QÜESTIIÓ, Vol 16, 1,2,3 pp. 27-42, 1992

GEOMETRÍA ESTADÍSTICA EN LOS ESPACIOS DE DISTANCIA Y SECUENCIA: DOS APLICACIONES

ELADIO BARRIO, CELIA BUADES y ANDRÉS MOYA

La Geometría estadística es un método complementario a los desarrollados hasta el momento para la inferencia y evaluación de relaciones filogenéticas entre entidades emparentadas, y que permite decidir si la estructura filogenética obtenida tiene una configuración de árbol, de estrella o de red.

El objetivo de este trabajo consiste en poner de manifiesto que, si bien la geometría estadística puede ayudar a decidir entre grandes topologías, no puede decidir entre tipos específicos de topologías. Para ello utilizamos dos ejemplos: el estudio por análisis de restricción del DNA mitocondrial del grupo obscura *de* Drosophila *como ejemplo de topología en forma de árbol y la evidencia de un fenómeno de transferencia genética horizontal de la IPNS (Isopenicilina-N sintetasa) de bacterias del género* Streptomyces *a determinados hongos filamentosos como ejemplo de topología en forma de red.*

Statistical geometry in distance and sequence spaces: two applications.

Keywords: Statistical Geometry, phylogeny, molecular evolution, mitochondrial DNA restriction analysis, horizontal gene transfer.

-Departament de Genètica, Facultad de Biologia. Universitat de València. C/Dr. Moliner, 50. 46100 Burjassot, València.

-Article rebut el desembre de 1991.

-Acceptat el juny de 1992.

1. INTRODUCCIÓN

Hay una gran cantidad de métodos para la reconstrucción de la filogenia de un conjunto de organismos, genes o cualquier tipo de entidades relacionadas por ancestralidad, que dependen de un conjunto particular de asunciones evolutivas y/o matemáticas (hay una revisión reciente en Felsenstein, 1988). También es muy amplia la literatura estadística para la comparación de filogenias alternativas así como para la determinación de la significación de los puntos de ramificación o nodos en una filogenia (véase también Felsenstein, 1988). Un método reciente es el de la geometría estadística en los espacios de distancia y secuencia (Winkler-Oswatitsch *et al.*, 1986; Eigen *et al.*, 1988, 1989; véase también Maynard-Smith, 1989). El método suministra reglas sencillas para seleccionar la topología apropiada y para decidir si una topología filogenética, es decir una sucesión de puntos de ramificación, es propiamente una filogenia en forma de árbol, donde la ramificación es binaria, en forma de estrella, donde la ramificación es de orden superior a dos, o una filogenia en forma de red, donde pueden reunirse ramas que proceden de nodos ancestrales distintos. El método ha sido aplicado con cierto éxito a la evolución de los RNA de transferencia, a las moléculas de RNA ribosomal (Winkler-Oswatitsch *et al.*, 1986; Eigen *et al.*, 1989) y más recientemente a la evolución de viroides y virusoides (Elena *et al.*, 1991). No obstante, no conocemos hasta el momento otros ejemplos de aplicación, y específicamente de validación que no sean los basados en ejemplos estrictamente teóricos, y que estén orientados a la evaluación del método. En el presente trabajo pretendemos su aplicación a dos casos bien distintos. El método de espacio de distancia será aplicado a los resultados del análisis de restricción del DNA mitocondrial (mtDNA) de diversas especies de *Drosophila* del grupo *obscura* (Barrio *et al.*, 1992), en donde disponemos de una matriz de presencia/ausencia de sitios que caracterizan cada una de ellas. Para la aplicación del método de espacio de secuencias, disponemos de la secuencia del gen de la IPNS (también llamada ciclasa) y del RNA ribosómico 5S (5SrRNA) de bacterias del género *Streptomyces* y tres hongos filamentosos (Peñalva *et al.*, 1990). Con el primer ejemplo evaluamos la eficiencia del método de distancias para discriminar entre posibles topologías en forma de árbol. Con el segundo ejemplo, tratamos de poner de manifiesto cómo la existencia de una fenómeno de transferencia genética horizontal debe implicar una topología en forma de red.

