Agronomía colombiana, 1995, Volumen XII No.1; pág 66-71

MÉTODOS ESTADÍSTICOS MULTIVARIADOS EN BIOLOGÍA MOLECULAR Y SU APLICACIÓN EN INVESTIGACIÓN AGRÍCOLA

Orlando Martínez Wilches[1]

RESUMEN

Los métodos estadísticos como elementos de apoyo en la investigación agrícola son fundamentales, puesto que cuantifican y cualifican objetivamente los resultados de la investigación. Estos métodos y procedimientos estadísticos, varían según la naturaleza y estructura del resultado experimental. Así, si las ciencias biológicas básicas proponen e innovan procedimientos y técnicas que describan la variabilidad de poblaciones agronómicas, entonces es necesario proveer de herramientas estadísticas a las nuevas propuestas experimentales.

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

MULTIVARIATE STATISTICAL METHODS IN MOLECULAR BIOLOGY: AND THEIR USE IN AGRONOMIC RESEARCH

SUMMARY

Statistical methods as support elements in the agronomic research are basic. The importance rely on their objetive capacity of quantify and qualify the results of the investigations. Statistical methods to be applyed vary according to the structure and nature of the experimental result. Hence, if the basic biological sciences propose or introduce methods and techniques that describe the variability of agronomic populations, it is necessary to provide of statistical tool to the new experimental biological propositions.

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.

INTRODUCCIÓN

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 ambiente donde los recursos naturales disponibles ya son finitos; y 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, la evolucion y el funcionamiento de los organismos biológicos. Inicialmente, se preveia 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 prevé serán promisorios y exitosos.

TÉCNICAS ESTADÍSTICAS MOLECULARES:

Los métodos y procedimientos estadísticos disponibles para el análisis de los resultados provenien-

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

tes de ensayos biotecnológicos se pueden agrupar en las siguientes categorías:

1. Aquellos que tienen como propósito evaluar la variabilidad, clasificación, estructura y composición genética de las poblaciones,
2. Los desarrollados para la construcción de mapas cromosómicos o genómicos, cuando se utilizan marcadores genéticos moleculares, y
3. Lo denominados QLT (Quantitative trait loci), los cuales 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 aquéllos que, en general, describen la variabilidad genética de las poblaciones. En particular, su uso se enfatiza 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 GENÉTICOS

Las variables, los caracteres o parámetros, que se han utilizado para observar y detectar la variabilidad presente en los seres vivos, son numerosas. Los marcadores genéticos son una clase de éstos 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, por ejemplo altura de planta, número de pétalos, longitud de la mazorca. Para su evaluación, se requiere de una medida, conteo o calificación.

Los marcadores bioquímicos están constituidos por las isoenzimas y las proteínas. Mediante la técnica de la electroforesis en gel, se hace posible el estudio de la variación de las proteínas y enzimas en organismos vivos, así: Las muestras de tejidos se homogenizan (muelen) para liberar las enzimas y proteínas de las células. El sobrenadante del homogenizado (parte líquida), se coloca en un gel de almidón, agar, poliacrilamida o alguna sustancia gelatinosa. El gel se somete, durante horas, a corriente eléctrica continua y cada proteína del gel migra en una dirección y velocidad, la cual 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 el 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 (por ejemplo 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 patron de migración de los fragmentos se observa directamente mediante manchas coloreadas de una manera similar a las isoenzimas y proteínas (Kochet, 1994).

Los marcadores moleculares, conocidos como RAPDS o 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 oligonucleotidos. 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 y 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).

CUANTIFICACIÓN DE LOS MARCADORES BIOQUÍMICOS Y MOLECULARES

Los resultados experimentales de un ensayo biológico donde se utilicen las proteínas, enzimas, RFLPS o RAPDS es el mismo son un conjunto de bandas coloreadas en el gel que representan el comportamiento de la variabilidad. Como ilustración, se consideran cinco colecciones de una especie agrícola, por ejemplo cacao, las cuales se sometieron a un estudio de diversidad enzimática. En la Figura 1, se presentan los resultados correspondientes a una corrida de la - esterasa y se observa el patrón (las bandas) de variación de las colecciones y, en la última columna, corresponde al estandard, el cual expresa todas las bandas posibles producidas por las cinco colecciones. El problema es como 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, ésto es, asignar el valor 1 si la banda está presente y cero si ésta no lo está. Al aplicar dicha función al ejemplo de la - esterasa, se obtiene el Cuadro 1 y ella refleja la variabilidad de las bandas pero ya de una forma cuantitativa y numérica, a la cual se le pueden proponer medidas estadísticas que expresen la diversidad enzimática entre las colectas en estudio.

