ANALISIS ESTADISTICO EN BIOLOGIA MOLECULAR: USOY APLICACION EN POBLACIONES VEGETALES

Orlando Martínez Wilches[1]

RESUMEN

Cada vez que los investigadores de las ciencias básicas biológicas producen, innovan y proponen métodos y técnicas que describen, cualquiera que sea, la variabilidad de las poblaciones naturales y experimentales, es necesario analizar, revisar y evaluar los procedimientos estadísticos disponibles que se adecúan a tales circunstancias; ó bien si se requiere, desarrollar técnicas bioestadísticas alternas para el análisis e interpretación de los resultados provenientes de experimentos, que involucran los nuevos métodos biológicos.

En el caso de la Agronomía, la biología molecular y disciplinas afines han presentado reciéntemente los métodos de isoenzimas, RFLPS y los RAPDS, para determinar la variabilidad, composición y estructura genética de individuos, poblaciones naturales y experimentales. Se analiza y discute el uso de las distancias genéticas, índices de similitud, dendogramas y escalas multidimensionales como técnicas estadísticas para experimentos agronómicos que usan isoenzimas, RFLPS y RAPDS como marcadores genéticos.

---

1 Profesor titular. Facultad de Agronomía. Universidad Nacional de Colombia, Bogotá.

STATISTICAL ANALYSIS IN MOLECULAR BIOLOGY: USE AND APPLICATIONS IN AGRONOMIC POPULATIONS

SUMMARY

When the researches in basic biological sciences propose, produce or introduce methods and techniques that describe the variability of natural or experimental populations, it is necessary to review, analyze and evaluate the statistical procedures available and adequate for these circumstances. Sometimes it is required to develop statistical techniques for the analysis and interpretation of the data originated in experiments involving biological innovations.

In agronomic research, the molecular biology and similar disciplines have proposed the isoenzymes, RFLP'S and the RAPDS to evaluate the variability, composition and genetic structure of natural and domesticated populations.

In this review, it is discussed and described the use of genetic distances, coefficients of similarity, dendograms and multidimensional scaling as statistical techniques in agronomic experiments which use isoenzymes RFLP'S and RAPDS as genetic markers.

INTRODUCCION

McCalla (1994) señala cuatro grandes tendencias de la agricultura en los últimos años así: La interdependencia global e integral de los países por el mercado de bienes y servicios; el desarrollo acelerado de las comunicaciones y la información tecnológica en la agricultura tanto a nivel de productor como en las negociaciones de las multinacionales; el consenso mundial y la preocupación de los países por la ecología y el medio-ambiente donde los recursos naturales disponibles ya son finitos; finalmente, la revolución de la biología molecular y su acelerado desarrollo en los últimos 20-30 años. Esta disciplina y otras afines a ella, han ampliado el conocimiento de la genética, evolución y el funcionamiento de los organismos biológicos. Inicialmente se preveía, que mediantes estas técnicas biotecnológicas se obtendría una rápida transformación de la agricultura. Sin embargo, tales observaciones estaban sobre estimadas y se considera que nos encontramos en los primeros estados del impacto y aplicación que estas tecnologías puedan causar en el desarrollo y la productividad agrícola de los países. Los próximos años se preveé serán promisorios y éxitosos.

TECNICAS ESTADISTICAS MOLECULARES:

Los métodos y procedimientos estadísticos disponibles para el análisis de los resultados provenientes de ensayos biotecnológicos, se pueden agrupar en las siguientes categorías:

1. Aquellos que tienen como propósito el de evaluar la variabilidad, clasificación, estructura y composición genética de las poblaciones,

2. Los desarrollados para la construcción de mapas cromosómicos ó genómicos, cuando se utilizan marcadores genéticos moleculares, y

3. Lo denominados QLT (Quantitative trait loci), que son loci asociados con caracteres cuantitativos de importancia económica, como el rendimiento, y que proveen al fitomejorador de una herramienta molecular ágil, precisa y oportuna de selección indirecta por los caracteres cuantitativos de interés envueltos en el programa de fitomejoramiento.

Este escrito solo se ocupa de los primeros, es decir, de aquellos que en general describen la variabilidad genética de las poblaciones. En particular, se enfatiza su uso en poblaciones, que convencionalmente se reconocen como "recursos genéticos naturales", las cuales son indispensables, como su nombre lo indica, para el desarrollo y progreso futuro de la agricultura.