2. GEOMETRÍA ESTADÍSTICA EN EL ESPACIO DE LAS DISTANCIAS

La base de datos de la que se parte es una matriz de caracteres discretos. A partir de dicha matriz se calcula la distancia genética entre pares de taxones. Posteriormente se efectuan todas las comparaciones posibles entre conjuntos de cuatro taxones (cuartetos), descomponiendo las seis distancias genéticas entre cada par de taxones que componen el cuarteto en cuatro ramas interiores, que conectan los cuatro nodos internos, y las cuatro ramas terminales (Figura 1). Dado que las ramas internas son iguales en longitud dos a dos, sólo hablamos de las distancias o longitudes internas de las ramas x e y (véase Winkler-Oswatitsch *et al.*, 1986, Figura 7 y págs. 62–63). Las seis distancias genéticas estimadas nos llevan a la resolución para cada cuarteto de las longitudes de las ramas que lo componen (x, y) y las cuatro ramas terminales (Figura 1). Para una topología dada, x es una medida de la desviación respecto de un verdadero árbol filogenético, e y corresponde a la auténtica longitud de la rama interior de cada cuarteto. Si x e y son cero, la topología es una estrella. Si x e y son del mismo orden, la topología tiene la forma de red. Y si x es mucho menor que y la topología es la del árbol. Valores positivos de x aparecen por intercambio genético o por mutaciones paralelas reversas. Según Eigen *et al.* (1988) razones $x/y = 0.5$ o mayores nos pueden hacer sospechar de que la auténtica topología sea de tipo árbol.

Nosotros hemos aplicado la geometría estadística en el espacio de distancias a un conjunto de 16 haplotipos de mtDNA, según los mapas de restricción correspondientes a 14 taxones del grupo *obscura* de *Drosophila*. A partir de la matriz de presencia/ausencia de sitios de restricción de los citados 16 haplotipos se estiman las distancias genéticas (Tabla 1) entre pares de taxones, siguiendo para ello el procedimiento de estima por máxima verosimilitud para sitios de restricción desarrollado por Nei (1987, pág. 104). La Figura 2 muestra un agrupamiento UPGMA de los 16 haplotipos derivados de la hemimatriz de distancias genéticas. Los recuadros sombreados que se situan sobre los nodos hacen referencia a los errores estandar de los mismos (Nei *et al.*, 1985). La correlación cofenética entre las matriz de distancias genéticas utilizadas para la construcción del agrupamiento por el método UPGMA (Tabla 1) y la de las distancias estimadas que se derivan como consecuencia de su utilización fue de 0.9, una bondad de ajuste muy buena. Como puede observarse hay dos agrupamientos monofiléticos de especies estrechamente emparentadas (el subgrupo Neártico de las *affinis*, haplotipos I–VI, y el subgrupo *pseudoobscura*, también neárticas, haplotipos XII–XVI), más un conjunto heterogéneo de especies más alejadas entre sí que corresponden a las especies paleárticas del grupo (subgrupo *obscura*, haplotipos VII–XI). La matriz de sitios de restricción fue también utilizada para

la construcción de un árbol, siguiendo el criterio de Wagner, que minimizase el número de mutaciones requeridas para conectar todos los haplotipos. Para ello utilizamos el programa MIX (paquete PHYLIP de inferencia filogenética, versión 3.4). El árbol más parsimónico encontrado, también mostró un agrupamiento entre los subgrupos *affinis* y *pseudoobscura*, así como la heterogeneidad del subgrupo *obscura*. Un remuestreo de 100 repeticiones por el método "bootstrap" (programa BOOT, paquete PHYLIP, versión 3.4) mostró que en el 100% de ellas, las especies del subgrupo *pseudoobscura* siempre aparecían juntas. No hay duda del carácter monofilético (es decir, origen único) de este subgrupo. En un 81% de las repeticiones, las especies del grupo *affinis* aparecían relacionadas. Sin embargo, el resto de nodos, que conectan las especies del subgrupo *obscura* entre si y con los dos subgrupos señalados, no alcanzaba el valor necesario (95% o mayor) para admitir que se trata de grupos monofiléticos.