Figura 1. Patrón de variabilidad de cinco colecciones de cacaco asociados con la ( ) esterasa.

	Colección					
ORDEN	A	B	C	D	E	STANDARD
1		-		-		-
2	-		-	-		-
3	-	-	-	-		-
4		-				-
5	-	-	-		-	-
6	-	-	-		-	-
7		-			-	-
8	-		-	-		-
9	-		-	-		-
10			-	-		-

Cuadro 1. Cuantificación de la α–β esterasa en cinco colecciones de cacao.

	Colecciones					
ORDEN	A	B	C	D	E	ESTANDARD
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 del Cuadro 1, sería aquella que relacionara el número de bandas (unos o ceros) que simultáneamente compartan las dos accesiones. El siguiente cuadro provee la información necesaria para relacionar las ausencias y presencias comunes entre el par de accesiones.

		B		
		1	0	
A	1	a	b	n= a+b+c+d
	0	c	d	

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

$S_{AB} = a/n$

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

entonces se puede calcular un cuadro (matriz) de coeficientes de similitud entre todas las colecciones.

Adicionales a las anteriores, se han propuesto diferentes índices de similitud. En el Cuadro 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, al evaluar el comportamiento ante una serie de estímulos congnoci-

tivos, la presencia y la ausencia de características son comunes.

Cuadro 2. Coeficientes de similitud.

Coeficiente	Interpretación	Autor
$\frac{a+b}{n}$	Igual peso a 0-0 y 1-1	Sokal, Michener 1958
$\frac{a}{a+b+c}$	No contabiliza 0-0	Jaccard, 1908
$\frac{2a}{2a+b+c}$	Doble peso a 1-1 no contabiliza 0-0	Dice, 1945
$\frac{2a}{b+c}$		Nei, 1987

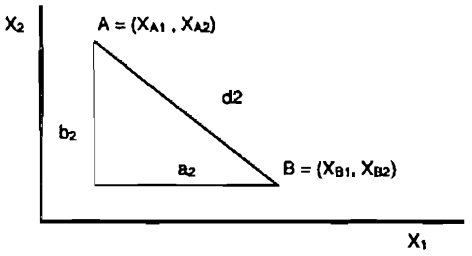
De los coeficientes o índices de similitud que se consignaron en el Cuadro 2, el de Jaccard posee ciertas características matemáticas y estadísticas que lo hacen más útil, sokal y Sneath (1963), Gower (1966). Por tal motivo se construyo la matriz de similitud entre todas las accesiones de cacao, con los resultados del Cuadro 1. La matriz de coeficientes de similitud de Jaccard se expresa a continuación:

A	1.00				
B	0.33	1.00			
C	0.86	0.30	1.00		
D	0.50	0.25	0.63	1.00	
E	0.29	9.29	0.25	0.13	1.00

Matriz de coeficientes de similitud de Jaccard para las cinco accesions de cacao.

DISTANCIAS EUCLIDIANA - GEOMÉTRICA - GENÉTICA

La distancia Euclidiana entre dos colectas es la aplicación del Teorema de Pitagoras, a dos características (X_1 y X_2) de A y B, asi:



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

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

para k - características.

$$D_{AB}{}^2 = \sum_k (X_{AK} - X_{BK})^2$$

La distancia euclidiana es intuitivamente atrayente, fácil de entender, es una medida geométrica que posee numerosas características algebráicos - 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 o colectas, divergencia atribuible exclusivamente a genes o 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:

$$D_{AB}{}^2 \sum_i (P_i - q_i)^2$$

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

DISTANCIAS E INDICES DE SIMILITUD

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

$D^2_{ij} = 1 - 2S_{ij}$

$D^2_{ij} = 2\,(1 - S_{ij})$

$D^2_{ij} = 1\,(1 - S_{ij})$

Sin embargo, tal como lo muestra Gower (1966), no siempre es posible calcular distancias euclideanas a partir de similaridades. Para lograr la conversión, la matriz de similaridad tiene que ser definida semi positiva. De las similaridades expresadas en el Cuadro 2, solamente, las definidas por Sokal y Michener (1958) Jaccard (1908) poseen esta condición. Gower (1966) enfatiza también en usar la expresión $2(1 - S_{ij})$.