MARCADORES GENETICOS

Son numerosas las variables, los caracteres ó parámetros, que se han utilizado para observar y detectar la variabilidad presente en los seres vivos. Los marcadores genéticos son una clase de estos y con ellos se espera que reflejen la variabilidad debida principalmente a los genes.

Los marcadores morfológicos - cuantitativos se consideran como el resultado de los efectos combinados de muchos genes y el ambiente p.e. altura de planta, número de pétalos, longitud de la mazorca. Para su evaluación se requiere de una medida, conteo ó calificación.

Los marcadores bioquímicos, están constituidos por las isoenzimas y las proteinas. Mediante la técnica de la electroforesis en gel, se hace posible el estudio de la variación de las proteinas y enzimas en organismos vivos así: Las muestras de tejidos se homogenizan (muelen) para liberar las enzimas y proteinas de las células. El sobrenadante del homogenizado (parte liquida), se coloca en un gel de almidón, agar, poliacrilamida ó alguna sustancia gelatinosa. El gel se somete durante horas a corriente eléctrica continua, cada proteína del gel migra en una dirección y velocidad que depende de la carga eléctrica neta de la proteína y del tamaño molecular. Después el gel se trata con una solución química con un sustrato específico para la enzima en estudio y una sal que produce una mancha (banda) coloreada, que refleja la migración de la enzima. La utilidad del método radica en el hecho de que el genotipo del locus genético que codifica la enzima puede ser inferido a partir del número y posiciones de las bandas observadas en los geles (Ayala y Kiger 1984).

Los marcadores moleculares de mayor uso en la detección de la variabilidad genética, lo constituyen los RFLPS y los RAPDS. Los RFLPS, son una clase de enzimas, llamadas enzimas de restricción. Son nucleasas producidas por diferentes microorganismos y tienen la capacidad de reconocer ciertos sitios (sitios de restricción) constituidos por secuencias de bases específicas en el ADN. Si una secuencia específica de bases está presente en el sitio de restricción, la enzima de restricción corta al ADN en ese sitio. Por lo tanto, una cadena larga de ADN se puede reducir a una serie de fragmentos de tamaño finito según el corte de la enzima de restricción. El número de fragmentos producidos y

el tamaño de cada fragmento refleja los sitios de restricción en la cadena del DNA. Los fragmentos de restricción producidos por el corte de la endonucleasa (p.e. Hind III) de un tejido, se someten al proceso de electroforesis en agar; los fragmentos migran con la presencia de la corriente eléctrica y la velocidad de migración depende del peso molecular de cada fragmento. Posteriormente, el gel se colorea con bromuro de etidio y el patrón de migración de los fragmentos se observa directamente mediante manchas coloreadas de una manera similar a las isoenzimas y proteinas (Kochet 1994).

Los marcadores moleculares conocidos como RAPDS ó AP-PCR tienen como base la reacción en cadena de la polimerasa (una enzima, que bajo ciertas circunstancias produce replicas de cadenas sencillas de ADN). Los RAPDS (segmentos, amplificados, aleatorios de ADN) es una técnica para estudiar la variabilidad genética, la cual permite la detección de secuencias polimórficas de ADN, utilizando cebadores (Primers) sencillos con secuencias arbitrarias de oligonucleótidos. Las secuencias se amplifican o se generan con la información ADN del tejido de la especie en estudio y mediante la reacción en cadena de la polimerasa. Al igual que las isoenzimas, el material procesado se somete a electroforesis en agar y los segmentos amplificados migran por la acción de la corriente eléctrica, la velocidad de migración depende de su peso molecular. Después el gel se colorea con bromuro de etidio y el patrón de la migración de los segmentos de ADN, se observa directamente mediante manchas coloreadas (Williams et al 1990, Welsh and Mcclelland 1991).

CUANTIFICACION DE LOS MARCADORES BIOQUIMICOS Y MOLECULARES

Los resultados experimentales de un ensayo biológico donde se utilicen las proteinas, enzimas, RFLPS o RAPDS es el mismo: un conjunto de bandas coloreadas en el gel que representan el comportamiento de la variabilidad. Como ilustración, considere 5 colecciones de una especie agrícola, p.e. cacao, las que se sometieron a un estudio de diversidad enzimática. En la Figura 1 se presentan los resultados correspondientes a una corrida de la α-β esterasa; en la figura se observa el patrón (las bandas) de variación de las colecciones, la última columna corresponde al estandard el cual expresa todas las bandas posibles producidas por las cinco colecciones. El problema es cómo

cuantificar las bandas y una vez cuantificadas proponer medidas estadísticas que expresen la variabilidad entre las colectas en estudio.