La Tabla 2 muestra los resultados obtenidos con el método de la geometría estadística en el espacio de distancias. La media x/y ha sido calculada para subconjuntos de haplotipos, y los resultados aparecen en las diez primeras filas de la Tabla 2, ordenados de menor a mayor valor. Como puede observarse la razón x/y es inferior a 0.5 en 9 de los 10 subconjuntos de haplotipos. También se puede observar que la cuarta fila corresponde al promedio de razones x/y de todos los posibles cuartetos de haplotipos. El método de geometría estadística en el espacio de distancias se ha probado aleatorizando la matriz de presencia/ausencia de sitios de restricción 100 veces, repitiendo otras tantas veces el proceso de cálculo de distancias genéticas y parámetros de geometría estadística, obteniendo con ello los valores medios para x, y y la razón x/y. Estos resultados aparecen reflejados en la fila 11 de la Tabla 2. Se observará que la razón obtenida es la predicha por Eigen *et al.* (1988). Los resultados obtenidos los interpretamos de la siguiente manera

(1) El valor límite de 0.5 es apropiado para suponer que una determinada topología es de tipo árbol.

(2) Valores inferiores a 0.5 no pueden ser tomados, no obstante, como confirmación estadística de una configuración topológica de auténtica filogenia. De hecho se han obtenido razones x/y menores de 0.5 para agrupamientos filogenéticos incompatibles. Así, de acuerdo con el resultado del remuestreo "bootstrap", en ninguno de los nodos, excepto el que corresponde al subgrupo *pseudoobscura* (Figura 2), se garantiza que las agrupaciones que surgen de ellos sean monofiléticas. Cualquier reordenamiento de los nodos localizados en la Figura 2 con valores de distancia superior a la que corresponde al nodo del subgrupo *affinis* (incluyendo la coalescencia de todos ellos en una estrella) es compatible con los datos. Pero la geometría estadística en el espacio de distancias nos lleva a valores inferiores a 0.5

para todos los nodos mostrados en la Figura 2, así como para el caso de cualquier topología alternativa.

3. GEOMETRÍA ESTADÍSTICA EN EL ESPACIO DE SECUENCIAS

Los datos de partida en este caso son una matriz de secuencias de caracteres discretos, previamente alineadas. Como para el caso de las secuencias de ácidos nucleicos los estados posibles de cada carácter son cuatro, correspondientes a cada uno de los cuatro nucleótidos presentes en los mismos. Eigen y colaboradores transforman la matriz de datos cuaternarios en una de datos binarios (agrupando las bases según sean purinas o pirimidinas). Por lo tanto, en esencia, la matriz de partida es una matriz de datos binarios y, en consecuencia, los datos analizados en el apartado anterior son susceptibles de ser analizados mediante este método, y viceversa. Al igual que en el método anterior, éste se basa en el cálculo de parámetros de todos los cuartetos posibles para los taxones o entidades estudiadas. La diferencia está en que el número de parámetros no es 6, sino 8 (Figura 3). Se trata de las longitudes de las ramas que conectan los nodos internos con los taxones, más tres parámetros de longitudes internas. De hecho, los primeros son el número de caracteres que identifican a cada uno de los taxones (el carácter está presente o ausente en solo uno de los cuatro taxones). Los otros tres parámetros corresponden al número de caracteres con dos secuencias en un estado y dos en otro. Al ser tres las posibilidades de combinación, son tres los parámetros, y se definen como l, m y s, por asignación al conjunto con mayor, medio y menor número de caracteres compartidos, respectivamente. Los caracteres presentes o ausentes en los cuatro taxones a la vez son considerados como no informativos y se excluyen, por tanto, del análisis. La Tabla 3 muestra las relaciones esperadas entre parámetros en el espacio de secuencias bajo tres situaciones: ideal, aleatoria y real; y definiendo las tres configuraciones topológicas de las que venimos hablando: árbol, estrella y red. Los resultados de secuencias reales se comparan con los de las mismas secuencias aleatorizadas. Para ello, y como ocurría en el caso de la geometría en el espacio de distancias, se procede a una aleatorización, en este caso horizontal, y al recálculo de los parámetros de geometría en el espacio de secuencias.

