MÉTODOS ESTADISTICOS APLICADOS EN BIOLOGÍA MOLECULAR

Agosto 8-10, 17 2.001

CIAT

M.C. Duque- CIAT

MÉTODOS ESTADISTICOS APLICADOS EN BIOLOGÍA MOLECULAR

Agosto 8-10, 17 2.001

CIAT

M.C. Duque- CIAT

CONTENIDO

1. Introducción
2. Características de los datos en estudios con Marcadores Moleculares
3. Análisis de Clasificación - Principios Generales
4. Software: NTSYS aspectos básicos.
5. Variabilidad en una clasificación: Métodos de Remuestreo
6. Software: Clasifi-Winboot
7. Análisis de Correspondencia Múltiple
8. Software: SAS para Análisis de Correspondencia
9. Diversidad - Heterogeneidad
10. QTL's : Nociones Básicas
11. Muestreo de Poblaciones

INTRODUCCION

Década :60's (segunda mitad)

Eventos : Introducción de técnicas bioquímicas y moleculares tales como: Clonación, Secuenciación de ADN y Uso de Enzimas para Restricción.

Resultados: Grandes logros sobretodo en estudios de evolución molecular y diversidad genética

El estudio del ADN:

* Ha aclarado interrogantes sobre cambio evolutivo de genes y dinámica de poblaciones

* Permite el desarrollo de nuevas teorías sobre la evolución de caracteres morfológicos, fisiológicos y conductuales.

Introducción (continuación...)

Nei (1987) señala que la Genética Evolutiva es una ciencia multidisciplinaria que compromete:

-Diferentes ramas de la Biología

-Matemática

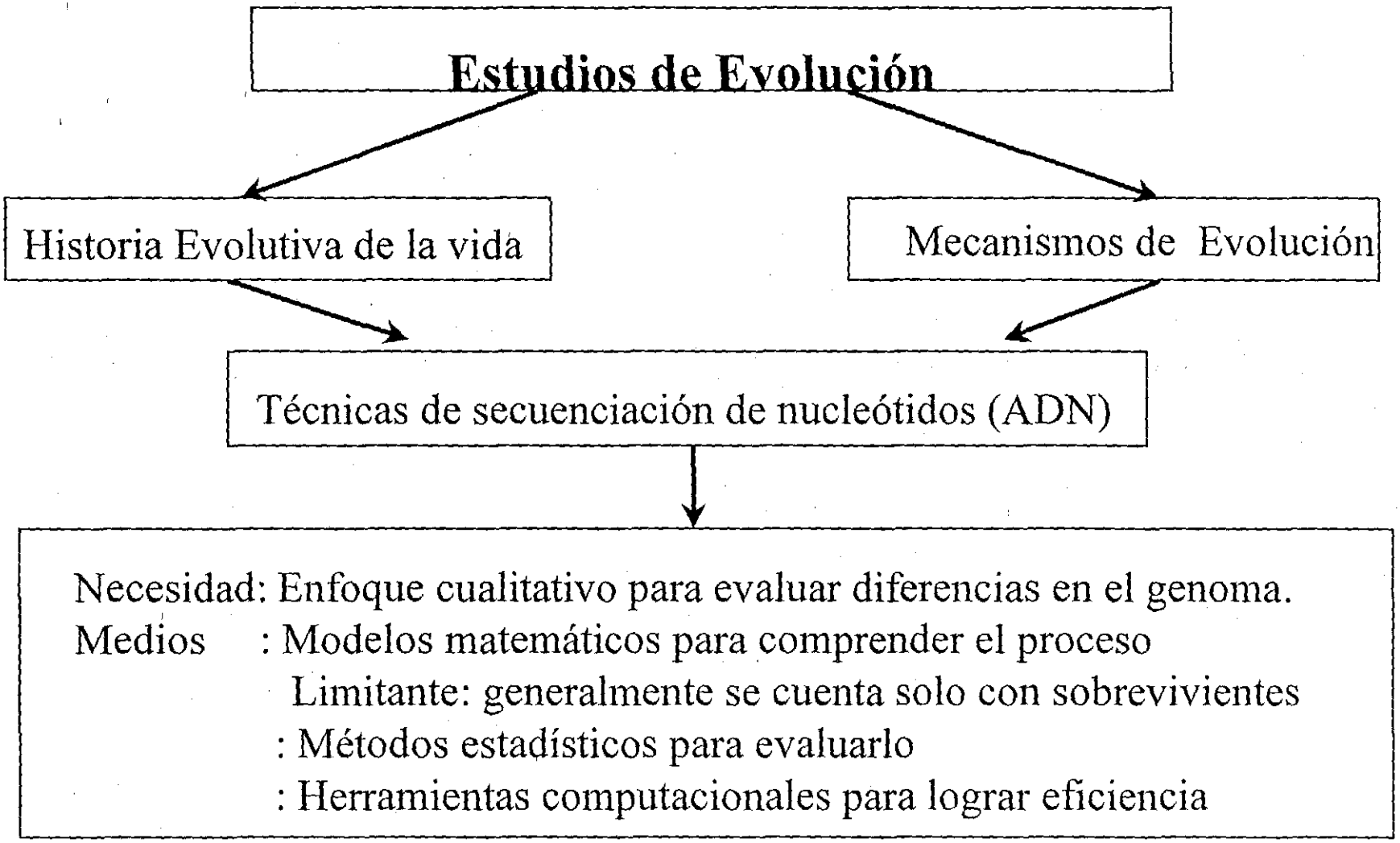
-Estadística

-Ciencias de la computación

tanto por razones de enfoque como de medios

Introducción (continuación)

Razones de Enfoque



Introducción (continuación...)

Razones de Medios

Las relaciones y la diversidad genética **dentro y entre** poblaciones y la comprensión de la filogenia y evolución debían apoyarse **antes** en variables morfológicas.

Hoy, con con el desarrollo de técnicas que permiten aislar **macromoléculas celulares** se cuenta con el uso de **marcadores moleculares** para resolver estas inquietudes.

Bachmann 1994; Thormann et al. 1994; dos Santos et al. 1994; Tohme et al. 1994; Mazur y Tingey 1995.

Introducción (continuación...)