Las bandas de la Figura 1 se pueden cuantificar mediante una función indicadora, esto es, asignar el valor 1 si la banda está presente y cero si no lo está. Al aplicar dicha función al ejemplo de la α-β esterasa, se obtiene la Tabla 1; ella refleja la variabilidad de las bandas pero ya de una forma cuantitativa, numérica, a la cual se le pueden proponer medidas estadísticas que expresen la diversidad enzimática entre las colectas en estudio.

FIGURA 1. PATRON DE VARIABILIDAD DE CINCO COLECCIONES DE CACAO ASOCIADOS CON LA α-β ESTERASA

	COLECCIONES					
ORDEN	**A**	**B**	**C**	**D**	**E**	**ESTAND.**
1		-		-		-
2	-		-	-		-
3	-	-	-	-		-
4		-				-
5	-	-	-		-	-
6	-	-	-		-	-
7		-			-	-
8	-		-	-		-
9	-		-	-		-
10			-	-		-

TABLA 1. CUANTIFICACION DE LA α-β ESTERASA EN CINCO COLECCIONES DE CACAO.

	COLECCIONES					
ORDEN	A	B	C	D	E	ESTAND.
1	0	1	0	1	0	1
2	1	0	1	1	0	1
3	1	1	1	1	0	1
4	0	1	0	0	0	1
5	1	1	1	0	0	1
6	1	1	1	0	1	1
7	0	1	0	0	1	1
8	1	0	1	1	1	1
9	1	0	1	1	0	1
10	0	0	1	1	0	1

INDICES O COEFICIENTES DE SIMILITUD:

Una medida de semejanza para comparar dos colecciones (la A y la B), utilizando los resultados de la Tabla 1, sería aquella que relacionara el número de bandas (unos o ceros) que simultáneamente compartan las dos colectas. La siguiente tabla provee de la información necesaria para relacionar las ausencias y presencias comunes entre el par de colecciones.

		B: 1	B: 0
A	1	a	b
	0	c	d

n = a+b+c+d

Dos medidas de semejanza (S_{AB}) entre A, B serían:

$$S_{AB} = a/n$$
$$S_{AB} = (a+d)/n$$

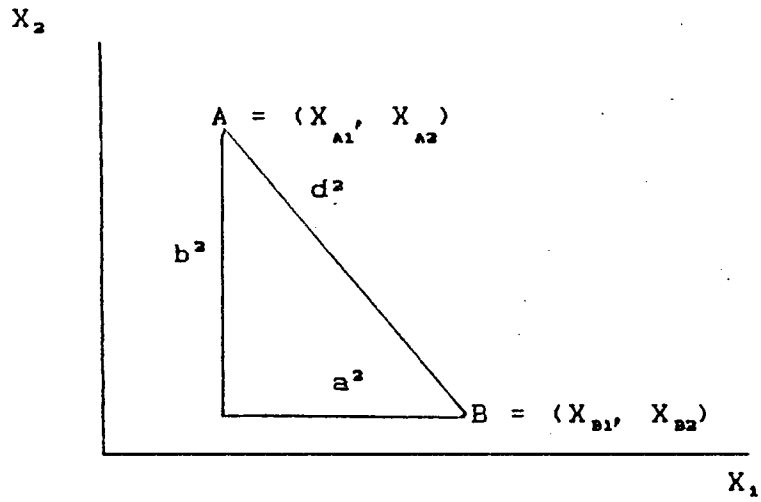
se puede entonces calcular una tabla (matriz) de coeficientes de similitud entre todas las colecciones.

Adicionales a las anteriores, se han propuesto diferentes índices de similitud. En la tabla 2 se expresan los más comunes, su interpretación y el autor. Estos índices fueron originalmente creados para estudios de poblaciones de insectos, ecología y en la especie humana donde la presencia y ausencia de características es común al evaluar el comportamiento ante una serie de estímulos cognocitivos.