Para ejemplificar y criticar el método en cuestión, vamos a aplicarlo al caso relativamente bien documentado de transferencia génica horizontal de un gen, el gen de la IPNS, implicado en la síntesis de antibióticos betalactámicos, presente tanto en bacterias del género *Streptomyces,* como en antecesores de determi-

nado grupo de hongos filamentosos. Hemos utilizado las secuencias del gen de la IPNS de 7 organismos (tres hongos filamentosos: *Acremonium chrysogenum, Penicillium chrysogenum,* y *Aspergillus nidulans,* tres bacterias del género *Streptomyces: S. clavuligerus, S. lipmanii* y *S. jumonjinensis,* así como una bacteria gram negativa del género *Flavobacterium*) y la secuencia del RNA ribosómico 5S de cinco organismos (los tres hongos filamentosos, un *Streptomyces* y una bacteria gram negativa). Para adquirir una mayor información sobre el tema, el lector puede acudir al artículo de Peñalva y col. (1990) y las referencias allí citadas. Hay que señalar que todas las secuencias analizadas fueron obtenidas de la base de datos del EMBL en Heildelberg, Alemania. La Tabla 4 muestra los resultados obtenidos tanto de los genes de la IPNS en un conjunto de especies, así como los relativos a los genes 5SrRNA, muy utilizados en la literatura como reloj molecular y de los que, por supuesto, se sabe que no han sufrido transferencia horizontal alguna. Es evidente que las relaciones entre los parámetros definen más nítidamente una topología en forma de árbol para el caso de los genes 5SrRNA que para el de los de la IPNS (véase relaciones esperadas en la Tabla 3). Pero, al igual que ocurría con la geometría estadística en el espacio de la distancias, la geometría estadística en el espacio de secuencias solo sugiere la posibilidad, pero no discrimina nítidamente. En otras palabras, la forma de árbol del 5SrRNA está más cerca de las relaciones ideales para dicha topología que el caso de la topología obtenida con los genes de la IPNS. Sin embargo, ello no nos permite concluir que el segundo caso no corresponda a una topología de árbol como lo es el primero. Está claro que cuando se produjo el suceso de transferencia, la topología correspondiente a la evolución del gen de la IPNS era en forma de red, y en forma de árbol para la de los genes 5SrRNA. Pero desde entonces se ha dado un proceso de evolución gradual y acumulación de mutaciones que nos ha puesto frente a una nueva topología que no es exactamente una red, pero tampoco exactamente un árbol. Tal incertidumbre no la puede resolver la geometría estadística en el espacio de las secuencias, y por ende, hemos de recurrir a otros procedimientos.

Agradecimientos

Agradecemos al SERBIO (Servicio de Bioinformática de la Universitat de Valencia) los servicios de computación. El trabajo ha sido financiado por los proyectos BIO89-0668-C03-03 y PB90-0491 de la CICYT y DGICYT, respectivamente, así como por una beca de la Conselleria de Cultura, Educació i Ciència de la Generalitat valenciana otorgada a E. B.

BIBLIOGRAFÍA

[1] **Barrio, E.; Latorre, A.; Moya, A.** and **Ayala, F.J.** (1992). "Phylogenetic reconstruction of the *Drosophyla obscura* group on the basis of mitochondrial DNA." *Mol. Biol. Evol.*, **9**, 621–635.

[2] **Eigen, M.; Lindemann, B.F.: Tietze, M.; Winkler-Oswatitsch, R.; Dress, A.** and **von Haeseler, A.** (1989). "How old is the genetic code? Statistical geometry of tRNA provides an answer". *Science,* **244**, 673–679.

[3] **Eigen, M.; Winkler-Oswatitsch, R.** and **Dress, A.** (1988). "Statistical geometry in sequence space: a method of quantitative sequence analysis". *Proc. Natl. Acad. Sci. USA,* **85**, 5913–5917.

[4] **Elena, S.F.; Dopazo, J.; Flores; R.; Diener, T.O.** and **Moya, A.** (1991). "Phylogeny of viroids, viroidlike satellite RNAs, and the viroidlike domain of hepatitis delta virus". *Proc. Natl. Acad. Sci. USA,* **88**, 5631–5634.