Macromoléculas : Proteínas y Acido Desoxirribonucléico ADN

1. Proteínas:

-**Isoenzimas:** Enzimas con diferente estructura e igual función.

-**Alozimas** : Isoenzimas derivadas de diferentes formas alélicas del mismo locus.

Los polimorfismos de proteínas sólo reflejan variación de las regiones codificantes del genoma por lo que se ha dudado de su capacidad para reflejar el valor real de la diversidad genética.

Reflejan interacciones genotipo ambiente.

Introducción (continuación...)

Marcadores Moleculares para ADN

Las técnicas más representativas para al obtención de Marcadores Moleculares para ADN se mencionan a continuación:

RFLP: "Restriction Fragment Length Polymorphisms"

RAPD: "Random Amplified polymorphic DNA"

AFLP: "Amplified Fragment Length Polymorphisms"

SSR : "Simple Sequence Repeat Polymorphisms"

En términos generales, ellos detectan polimorfismos evaluando variaciones en la secuencia del ADN en algunos sectores del genoma, los cuales se manifiestan con diferentes patrones de bandeo en electroforesis .

Introducción (continuación...)

Marcadores Moleculares para ADN
Generación de bandas

La generación de bandas se logra de dos formas principalmente :

- Cortando el ADN con Enzimas de Restricción

-Utilizando la Reacción en Cadena de la Polimerasa (PCR) ara amplificar selectivamente fragmentos de ADN.

Introducción (continuación...)

Marcadores Moleculares para ADN
Generalidades

RESTRICCION:

RFLP :Digestión con endonucleasas específicas para restricción del genoma e hibridación con SONDAS marcadas o no radiactivamente.

Se evalúa el polimorfismo de la longitud de los fragmentos de restricción.

Características: Lenta. A veces no refleja mucho polimorfismo. Dispendiosa. 1-2 Locus . Codominante

Usos :Grupos Cercanos

Introducción (continuación...)

Marcadores Moleculares para ADN
Generalidades

PCR

RAPD: Amplificación aleatoria de fragmentos.

El polimorfismo refleja diferencia en la secuencia (la amplificación se produce en diferentes sitios).

Características: Detectan polimorfismos más efectivamente que RFLP. Rápidos. Dominantes

Usos: Mejoramiento vegetal, estudios genéticos

SSRP: Loci de secuencias cortas repetidas.

Polimorfismos significan variación en el número de copias

Características:1-2 Locus. Hipervariables. Lentos Costosos . Detectan cambios muy sutiles

Usos : Mamíferos, Poblaciones cercanas.

Sigue...

Introducción (continuación...)

Marcadores Moleculares para ADN
Generalidades

PCR

AFLP:Detectan fragmentos de restricción por medio de amplificación con PCR Por lo tanto son combinación de restricción y amplificación.

Polimorfismos detectan presencia o ausencia del fragmento de restricción.

Características: Todo el genoma. Detectan altos niveles de polimorfismo.

Nota: PCR necesita menos cantidad de ADN para cada reacción

Introducción (continuación...)

Marcadores Moleculares para ADN
Comparación

	AFLP	RFLP	RAPD	SRRP
	Hasta 120 locus	1-2 locus	1-8 locus	1-2 locus
RESTRICCIÓN	Diferentes tamaños de restricción	Diferentes tamaños de restricción		
AMPLIFICACION POR PCR	Diferente longitud del sector amplificado		Diferente longitud del sector amplificado	
				Diferente número de secuencias repetidas
			Diferentes sitios de unión del sebador con ADN	

Principales usos actuales de los marcadores moleculares

Mapeo genético

Características de importancia económica en plantas y animales.

Mejoramiento asistido por marcadores

Dactiloscopia del genoma (fingerprinting) (Criminalística-Paternidad) Búsqueda de genes afectados por enfermedades humanas, Pedigree animal-Variedades nuevas o mejoradas o depuradas ...)

Estudio de relaciones genéticas en poblaciones

Procesos evolutivos

Caracterización y evaluación de recursos genéticos

Diseño de fármacos - Métodos de diagnóstico.

M.C. Duque- CIAT

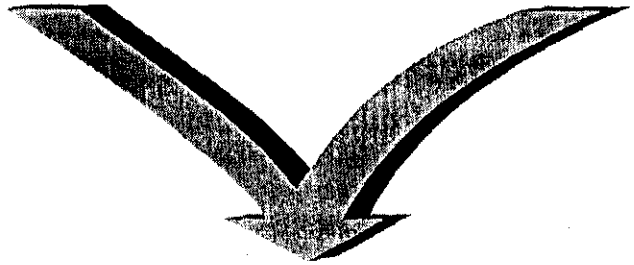
Características de los datos en estudios con

Marcadores Moleculares

M.C. Duque- CIAT

PROCESOS BIOLOGICOS

Observación **Medición**



ESPACIO MULTIDIMENSIONAL

- Gran número de variables
- Complejidad de relaciones

Los objetivos ocasionalmente requieren la formación de grupos descubriendo estructuras sugeridas por los datos.

No hay supuestos sobre el número ni sobre las características de los grupos.

BASE : Los elementos de un grupo deben ser similares entre sí y diferentes de los individuos en otro grupo.

CÓMO SE DEFINE SIMILARIDAD ?