A continuación, se expresa la matriz de distancias entre las colectas de cacao, la matriz se calcula mediante la expresión $2(1 - S_{ij})$ propuesta por Gower y a partir de la matriz de coeficientes de similitud (S_{ij}) previamente estimados.

A	0.00				
B	1.34	0.00			
C	0.28	1.40	0.00		
D	1.00	1.50	0.74	0.00	
E	1.42	1.42	1.50	1.74	0.00

Matriz de distancias entre las accesiones de cacao.

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, por ejemplo 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. Entonces el siguiente paso, es 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 gráfico- estadísticos que proveen al investigador que usa marcadores moleculares y bioquímicos en la descripción de la variabilidad de poblaciones biológicas de buenas guías. Los dendogramas y las escalas multidimensionales resumen la matriz de distancias.

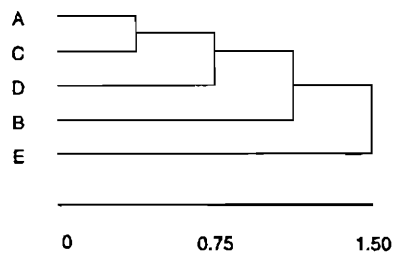
DENDOGRAMAS - CONGLOMERADOS

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

Los grupos establecidos por el análisis forman particiones y subdivisiones en conjuntos menores o 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 o la estructura en árbol, se puede representar en un diagrama o figura bidimensional y a tal representación se conoce como "dendograma".

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

A continuación, se presenta el dendograma elaborado a partir de la matriz de distancias de las cinco colectas de cacao. El dendograma se obtuvo mediante el método de distancia mínima (Single linkage). En cuanto a los métodos de construcción (algoritmo) de los dendogramas, son diversos y remitiremos al lector consultar el lobri de Sokal y Sneath (1963) o también otros más recientes de análisis multivariado, los cuales describen con detalle los procedimientos existentes para la elaboración de los dendogramas.



El dendograma permite establecer relaciones de similitud entre las colectas A y B e incluso de estas con D y también cierto grado de diferenciación entre las colectas B y E con el resto de accesiones. Las bandas de la Figura 1, poco o nada ofrecían al

investigador en cuanto a similitud y diferenciación entre las accesiones.

Por lo tanto, al patrón de bandas electroforéticas provenientes de las isoenzimas, los RFLPS o los RAPDS, se le 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.

COORDENADAS PRINCIPALES

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

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 kilometros. 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 euclidianas no existe una representación única de puntos que origine las distancias y, así, si conocemos la distancia entre Cali - Ibagué no sabemos si Cali está al oriente - occidente - norte o 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 origen. 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.

BIBLIOGRAFÍA

1. **Ayala, J.F. y J.A. Kiger.** Genética moderna. 1984.
2. **Gower, J.C.** Some distance properties of latent root and vector methods used in multivariate analysis Biometrika: 53:325-328. 1966.
3. **Kochet, G.** Introduction to RFLP mapping and plant breeding applications. 1994.
4. **Nei, M.** Molecular Evolutionary Genetics. Columbia University Press, N.Y. 1994.
5. **Sokal, R.R.** and Sneath, P.H.A. Principles of numerical taxonomy, London: Freeman. 1987.
6. **William, J.G.K.** et al. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers Nucleic Acids Research. 18-6531-6535. 1963.
7. **Welsh, J.** and M. McClelland. Finger printing genomes using PCR with arbitrary primers. Nucleic Acids Research. 18:7213-7218. 1990.
8. **Welsh, J.** and M. McClelland. Finger printing genomes using PCR with arbitrary primers. Nucleic Acids Research. 18:7213-7218. 1970.