[5] **Felsenstein, J.** (1988). "Phylogenies from molecular sequences: inferences and reliability". *Annual Review of Genetics,* **22**, 521–565.

[6] **López-Bueno, J.A.** and **Moya, A.** (1992). "GEOSEQ: A Pascal program to calculate statistical geometry parameters of aligned nucleic acid sequences". *Computer Applications in the Biosciences,* en prensa.

[7] **Maynard-Smith, J.** (1989). "Trees, bundles or nets?". *Trends in Ecology and Evolution,* **4**, 302–304.

[8] **Nei, M.** (1987). *Molecular Evolutionary Genetics.* Columbia University Press, New York.

[9] **Nei, M.; Stephens, J.C.** and **Saitou, N.** (1985). "Methods for computing the standar errors of branching points in a evolutionary tree and their application to molecular data from human and apes". *Molecular Biology and Evolution,* **2**, 66–85.

[10] **Peñalva, M.A.; Moya, A.; Dopazo, J.** and **Ramón, D.** (1990). "Sequences of isopenicillin N synthetase genes suggest horizontal gene transfer from prokaryotes to eukaryotes". *Proc. Roy. Soc. Lond.* **B 241**, 164–169.

[11] **Winkler-Oswatitsch, R.; Dress, A.** and **Eigen, M.** (1986). "Comparative sequence analysis". *Chemica Scripta,* **26B**, 59–66.

Tabla 1

Hemimatriz de divergencia nucleotídica entre los haplotipos mitocondriales determinados en las especies de *Drosophila* del grupo *obscura* (para más información sobre la correspondencia entre haplotipo y especie, véase Figura 2). Las estimas de distancia genética se calculan a partir de los datos de presencia/ausencia de los sitios de restricción de trece endonucleasas en el DNA mitocondrial de estas especies (Barrio *et al.*, 1992), y utilizando el procedimiento de máxima verosimilitud desarrollado por Nei (1987, pág. 104).

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI
I	–															
II	0.013	–														
III	0.010	0.003	–													
IV	0.023	0.021	0.018	–												
V	0.027	0.024	0.021	0.036	–											
VI	0.051	0.041	0.038	0.047	0.039	–										
VII	0.106	0.103	0.100	0.080	0.107	0.097	–									
VIII	0.109	0.095	0.092	0.092	0.110	0.090	0.086	–								
IX	0.085	0.082	0.080	0.088	0.084	0.077	0.093	0.096	–							
X	0.114	0.111	0.108	0.107	0.115	0.106	0.114	0.107	0.077	–						
XI	0.094	0.100	0.097	0.097	0.093	0.095	0.124	0.095	0.100	0.119	–					
XII	0.104	0.100	0.098	0.097	0.115	0.126	0.126	0.106	0.111	0.141	0.091	–				
XIII	0.108	0.105	0.102	0.101	0.121	0.133	0.132	0.111	0.117	0.148	0.104	0.012	–			
XIV	0.107	0.104	0.101	0.100	0.118	0.130	0.129	0.109	0.115	0.144	0.094	0.008	0.015	–		
XV	0.097	0.094	0.091	0.091	0.108	0.120	0.118	0.099	0.104	0.135	0.085	0.006	0.012	0.009	–	
XVI	0.098	0.094	0.091	0.082	0.097	0.110	0.108	0.099	0.094	0.116	0.094	0.028	0.037	0.032	0.022	–

Tabla 2

Parámetros obtenidos mediante el método de la geometría estadística en el espacio de las distancias para diferentes agrupamientos de especies. El cociente x/y determina la desviación de la topología de árbol (para más detalles ver el texto). Abreviaturas:
aff = *affinis*, amb = *ambigua*, bif = *bifasciata*, gua = *guanche*, obs = *obscura*, pse = *pseudoobscura* y sub = *subobscura*.