TABLA 2: Coeficientes de Similitud

	COEFICIENTE	INTERPRETACION	AUTOR
1.	$\frac{a+d}{n}$	Igual peso a 0-0 y 1-1	Sokal, Michener 1958
2.	$\frac{a}{a+b+c}$	No contabiliza 0-0	Jackard, 1908
3.	$\frac{2a}{2a+b+c}$	Doble peso a 1-1 no contabiliza 0-0	Dice, 1945
4.	$\frac{2a}{b+c}$		Nei, 1987
5.	$\frac{a+d}{a+d+2(b+c)}$	Doble peso a 0-1. 1-0	Rogers y Tamimoto, 1960
6.	$\frac{a}{n}$	No. 0-0 en numerador	
7.	$\frac{a}{a+2(b+c)}$	No. 0-0 en numerador doble peso a 100 y 0-1	
8.	$\frac{a}{b+c}$	0-0 excluído, relaciona presencia con ausencia	

DISTANCIAS: Euclideana - Geométrica - Genética. La distancia Euclideana entre dos colectas es la aplicación del teorema de Pitagoras así:



$$D^2_{AB} = d^2 = a^2 + b^2$$

$$D^2_{AB} = (X_{A1} - X_{B1})^2 + (X_{A2} - X_{B2})^2$$

$$D^2_{AB} = \Sigma_k (X_{Ak} - X_{Bk})^2$$

La distancia euclideana es intuitivamente atrayente, fácil de entender, es una medida geométrica que posee numerosas características algebráicas - matemáticas, de allí su amplio uso en investigaciones en las ciencias biológicas, económicas y sociales.

La distancia genética es una medida que expresa la divergencia, entre dos poblaciones, razas ó colectas, divergencia atribuible exclusivamente a genes ó a conjuntos de los mismos. Si pi es la frecuencia del i-ésimo gene de la población A y qi lo es para la población B entonces una medida de distancia genética entre A y B es la distancia euclideana aplicada a la frecuencia de los genes así:

$$D^2_{AB} = \Sigma_i (p_i - q_i)^2$$

Se han propuesto diferentes medidas de distancia genética como son la de Rogers, Prevosti, Cavalli-sforza, Nei etc.; para su construcción se ha considerado aspectos geométricos, matemáticos y biológicos entre otros (Nei 1987).

DISTANCIAS Y SIMILARIDADES. Los índices ó coeficientes de similaridad, son medidas de semejanza entre bandas electroforéticas; algunos de ellos están relacionados con las distancias, fmediante funciones algebraicas. Es decir, bajo ciertas circunstancias es posible calcular distancias euclideanas a partir de los índices de similitud. Entre las expresiones que relacionan los coeficientes y las distancias se encuentran.

$$D^2_{\ i} = 1 - 2S_i$$

$$D^2_i = 2 (1 - S_i)$$

$$D^2_i = 1 \div (1 - S_i)$$

Sin embargo, tal como lo muestra Gower (1966, 1967) no siempre es posible calcular distancias euclideanas a partir de similaridades. La matriz de similaridad tiene que ser definida semipositiva para lograr la conversión. De las similaridades expresadas en la Tabla 2 solamente las definidas por Sokal

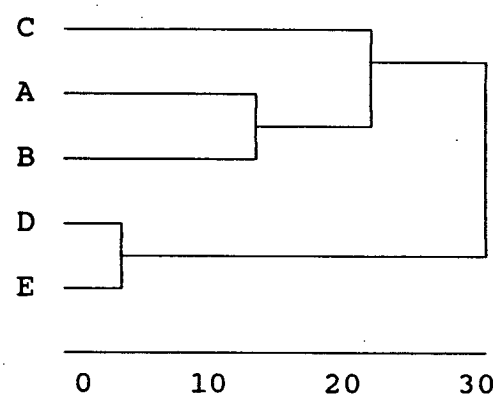
y Michener (1958) y Jacckard (1908) poseen esta condición. Gower (1966) también enfatiza en usar la expresión $2(1 - S_i)$.

Hasta ahora, para cuantificar el patrón de bandas electroforéticas se han propuesto diferentes coeficientes de similaridad, distancias genéticas, geométricas y euclideanas. Sin embargo, cuando se estudian varias poblaciones, p.e. 20 que es un tamaño más bien intermedio, el número total de distancias entre pares de poblaciones sería de (20 x 19)/2 = 190. Por lo tanto, se torna dispendioso resumir estas 190 distancias y a partir de ellas realizar las inducciones y deducciones poblacionales. El siguiente paso, es entonces el manejo de una matriz de distancias y con ella hacer inferencias estadísticas. Se discuten los dendogramas y las coordenadas principales, que son dos métodos gráficos - estadísticos que proveen de buenas guías al investigador que usa marcadores moleculares y bioquímicos en la descripción de la variabilidad de poblaciones biológicas. Los dendogramas y las escalas multidimensionales resumen la matriz de distancias.