La definición depende inicialmente del tipo de variables observadas o medidas

Cualitativas

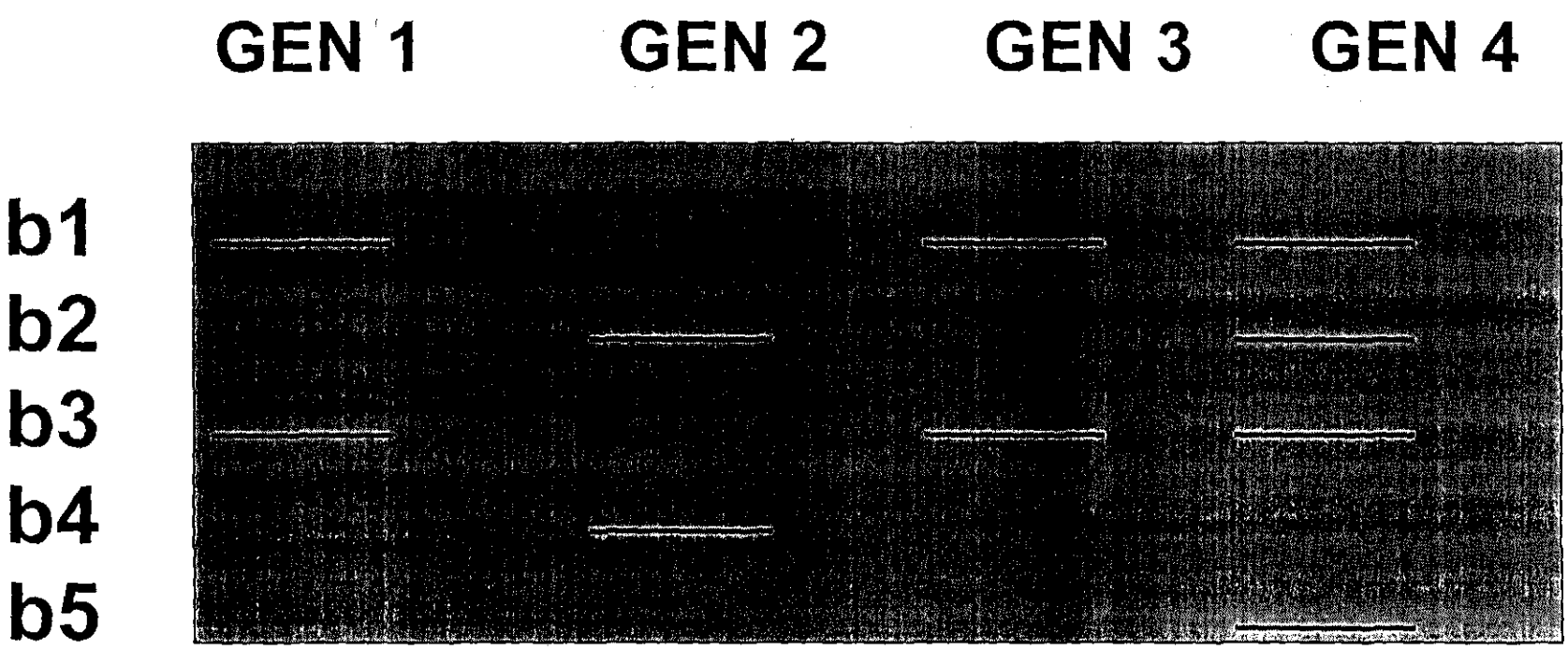
- **Binarias**
- **Categóricas**
  - **Nominales**
  - **Ordinales**

Cuantitativas

Ejemplo 1

Se tienen datos de autorradiografías que buscan realizar ´fingerprintings´ de genotipos de *Phaseolus vulgaris.*

Una representación esquemática es la siguiente:



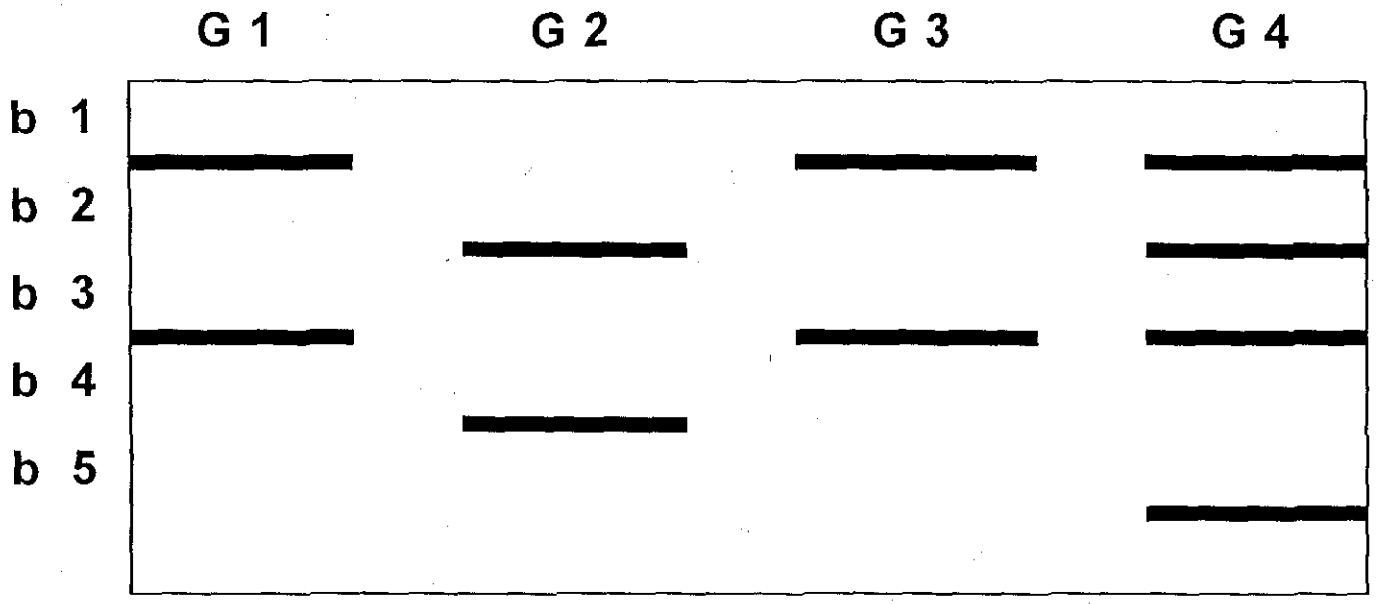
M.C. Duque- CIAT

De qué tipo son las variables ?

Cómo puedo establecer similaridad entre dos individuos ?

Paso 1

Organización de la información original



	banda 1	banda 2	banda 3	banda 4	banda 5
Gen. 1	1	0	1	0	0
Gen. 2	0	1	0	1	0
Gen. 3	1	0	?	?	?
Gen. 4	?	?	?	?	?

1 = Presencia 0 = Ausencia

ANALISIS DE CLASIFICACION

Principios Generales

M.C. Duque- CIAT

En qué se parecen, por ejemplo, los genotipos 1 y 4 ?

