Revista Colombiana de Estadística

No.33 y No.34

MEJORAMIENTO ESTADÍSTICO DE ARBOLES: UN CASO DE BIODIVERSIDAD

HERNANDO CHACÓN GONZÁLEZ

ESTADÍSTICO BIOMETRISTA

INTRODUCCIÓN

La biodiversidad es un recurso genético que posibilita, entre otras, una mejor producción agropecuaria e industrial en términos de cantidad, calidad y estabilidad a través del mejoramiento varietal y la biotecnología. En este recurso, Colombia es un país privilegiado con un 0.77% de la superficie emergida del planeta, sus fronteras albergan alrededor del 10% de especies vivas de plantas y animales Higgins (1991). El racional aprovechamiento de esa riqueza biológica ha motivado desarrollos más o menos específicos en los dominios de la evolución, la taxonomía, la filogenética, la biosistemática, y más recientemente en la biología molecular, Swofford y Olsen (1990). La estadística, en tanto que herramienta del método científico, también contribuye al estudio de la diversidad genética con métodos tanto generales como específicos. Los métodos de clasificación automática o taxonomía numérica, Sneath y Sokal (1973), y más específicamente la construcción de un **árbol** -objeto matemático utilizado clásicamente para representar en forma gráfica distintos procesos- están entre estos últimos, Barthelemy y Guenoche (1988). Este trabajo versa sobre el ajuste de un árbol ilustrado en un caso de diversidad genética.

Construcción de un árbol

El árbol como caso particular de una gráfica es la manera clásica de representar procesos evolutivos, taxonómicos, filogenéticos y biosistemáticos. Para su construcción, una de las posibilidades comprende las siguientes etapas:

1. Selección, observación y registro de las características o descriptores a partir de las cuales se va a estudiar las unidades taxonómicas, UT's;

2. Adopción de una medida de semejanza entre las UT's y

3. Utilización de un algoritmo de aglomeración de la UT's a partir de su grado de semejanza.

Entre los pasos 2 y 3 de este procedimiento hay una pérdida de información. Por este motivo es necesario buscar una estructura que represente mejor el valor del indice utilizado para medir la semejanza entre las UT's; este precisamente es el objetivo del presente trabajo.

En el otro enfoque , llamado genético, " ... se hacen fuertes hipótesis sobre un modelo genético de evolución. El principio de base de estos métodos denominados de parsimonia es considerar como más probable el árbol que permite oaganizar la diversidad genética de la forma más económica en eventos mutacionales. " Chacón (1993).

Metodología

La clasificación ascendente jerárquica -CAJ-, Benzecri (1973) y NJTREE, Nei (1987) son dos algoritmos de construcción de árbol que inducen respectivamente las distancias ultramétricas y aditivas de árbol. Estas distancias son una aproximación

representable del valor del índice para medir la semejanza entre las UT's[1]. En su lógica, estos dos algoritmos buscan en cada iteración la pareja de UT's más próxima para formar una nueva clase. La forma como se define esta proximidad es la principal diferencia entre los algoritmos. En general el valor de la distancia entre dos UT's sobre el árbol difiere del valor inicial del índice, lo que se define aquí como una pérdida de información; el objetivo es entonces minimizarla. Para ello se debe evaluar el criterio -mínimos cuadrados, por ejemplo - sobre todos los árboles posibles y retener el óptimo. El crecimiento exponencial del número de árboles posibles a partir de un número todavía pequeño de UT's dificulta este procedimiento a pesar de las cada vez menores dificultades de cálculo computacional. Se demuestra que, por ejemplo, para 10 UT's hay más de 2 millones de árboles posibles.

La idea puesta en práctica es entonces hacer una optimización global conservando la topología del árbol producido por CAJ y NJTREE. Se ajusta la distancia sobre los árboles mediante el criterio de mínimos cuadrados.

Adoptemos entonces la siguiente notación:

$d(i,j)$es el valor del índice de semejanza entre las UT's i y j;

$\delta(i,j)$es la distancia sobre el árbol entre las UT's i y j.