DENDOGRAMAS - CONGLOMERADOS: El propósito fundamental del análisis de conglomerados es el de proveer al investigador de "agrupaciones naturales" de un conjunto de individuos, razas, ó variedades. Se busca colocar conjuntos de individuos en grupos exhaustivos y mutuamente excluyentes, de tal forma que se puedan hacer inferencias estadísticas de semejanzas ó diferencias en y entre los grupos provistos por el análisis.

Los grupos establecidos por el análisis forman particiones y subdivisiones en conjuntos menores ó reagrupamientos en mayores y eventualmente se puede finalizar con una estructura jerárquica de agrupamiento. A esta estructura se le conoce como jerarquización en árbol.

La estructura jerárquica de agrupamiento ó la estructura en árbol, se puede representar en un diagrama ó figura bidimensional y a tal representación se conoce como dendograma así p.e.



Los dendogramas, en general, se construyen a partir de una matriz p x p de distancias ó de coeficientes de similitud. Entonces las p(p-1)/2 posibles distancias ó similitudes obtenidas de p poblaciones se condensan en el dendograma, lo que facilita y simplifica enormemente las inferencias de semejanza ó disimilitud entre los diferentes grupos y subgrupos de poblaciones en estudio.

Por lo tanto, al patrón de bandas electroforéticas provenientes de las isoenzimas, los RFLPS ó los RAPDS, se les ha provisto de métodos estadísticos formales (distancias, similitudes, dendogramas) de tal manera que su variabilidad, bioquímica, molecular y en general genética se puede discriminar y cuantificar.

ESCALAS MULTIDIMENSIONALES

COORDENADAS PRINCIPALES. Es un conjunto de técnicas estadísticas -matemáticas, para encontrar una configuración de puntos a partir de una matriz de distancias. Para usar el escalamiento multidimensional se requiere necesariamente que las distancias sean euclideanas.

Como ilustración de la técnica considere el siguiente ejemplo: suponga un mapa de Colombia y un conjunto de ciudades; se solicita construir una tabla (matriz) de distancias entre las ciudades; simplemente con una regla se medirían las distancias en el mapa y luego se convertirían a distancias reales en kilómetros por ejemplo. Ahora considere el problema inverso: dada una matriz de distancias entre las ciudades construya el mapa (las coordenadas).

En primer término, dado un conjunto de distancias euclideanas, no existe una representación única de puntos que origine las distancias; así, si conocemos la distancia entre Cali - Ibagué no sabemos si Cali está al oriente - occidente - norte ó sur de Ibagué. Técnicamente significa que no conocemos la localización y orientación de la configuración. El problema de localización se resuelve colocando el centro de gravedad de la configuración en el orígen. El problema de orientación se resuelve mediante una transformación ortogonal, de tal forma que los ángulos y distancias no se modifiquen.