	Agrupamientos	x	y	x/y
1.	Subgrupo *pseudoobscura* frente al resto	0.0023	0.0612	0.0377
2.	Par *ambigua-obscura* frente al resto	0.0033	0.0302	0.1103
3.	Subgrupo *affinis* frente al resto	0.0052	0.0450	0.1163
4.	Todas las especies estudiadas	0.0042	0.0336	0.1237
5.	Cuarteto sub-gua-amb-obs frente al resto	0.0052	0.0278	0.1877
6.	Subgrupo *obscura* frente al resto	0.0061	0.0274	0.2226
7.	Par *subobscura-guanche* frente al resto	0.0069	0.0243	0.2842
8.	Trío bif-obs-amb frente al resto	0.0078	0.0231	0.3379
9.	Cuarteto amb-obs-gua-sub frente a bif frente subgrupo pse frente subgrupo bif	0.0073	0.0171	0.4282
10.	Par amb-obs frente par sub-gua frente subgrupo aff frente subgrupo pse con bif	0.0097	0.0041	2.3742
11.	Agrupamientos al azar desde	0.0144	0.0079	0.5541
	a	0.0406	0.0261	6.7626

Tabla 3

Valores de referencia esperados para los tres grandes tipos de topologías según los parámetros medios de geometría estadística en el espacio de secuencias.

Situación	Estrella	Árbol	Red
Ideal	$l = m = s = 0$	$m = s = 0, l > 0$	$m = (a+b+c+d)/4$
Aleatorio	$l = m = s$	$m, s \geq (a+b+c+d)/4$	———
Real	$l \approx m \approx s$	$l \gg m, s$	$m \approx (a+b+c+d)/4$

Tabla 4

Valores medios de geometría estadística en el espacio de secuencias obtenidos en el estudio de dos genes: el de la IPNS (7 secuencias alineadas de 1024 nucleótidos, 35 cuartetos posibles) y el del RNA ribosómico 5S (5 secuencias alineadas de 123 nucleótidos, 5 cuartetos posibles). El procedimiento de aleatorización de las secuencias se siguió utilizando el algoritmo descrito por López-Bueno y Moya (1992). Para más información sobre las especies de procedencia de las secuencias consúltese Peñalva *et al.* (1990).

	Secuencias			
	IPNS		5SRNA	
Parámetros	Reales	Aleatorizadas	Reales	Aleatorizadas
a	82.54	131.45	16.20	16.04
b	54.23	129.44	8.20	15.80
c	42.83	126.26	4.00	15.36
d	70.23	127.06	4.40	16.48
l	76.14	139.35	17.20	18.24
m	23.71	128.62	2.60	14.52
s	17.63	118.45	0.60	11.64
$2l-(m+s)$	110.94	31.62	31.20	10.32

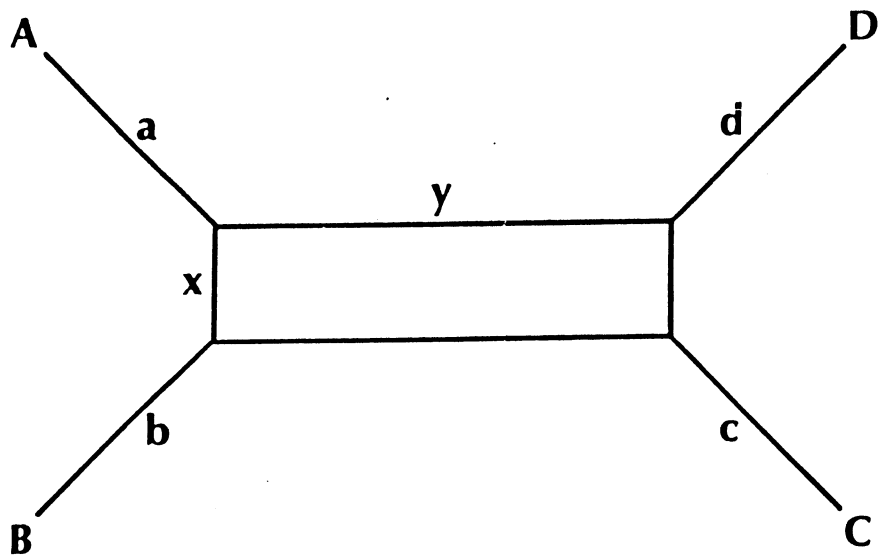

Figura 1.
Parámetros medios de geometría estadística en el espacio de las distancias. Véase texto para más información.