El ajuste del árbol se hace entonces minimizando:

$$\phi = \sum_{i<j} (d(i,j) - \delta(i,j))^2$$

Se trata de resolver un sistema de tantas ecuaciones como parejas de UT's, $\binom{N}{2}$

[1]El índice utilizado para medir el grado de semejanza entre dos UT's, de acuerdo a sus propiedades matemáticas, puede ser una desviación, una disimilaridad o una distancia. La literatura reporta muchos de estos índices, Perrier (1992) y Chacón (1993); su escogencia es uno de los aspectos cruciales en el procedimiento de construir un árbol.

donde N es el número de UT's iniciales), y la longitud de las ramas como incógnitas. A estas últimas se les impone la restricción de que sean positivas de acuerdo con una de las propiedades que caracterizan formalmente una distancia.

Para medir la bondad del ajuste después de la optimización global y comparar resultados entre la distancia Ultramétrica de la Clasificación Ascendente Jerárquica y la Distancia Aditiva de Arbol de NJTREE se ulitizan varios criterios:

1. DAM: Desviación Absoluta Media

$$DAM = (1/P) \sum_{i<j\leq N} |d(i,j) - \delta(i,j)|$$

2. DAMAX: Desviación Absoluta Máxima

$$DAMAX = Max_{\,i<j\leq N} |d(i,j) - \delta(i,j)|$$

3. DARM: Desviación Absoluta Relativa Media

$$DARM = \frac{(1/P)\sum_{i<j\leq N} |d(i,j) - \delta(i,j)|}{d(i,j)}$$

4. DCM: Desviación Cuadrática Media $DCM = (1/P)\sum_{i<j\leq N} (d(i,j) - \delta(i,j))^2$

Donde $P = N(N-1)/2$, con N igual al número de UT's.

Los datos experimentales

Para probar experimentalmente la idea central de este trabajo, se utilizan datos de Verniere (1992). En su doctorado, este investigador del Centro Internacional de Investigación Agronómica para el Desarrollo -CIRAD/Francia- estudiaba la Ecología

y la Epidemiología de Xanthomonas Campestrís pv. Esta bacteria se encuentra en la mayor parte de zonas de producción de agrios en el mundo y provoca daños importantes. El objetivo de ese estudio fue investigar, entre las cepas aisladas en diversas regiones del mundo, posibles grupos homogéneos, estudiar su relación y eventualmente deducir un modo de dispersión de la enfermedad desde la zona de origen.

La clasificación actual del patotipo cítrí de Xanthomonas Campestrís pv, se basa sobre la naturaleza de las variedades atacadas y sobre el origen geográfico. Se distinguen las razas A originaria de Asia, la raza B encontrada en Argentina y Uruguay y la raza C propia de Brasil. Estudios recientes cuestionan esta clasificación y dejan bastante abierto el tema de su coformación.

Tratando de identificar grupos homogéneos de cepas y su relación entre sí se propuso hacer varias clasificaciones, utilizando entre otras a CAJ y NJTREE. Verniere disponía de 47 cepas de distintos orígenes geográficos y observó su respuesta frente a 56 antibióticos de 16 familias. En concreto, él midió el diámetro de inhibición. Sobre el centro de una caja de Petrí se depositó el antibiótico y se observó qué tanto se acercaba al mismo la cepa. Entre más sensible y más resistente, deberían acercarse menos. En principio, todas las cepas tenían la misma escala de variación y se las quería clasificar en función de su respuesta diferencial a los antibióticos, por lo que se utilizó la Distancia Euclidiana Clásica no reducida como índice para medir su semejanza

Para el procesamiento computacional se utilizó particularmente el paquete ABCD Guenoche (1993) escrito en BASIC.