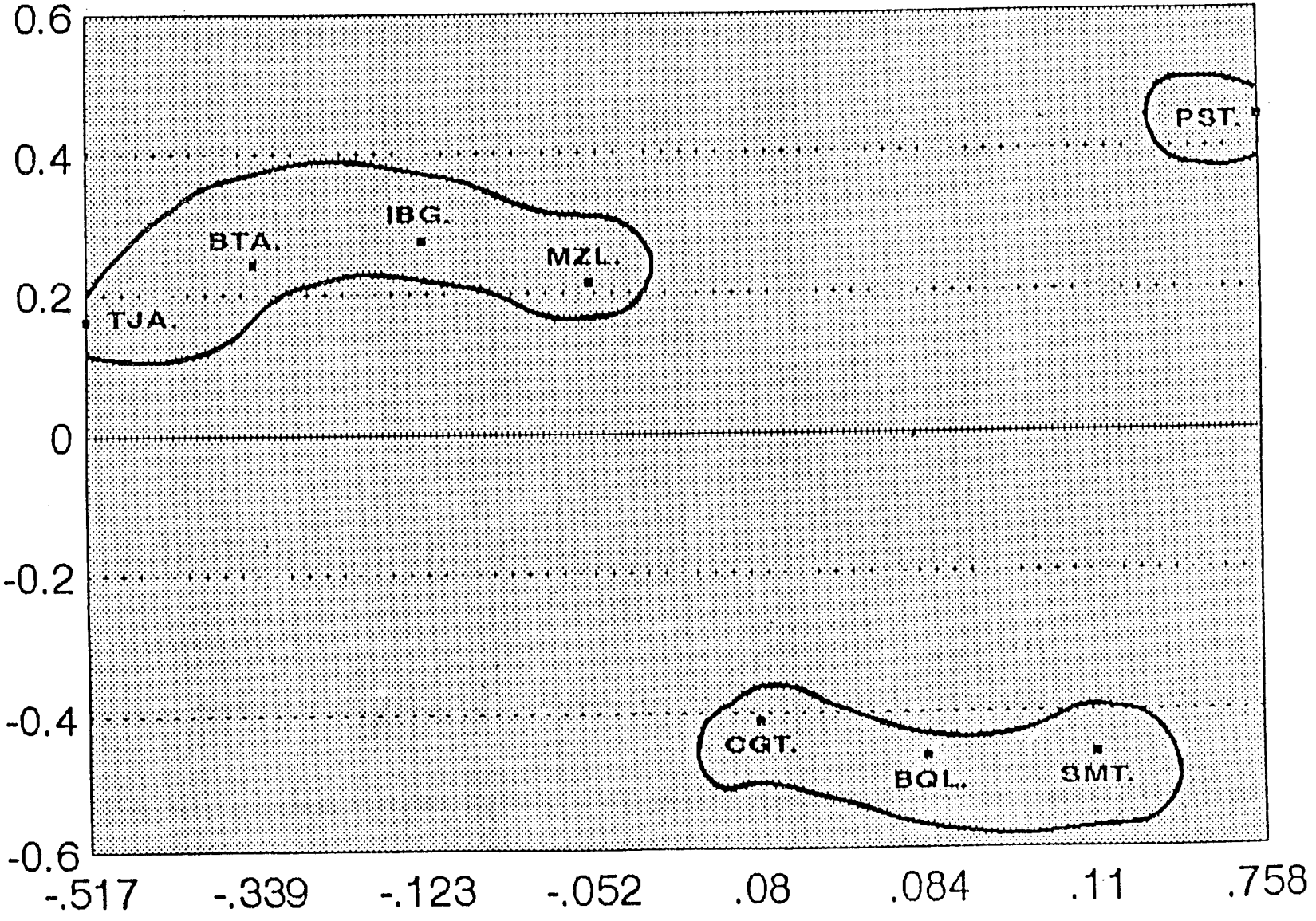
La aplicación de esta técnica estadística a los datos provenientes de ensayos biotecnológicos agrícolas, es inmediata ya que a partir de las bandas electroforéticas, se construyen índices de similaridad y con estos distancias euclideanas, a las cuales se aplican las escalas multidimensionales, para encontrar un plano de coordenadas principales donde las relaciones de semejanza y divergencia entre poblaciones biológicas se discriminan y cuantifican con cierto grado de sencillez.

A continuación, como ejemplo, se presenta una matríz de distancias entre algunas ciudades de Colombia y el uso de las escalas multidimensionales para reproducir un mapa (las coordenadas) de las ciudades. Se observa que el mapa preserva la distancia y localización de las ciudades. En la figura siguiente se observa la aplicación de las escalas multidimensionales a una colección de café, con datos provenientes de experimentos con RAPDS.