Genotipo 4

Genotipo 1		1	0	
	1	a=	b=	a+b
	0	c=	d=	c+d
		a+c	b+d	n=a+b+c+d

a: # de bandas presentes en ambos genotipos

b: # de bandas presentes sólo en el genotipo 1

c: # de bandas sólo en el genotipo 4

d: # de bandas ausentes en los 2 genotipos

n: # de bandas en el estudio

Algunas definiciones de similaridad

$$S = \frac{a + d}{n}$$

$$S = \frac{2a + 2d}{2a + b + c + 2d}$$

$$S = \frac{a}{a + b + c}$$

$$S = \frac{2a}{2a + b + c}$$

Cuál es el valor de similaridad de 2 genotipos idénticos ?

Y cuál la similaridad entre 2 genotipos totalmente diferentes ?

Los coeficientes que miden similaridad e incluyen el factor "d", para el ejemplo que consideramos, tienen poca importancia, debido a que la ausencia de la variable tiene poco significado biológico.

El coeficiente de Jaccard tiene en cuenta esta consideración, pero:

si hay falsos positivos o falsos negativos

el error (sesgo) en Jaccard >error (sesgo) en Nei-Li

Para el caso que estudiamos, por estimar similaridad, son importantes sólo los falsos positivos.

Además, el coeficiente de Nei-Li tiene completo significado a nivel de similaridad en términos de DNA:

es una estimación de la proporción esperada de los fragmentos amplificados que se comparten.

Ejemplo

Completar los valores faltantes en la matriz de distancia, utilizando $D_1 = 1\text{-}S$

$$D = \begin{matrix} D_{11}=0 & D_{12}=1 & D_{13}= & D_{14} = 0.33 \\ D_{21}= & D_{22}= & D_{23}=1 & D_{24}= \\ D_{31}=0 & D_{32}= & D_{33}= & D_{34}= \\ D_{41}= & D_{42}=0.67 & D_{43}= & D_{44}= \end{matrix}$$

Conclusión

La distancia es una función decreciente de la similaridad.

Algunas expresiones para el cálculo de la distancia son:

$$D_1 = 1 - S$$

$$D_2 = \sqrt{1 - S}$$

$$D_3 = \sqrt{(1 - S^2)}$$

Paso 3

Al definir similaridad entre individuos surge un concepto paralelo: DISTANCIA

Cómo debe ser la distancia entre dos individuos idénticos ?

Cómo se comporta la distancia a medida que la similaridad se reduce ?

$$A = \begin{vmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{vmatrix}$$

S con el coeficiente de Nei-Li :

$$S = \begin{vmatrix} S_{11}=1 & S_{12}= & S_{13}= & S_{14}=0.67 \\ S_{21}= & S_{22}= & S_{23}= & S_{24}= \\ S_{31}=1 & S_{32}= & S_{33}=1 & S_{34}= \\ S_{41}= & S_{42}=0.33 & S_{43}=0.67 & S_{44}= \end{vmatrix}$$

Ejemplo:

Se desea analizar el comportamiento de un grupo de genotipos de arroz, frente a diferentes aislamientos de *P. grisea.* Para tal fín, el conjunto de respuestas se ha transformado en una escala combinada, donde "R" significa resistente y "S" susceptible.

Tipo de lesión	% AFA	Calificación	Escala
0,1	1	HR	R
0,1,2	<10	HR	R
2	>10	R	R
3	<8	I	S
3	>=8	S	S
4	<5	S	S
4	>=5	HS	S

	Aislam 1	Aislam 2	Aislam 3	Aislam 4	Aislam 5
Arroz1	R	R	S	S	S
Arroz2	S	R	S	S	R
Arroz3	S	R	S	S	R
Arroz4	S	S	R	R	R
Arroz5	R	R	R	R	S
Arroz	R	S	R	R	S
Arroz7	R	S	R	S	R
Arroz8	S	S	S	S	S
Arroz9	R	R	R	R	R

	Aislam 1	Aislam 2	Aislam 3	Aislam 4	Aislam 5
Arroz1	1	1	0	0	0
Arroz2	0	1	0	0	1
Arroz3	0	1	0	0	1
Arroz4	0	0	1	1	1
Arroz5	1	1	1	1	0
Arroz	1	0	?	1	0
Arroz7	1	0	1	0	?
Arroz8	0	0	0	0	0
Arroz9	1	1	?	?	?

M.C. Duque- CIAT

Paso 4

Ya conocemos la distancia entre los individuos.

Cómo podemos formar grupos ?

Hay varias estrategias: Jerárquicas, no jerárquicas y de grupos traslapados. Sólo veremos las primeras:

Características de los datos en estudios de

VIRULENCIA

M.C. Duque- CIAT

Matriz de similaridad calculada con el coeficiente :