Resultados (Figura 1, Tabla 1)

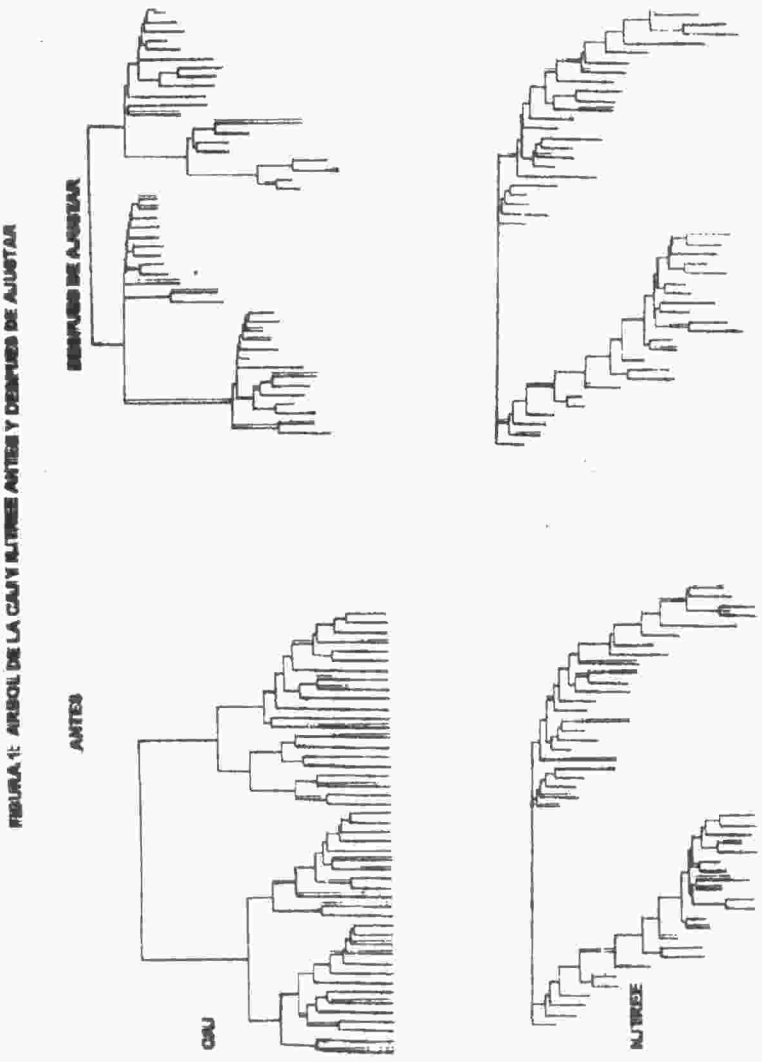

Tabla **1:** *Comparación de resultados del ajuste para las Distancias Ultramétrica y Aditiva según distintos criterios de evaluación y valor del índice de optimización*.*

DISTANCIA O ALGORITMO	ULTRAMÉTRICA CAJ			ADITIVA DE ARBOL NJTREE		
CRITERIO	SIN AJUSTE	CON AJUSTE	(SA-CA)/ SA*100	SIN AJUSTE	CON AJUSTE	(SA-CA)/ SA*100
1.DAM	13.27	6.17	53.5	3.87	3.20	17.3
2.DAMAX	62.94	28.30	55.0	16.10	16.86	-9.00
3.DARM	0.29	0.13	55.2	0.07	0.06	14.3
4.DCM	**362.38**	**65.08**	**82.0**	**24.07**	**16.58**	**31.00**

*: En las columnas 4 y 7 SA significa sin ajustar y CA con ajuste.

En general, mediante las clasificaciones hechas se encontró coherencia entre el origen geográfico de las cepas y su semejanza en la respuesta a los antibióticos. También hubo consistencia entre las clases obtenidas y el conocimiento proveniente de otras fuentes. Particularmente, las cepas provenientes de La Reunión, Rodrígues, Mauricio y Oman formaron un grupo compacto y estable.

Para nuestro objetivo estadístico, el análisis de resultados comprende 3 etapas:

1. Para cada algoritmo y a través de los indicadores o criterios propuestos, medir la distorsión entre la distancia sobre el árbol tal como lo producen los algoritmos normalmente -o sin ajustar- y la distancia sobre el árbol ajustado con base en la

optimización global, ambas con respecto al valor inicial que da el índice utilizado para medir la semejanza entre UT's.

2. Para cada criterio y algoritmo cuantificar en porcentaje la diferencia entre el valor con y sin ajuste con respecto al valor sin ajuste -columnas 4 y 7 de la tabla 1-.