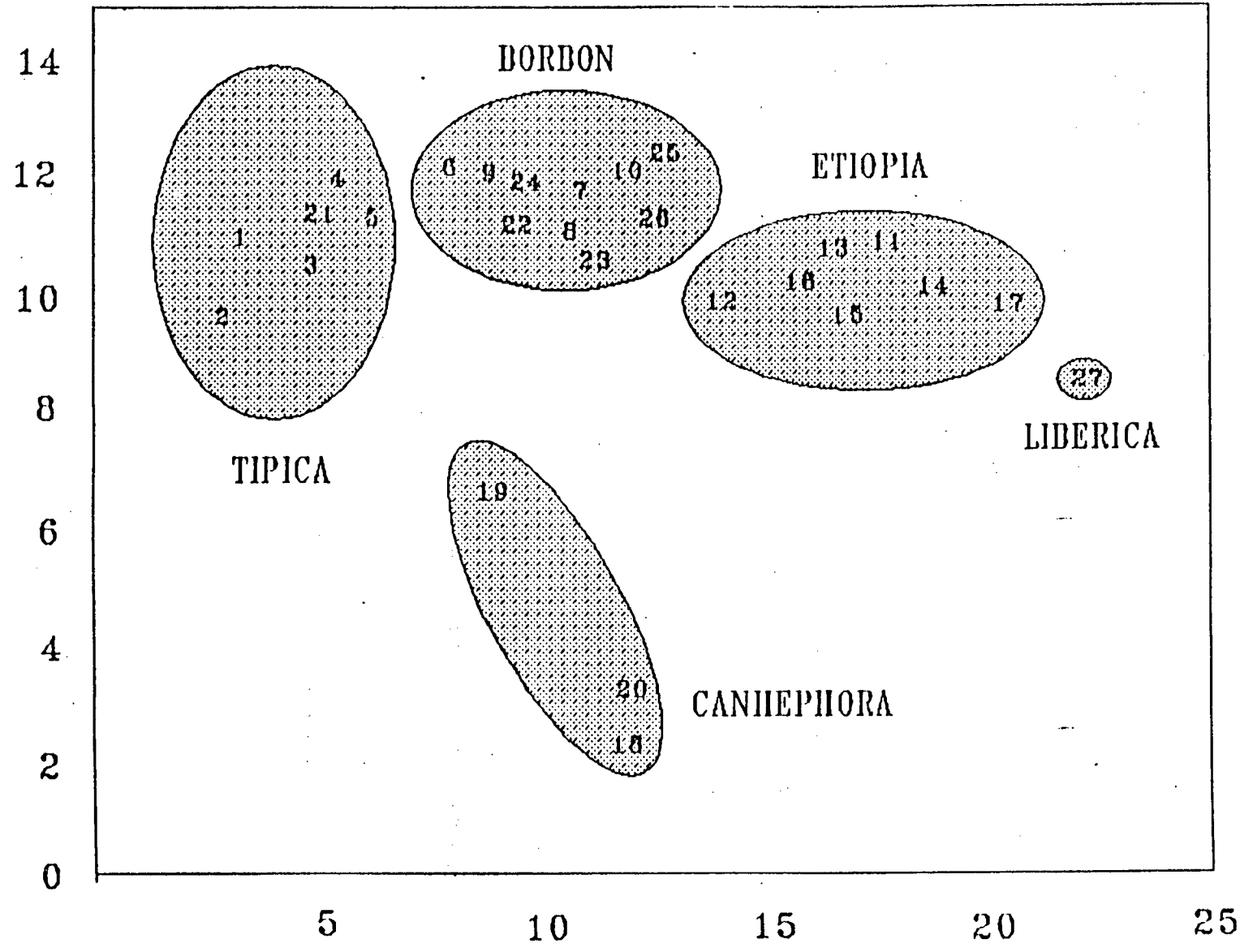
BTA	0							
IBG	1.2	0						
MZL	1.7	0.7	0					
PST	2.6	2.2	2.4	0				
CTG	3.4	3.2	2.9	5.2	0			
BQL	3.6	3.7	3.4	5.5	0.5	0		
STA	3.7	3.8	3.5	5.7	0.8	0.7	0	
TJA	0.7	1.2	1.4	3.4	2.9	3.2	3.3	0

DISTANCIA ENTRE CIUDADES DE COLOMBIA

MSD: COLOMBIA



COORDENADAS PRINCIPALES RAPDS



BIBLIOGRAFIA

1. Ayala, J.F. y J.A. Kiger. 1984. Genética moderna.

2. Gower, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis Biometrika: 53:325-328.

3. Kochet, G. 1994. Introduction to RFLP mapping and plant breeding applications.

4. McCalla, A.F. 1994. Priorites and Problems: the challenges facing world agriculture. V International Congress for Computer Technology in Agriculture. Royal Agricultural Society of England.

5. Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, N.Y.

6. Sokal, R.R. and Sneath, P.H.A. 1963. Principles of numerical taxonomy, London: Freeman.

7. William, J.G.K. et al. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers Nucleic Acids Research. 18-6531-6535.

8. Welsh, J. and M. McClelland. 1970. Finger printing genomes using PCR with arbitrary primers. Nucleic Acids Research. 18:7213-7218.