_____

dimensiones: ___filas x ___columnas

S21=

S31= S32=

S41= S42= S43=

S51= S52= S53= S54=

S61= S62= S63= S64= S65=

S71= S72= S73= S74= S75= S76=

S81= S82= S83= S84= S85= S86= S87=

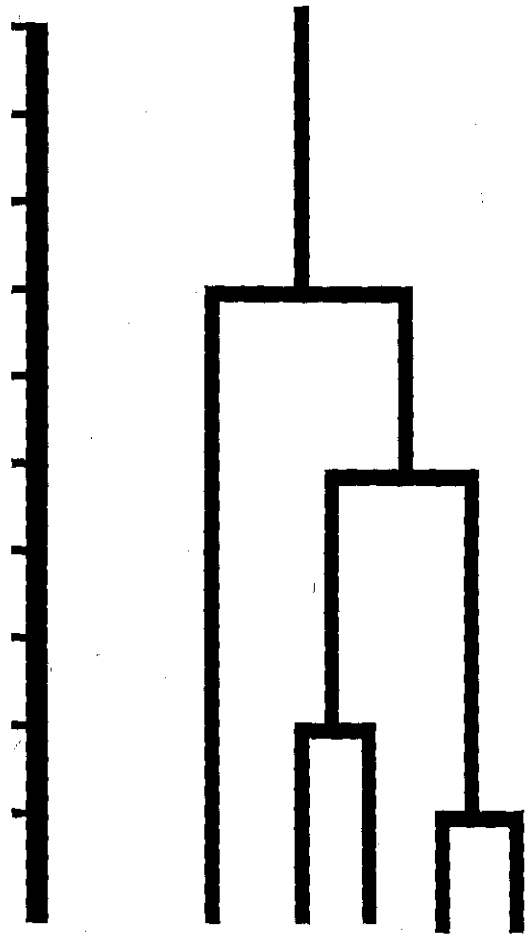
S91= S92= S93= S94= S95= S96= S97= S98=

Matriz de similaridad calculada con el coeficiente : Concordancia Simple

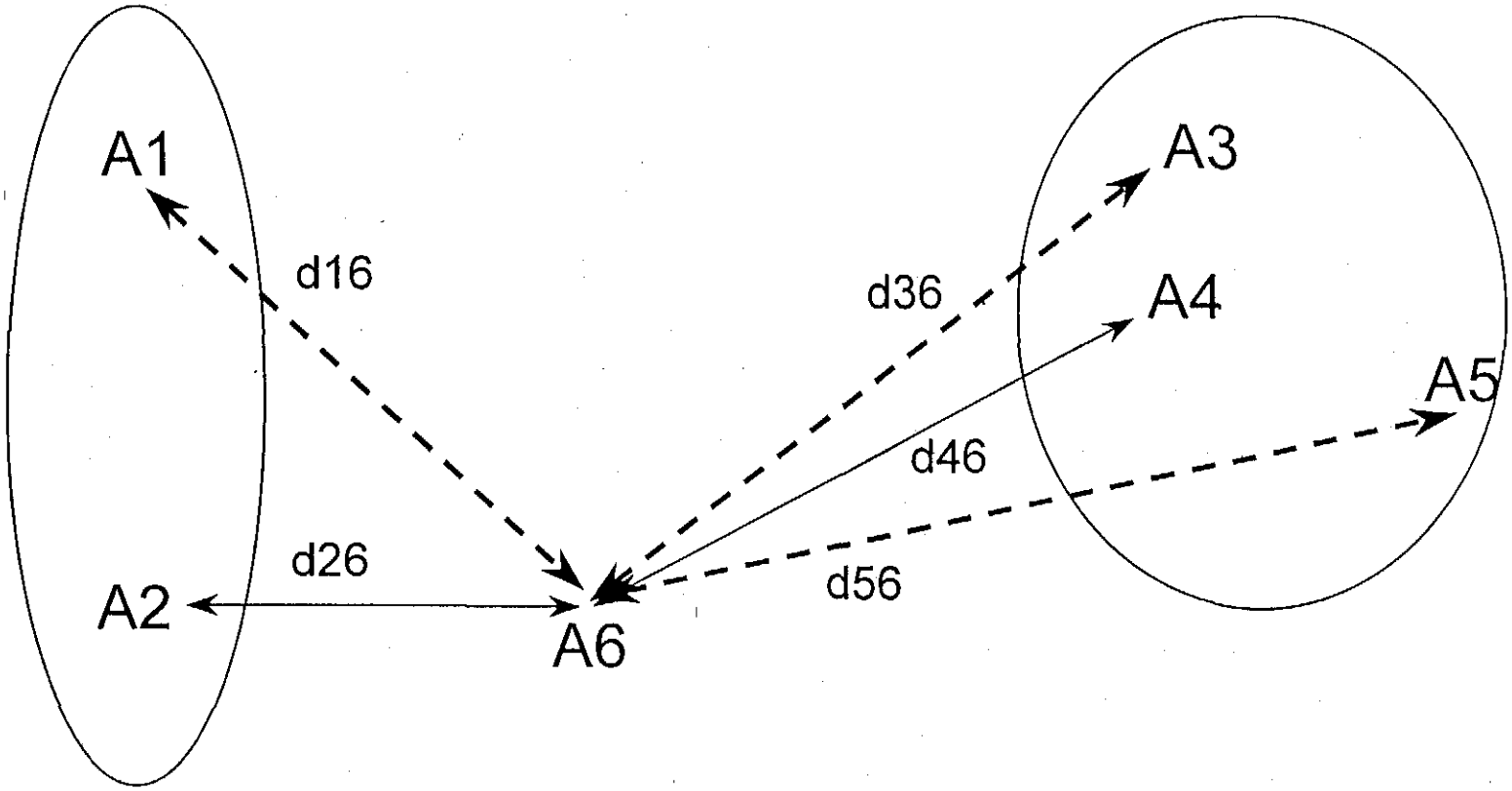
dimensiones: 9 filas x 9 columnas

0.60								
0.60	1.00							
0.00	0.40	0.40						
0.60	0.20	0.20	0.40					
0.60	0.20	0.20	0.40	0.60				
0.60	0.20	0.20	0.40	0.60	0.60			
0.60	0.60	0.60	0.40	0.20	0.60	0.60		
1.00	0.60	0.60	0.00	0.60	0.60	0.60	0.60	

- **El árbol se construye a partir de la consideración de cada individuo como un grupo, y uniendo los individuos o grupos más similares hasta llegar a un solo grupo.**