3. Comparar los resultados sobre los dos algoritmos.

Discusión, conclusiones y recomendaciones

Con respecto al objetivo biológico. La teoría de la evolución plantea la adaptación al medio ambiente como uno de los mecanismos de respuesta de los seres vivos ante fenómenos migratorios por ejemplo. La asociación entre el origen geográfico y las cepas y su respuesta a los antibióticos -igual origen implica similar respuesta- es entonces consiste con esta teoría general. En términos prácticos implica que para combatir la enfermedad habrá que considerar el origen geográfico de la cepa como parte del tratamiento diferencial a aplicar.

Con respecto al objetivo estadístico. El árbol producido normalmente por NJTREE representa mejor la distancia inicial: el valor de todos los indicadores sin ajuste siempre es inferior a aquellos de la CAJ. En la distancia ultramétrica de la CAJ es sustancial la mejora en el ajuste sobre todos los indicadores. La mejora en el ajuste es inferior a la distancia aditiva de Arbol de NJTREE, esto debido a su menor distorsión inicial. Ajustado o no, este algoritmo produce un árbol que deforma menos la semejanza entre dos UT's .

En cuanto a la lectura del árbol, esta es más fácil en la CAJ y también más pertinente en un objetivo de clasificación. En el caso de NJTREE Chacón (1993)

muestra que este algoritmo es más adecuado en una óptica filogenética: una pareja de UT's para formar una clase debe estar próxima y además cada uno de sus elementos alejados de los demás.

La generalidad de los resultados de este trabajo -obtenidos aquí con unos datos particulares- habrá que buscarla con investigación teórica. Se abre así una ventana para explorar una idea que mejora la aproximación del mundo real mediante su abstracción por medio de un árbol.

Agradecimientos

El autor agradece sinceramente a las siguientes entidades y personas: Centro Frutícola Andino -Cali, Universidad del Valle, COLCIENCIAS/ICETEX, La Embajada de Francia en Colombia y CIRAD/Francia. Carlos Arana, Claudia María Pelaez y Enrique Abadía, del Centro Frutícola Andino; Jorge Cabra de BIOTEC / Universidad del Valle; X. Perrier, C. Dubois y C. Verniere del CIRAD / Francia. C. Vernierees un investigador del CIRAD con la siguiente dirección: AV. DU VAL DE MONTFERRAND BP 5035 34032 MONTPELLIER FRANCIA.

REFERENCIAS

Barthelemy J.-P., Guenoche A. (1988). *Les arbres et les représentations des proximités. Méthodes + programmes.* Ed. Masson. 240 p.

Benzecri J.P. (1973). *L 'Analyse des Données I La taxinomie.* Ed. Dunod. 615 p.

Chacn H., (1993). *Approximation D 'Arbres de diversite genetique,* Memoria de DEA de Bioestadística, Universidad de Montpellier II, 54 p.

Guenoche A. (1993). *Manuel d 'utilisation du logiciel d 'Analyses Booléennes et Combinatoires de Données,* ABCD, GRTC-CNRS, Marseille. 31 p

Higgins (1991); *Informe de Colombia a UNCED/92, 1991, citado en Nuevas Tecnologías para Recrear el Agro,* COLCIENCIAS.

Nei M. (1987). *Phylogenetic Tree in Molecular Evolutionary Genetics. Ed. Columbia University Press (New York).* 512 p.

Nei M, Saitou N. (1987). *The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Tree. Mol.* Biol. Evol. 4(4) : 406-425 p.

Perrier X. (1992). *Les mesures de la diversité génétique chez les bananiers. Symposium International sur l 'amélioration génétique.* Montpellier, Sep. 92 (à paraître) . 17 p.

Sneath P.H.A.,Sokal R.R (1973). *Numerical Taxonomy. The principles and practice of numerical classification.* Ed. W.H. Freeman and Company (San Francisco). 573 p.

Swofford D., Olsen G. (1990). *Phylogeny Reconstruction in Molecular Systematics.* Ed. Hillis D.M. and Moritz C. 588 p.

Verniere C. (1992). *Le chancre bactérien des agrumes (Xanthomonas campestris pv. citri).* Etude épidémiologique et écologique dans le cadre de l 'île de la Réunion, Thèse de Doctorat de l 'Université de Paris - sud. 175 p.