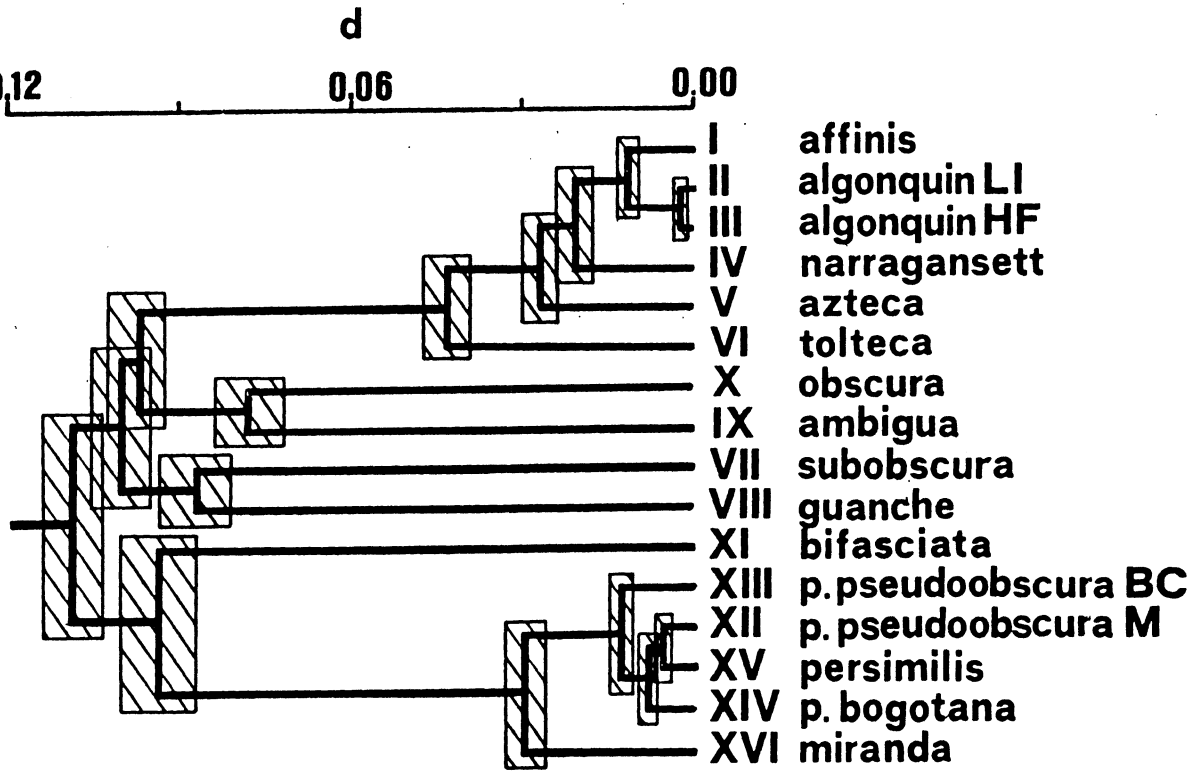

Figura 2.

Agrupamiento siguiendo el procedimiento UPGMA de las distancias entre haplotipos mitocondriales de las especies del grupo *obscura* de *Drosopila*.