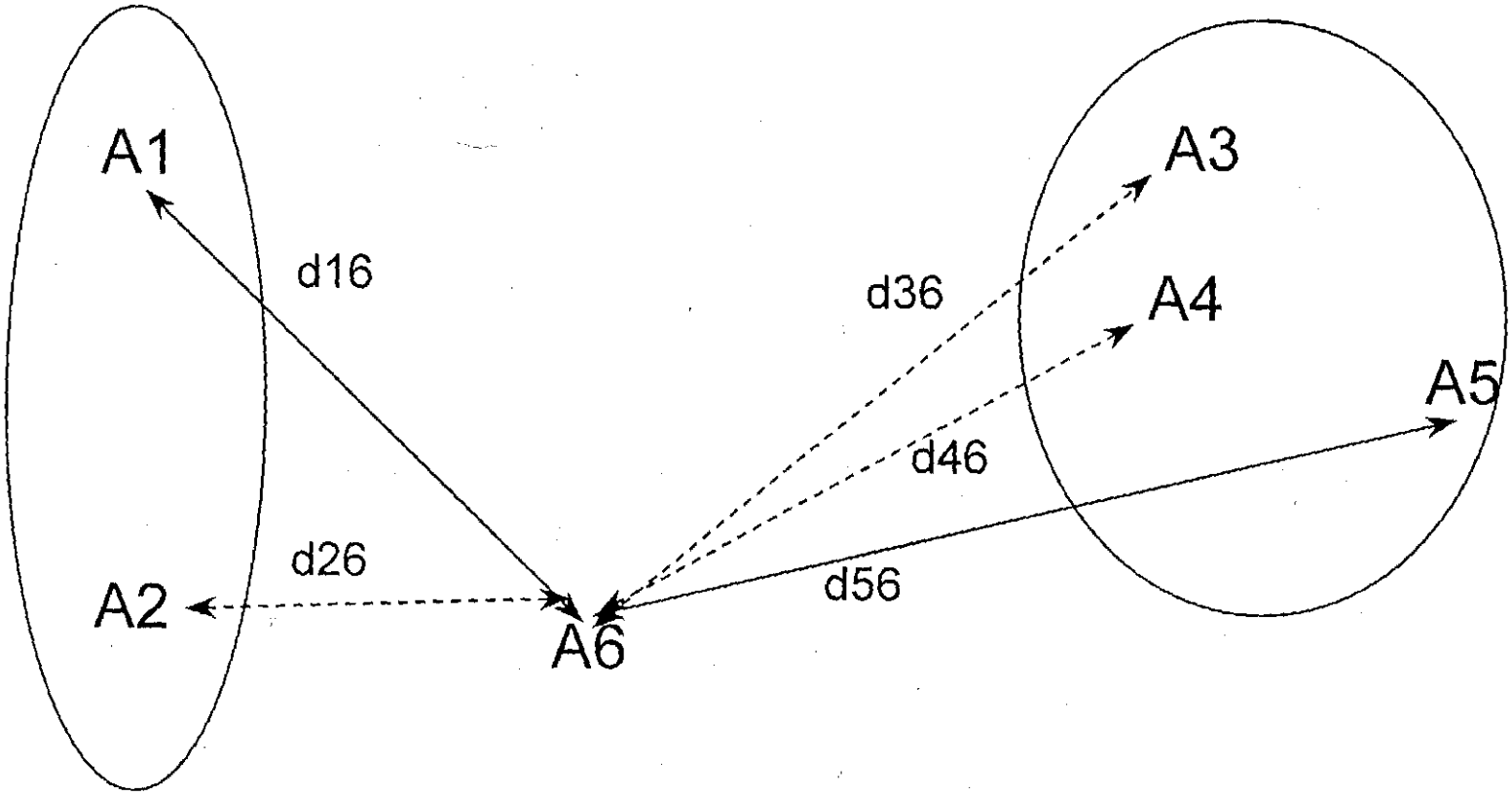
LIGAMIENTO O UNIÓN SIMPLE



d(A6, Grupo 1) = d26
d(A6, Grupo 2) = d46

d26 < d46

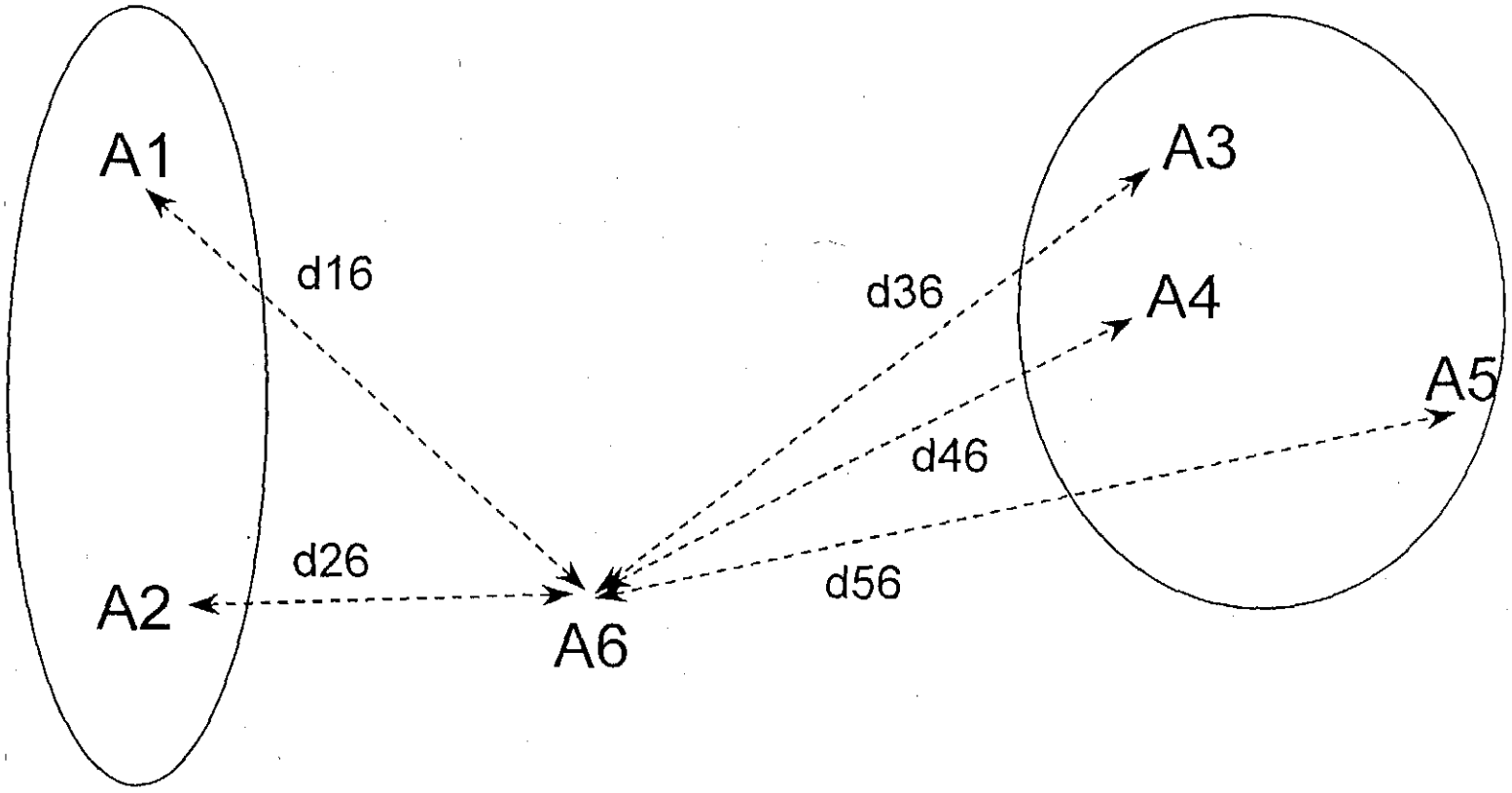
LIGAMIENTO O UNIÓN COMPLETA



d(A6, Grupo 1) = d16
d(A6, Grupo 2) = d56

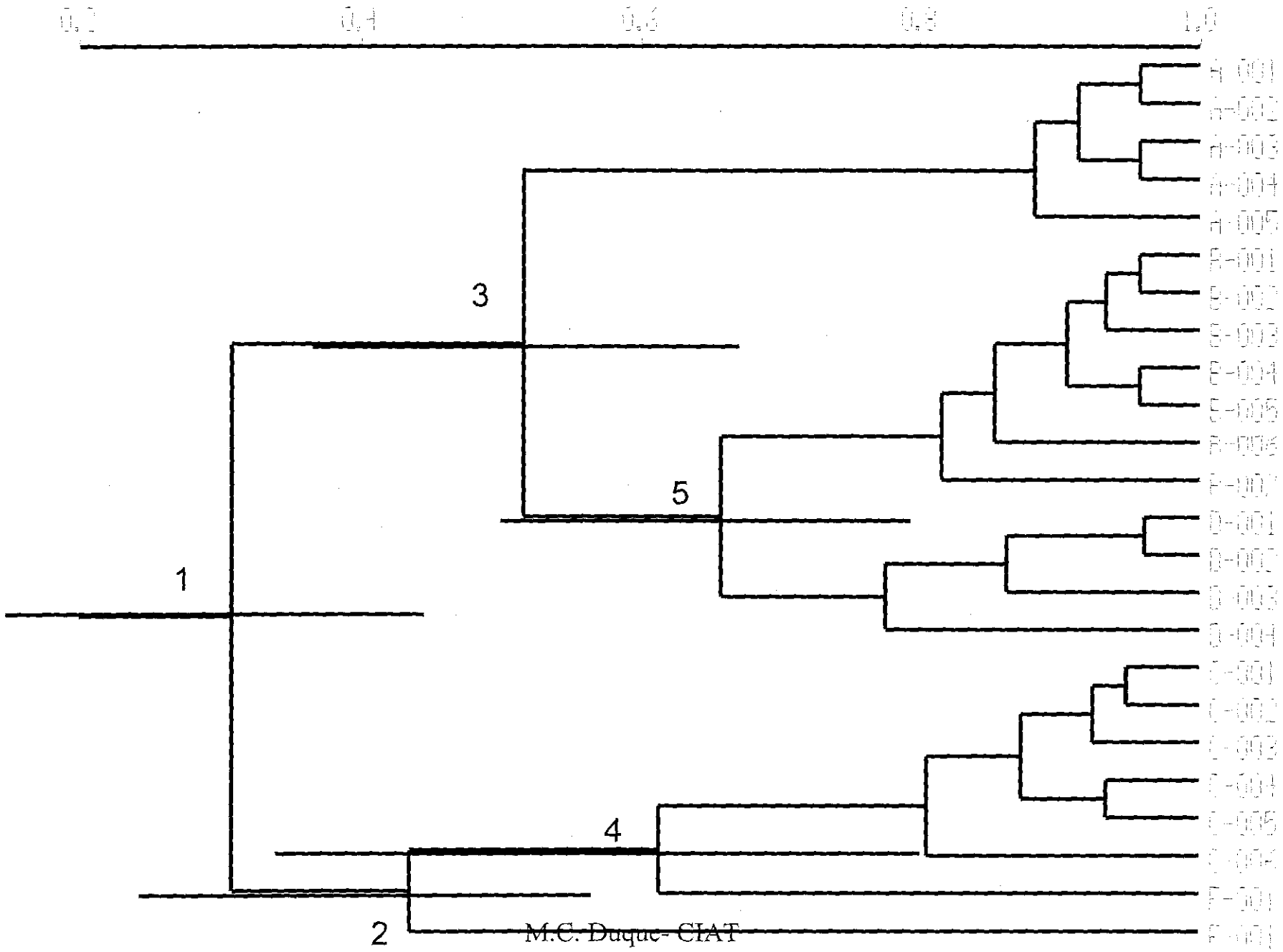
d16 < d56

LIGAMIENTO O UNIÓN MEDIA



d(A6, Grupo 1) = (d16 +d26) / 2

d(A6, Grupo 2) = (d36 + d46 + d56)/3


0,2
0,4
0,6
0,8
1,0
1
2
3
4
5
M.C. Duque- CIAT

Bootstrap

Modalidad 2:

Se pretende ver qué tan factible es cada rama formada.

