de un marco conceptual que permita utili-

zar nuestras inferencias estadísticas sobre los datos para constreñir los análisis basándose en la asunción de parsimonia nos parece de gran valor para el avance del razonamiento Sistemático.

AGRADECIMIENTOS

Agradecemos a D. Águedo Marrero su revisión crítica que nos ayudó a aclarar algunos puntos confusos en versiones anteriores del manuscrito.

REFERENCIAS

CAUJAPE-CASTELL, J., R. K. JANSEN, N. MEMBRIVES, J. PEDROLA-MONFORT, J. M. MONTSERRAT & A. ARDANUY, 2001.- Historical Biogeography of *Androcymbium* Willd. (Colchicaceae) in Africa: Evidence from cpDNA RFLPs. *Bot. J. Linn., Soc*. 136:379-392.

FELSENSTEIN, J., 1978a.- The number of evolutionary trees. *Systematic Zoology,* 27: 27-33.

- 1978b.- Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology,* 27: 401-410.

- 1984.- The statistical approach to inferring evolutionary trees and what it tells us about parsimony and compatibility. En T. Duncan y T. F. Stuessy (eds.) *Cladistics: perspectives on the reconstruction of evolutionary history*: 169-191. Columbia University Press, New York.

HENNIG, W., 1966.- *Phylogenetic Systematics*. University of Illinois Press, Urbana.

HILLIS, D. M., 1998.- Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology,* 47: 3-8.

HUELSENBECK, J. P., 1998.- Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved?. *Systematic Biology,* 47: 519-537.

KLUGE, A. G., 1997.- Testability and the refutation and corroboration of scientific hypotheses. *Cladistics,* 13: 81-96.

POPPER, K., 1968.- *La lógica de la investigación científica*. Ed. Tecnos, Madrid. 451 pp.

- 1992.- *El coneixement objectiu*. Edicions 62, Barcelona. 378 pp.

SWOFFORD, D. L., 1993.- *PAUP: Phylogenetic analysis using parsimony, version 3.1.1*. Illinois Natural History Survey, Champaign.

- G. J. OLSEN, P. J. WADDELL, & D. M. HILLIS, 1996.- Phylogenetic Inference. En Hillis, D. M, Moritz, C. y Mable, B. K. (eds.) *Molecular Systematics*, 407-514 Sinauer, Massachussets.