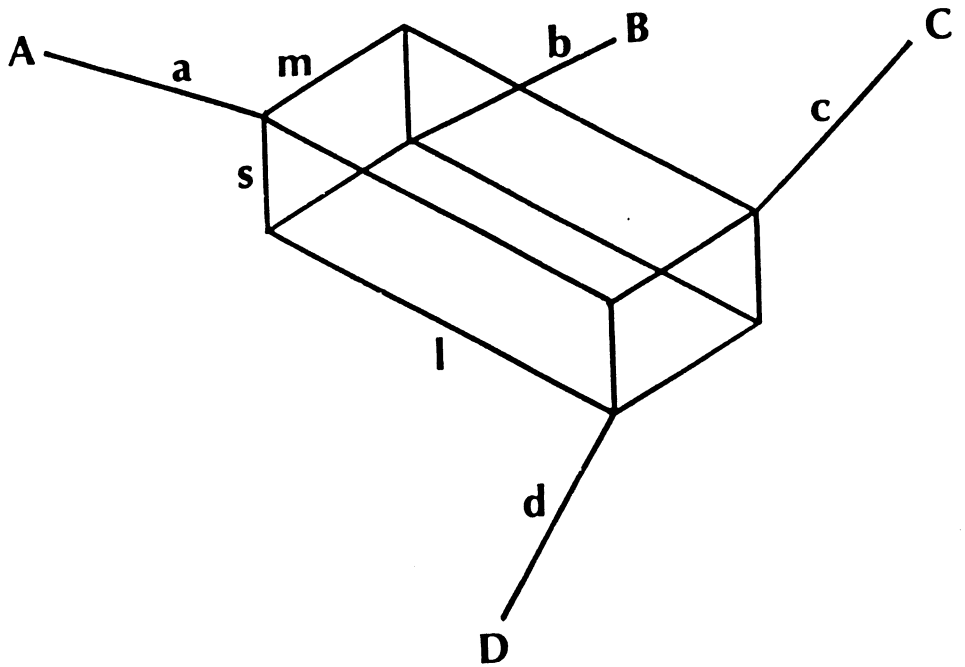

Figura 3.
Parámetros medios de geometría estadística en el espacio de las secuencias. Véase texto para más información.

ENGLISH SUMMARY:

STATISTICAL GEOMETRY IN DISTANCE AND SEQUENCE SPACES: TWO APPLICATIONS

Eladio Barrio, Celia Buades y Andrés Moya

1. INTRODUCTION

Statistical geometry is a new approach that complements existing methods for comparing alternative phylogenies and for assesing the significance of the topological location of the branching points in a phylogeny, allowing to select the appropiate topology; that is to say, a tree-like tipology, a 'bundle' (i.e., an effectively simultaneous split of all branches) or a 'net' (i.e., such as might arise by hibridization or lateral genetic transfer).

In the present paper we show that, although statistical geometry helps to choose between phylogenetic topologies, it cannot decide between specific topological types.

2. STATISTICAL GEOMETRY IN DISTANCE SPACE

Two methods are presented. The first one is the statistical geometry in distance space. This method calculates two internal parameters of distance, x and y, that are mean values obtained from all possible quartets of species (i.e., Operative taxonomical units, OTUs). The data available are genetic distances between species, and for each quartet there are six distances that can be partitioned in six unknown partial distances: x, y as internal distances, and four more corresponding to the external branches. The x/y ratio can be used to decide between the different topological types mentioned above. Data of genetic distances obtained from restriction site analysis of the mitochondrial DNA of the *Obscura* group of *Drosophila* are used to verify the efficiency of the statistical method in distance space. In fact x/y ratios less than 0.5 indicate tree-like topologies. The method tell us that the phylogeny for the total set of species is a tree-like topology, but when try to decide between specific sets, the ratios are mostly less

than 0.5 and some of such sets are clearly improbable from a phylogenetic point of view.

3. STATISTICAL GEOMETRY IN SEQUENCE SPACE

In this case the data set is a matrix of aligned nucleic acid sequences. The matrix is directly used for the method of statistical geometry in sequence space, but only two states are allowed for each character. If more than two states occur, they should be reduced to only two. A network connecting each possible quartet of binary sequences is generated and each position in a given quartet is assigned to one of the following eight parameters: a, b, c, d, l, m, s and x. $a, b, c,$ and d are the number of character for which the four species are odd man out, while l, m and s (for large, medium and small) are the numbers of characters with two sequences in one state and two in the other. x is the number of uninformative characters. As the previous one, we have used the method in order to choose between a tree, a bundle or a net.

To verify the efficiency of the method, data sequence of two different genes have been used: the 5SrRNA gene, a typical molecular clock used in molecular evolutionary studies to construct phylogenetic trees, and the IPNS gene (i.e., isopenicillin N synthetase), that has probably been horizontally transfered from *Streptomyces* to an ancient fungi. The relationship between internal mean values corresponding to quartets of 5SrRNA genes are closer than the IPNS genes to a tree topology. Horizontal gene transfer has previosly been put into evidence in the case of IPNS, but the statistical geometry in sequence space is not able to say why the 5SrRNA is a tree topology and IPNS a net topology. In other words, there is no statistical procedure to decide between both topological types.