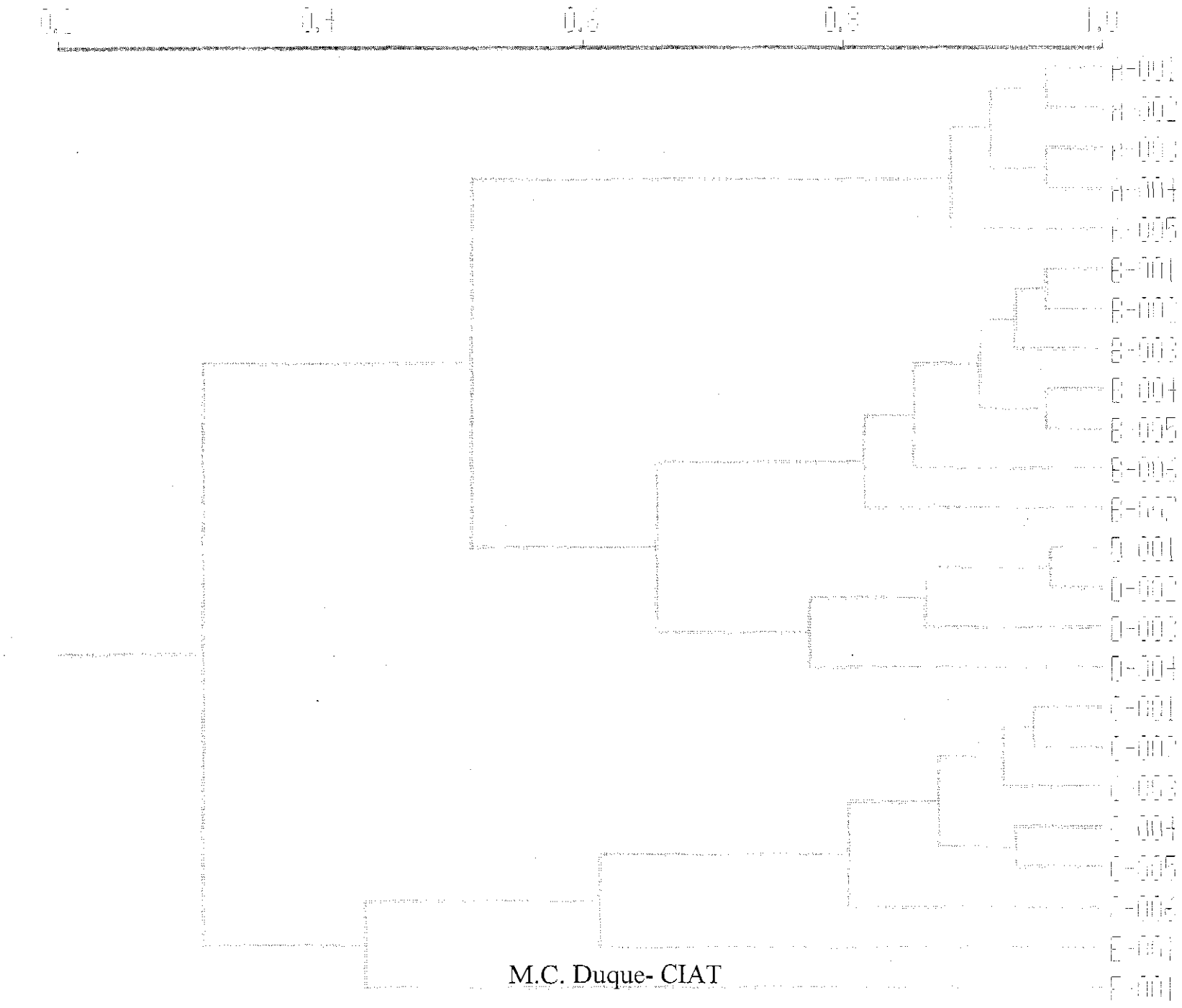
(Yap I, y Nelson R 1996)

ESTADISTICOS DE CADA NODO

Variable	Label	N	Mean	Variance	Std Dev	CV	Minimum	Maximum
M1	NODO-1	500	0.6945698	0.0072534	0.0851667	12.2617957	0.4286506	0.9365384
M2	NODO-2	500	0.5713613	0.0096804	0.0983888	17.2200620	0.3134264	0.9124318
M3	NODO-3	500	0.4888485	0.0097115	0.0985467	20.1589458	0.2407155	0.8162255
M4	NODO-4	500	0.3951715	0.0184582	0.1358608	34.3802068	0.0892774	0.8576923
M5	NODO-5	500	0.3464447	0.0081024	0.0900132	25.9819862	0.1375636	0.7185198

INTERVALOS DE CONFIANZA EN CADA NODO

OBS	N	MEAN	VAR	MEDIAN	P0_5	P2_5	P5	P95	P97_5	P99_5
1	500	0.69457	0.007253	0.68886	0.46907	0.52697	0.56278	0.83995	0.86626	0.92268
2	500	0.57136	0.009680	0.56978	0.35452	0.40133	0.41541	0.74506	0.77268	0.82404
3	500	0.48885	0.009711	0.48971	0.26911	0.30611	0.33796	0.65373	0.68754	0.77365
4	500	0.39517	0.018458	0.38660	0.10327	0.16429	0.19719	0.63007	0.68205	0.81740
5	500	0.34644	0.008102	0.33971	0.15353	0.18715	0.20715	0.50316	0.55155	0.65210



M.C. Duque- CIAT

Ejemplo:

Estudio de la estructura poblacional del patógeno del arroz *Xanthomonas oryzae* pv. *oryzae* con RFLP .

```
options ls=146 ps=62 ;
filename a   '~/acacat.prn'  lrecl =140;
filename c   '~/aagctc.prn';
filename d   '~/eaagmctt.prn' lrecl =140;
filename e   '~/acccta.prn'   lrecl =110;
filename f   '~/codmatr.prn'  lrecl =100;
*************************************************************;
*ENTRADA DE DATOS;
data a ; infile a ; input ide  1-4 (b1-b62)(2.);   s1=sum(of b1-b62);
data c ; infile c ; input ide  1-4 (b63-b96)(2.);  s2=sum(of b63-b96);
data d ; infile d ; input ide  1-4 (b97-b159)(2.); s3=sum(of b97-b159);
data e ; infile e ; input ide  1-4 (b160-b207)(2.);s4=sum(of b160-b207);
data f ; infile f  ;length genot $ 14. retro $40. especie $11.;
input ide   1-4 especie $ 5-15 genot $16-29 hibrido $30   retro $43-83;

proc sort data=a; by ide;proc sort data=c; by ide;
proc sort data=d; by ide;proc sort data=e; by ide;
proc sort data=f; by ide;

data a; merge a c d e f; by ide; ident=compress(ide||'-'||genot);
keep ident b1-b207;
*************************************************************;
options mautosource sasautos='~jgarcia/sas';
*************************************************************;
%clasifi(a,ident,1,5,1,1000,lclust=1,impr=1,std=0,tdistbi=1,salida=cart);
```

d. % CLASIFI (DATA1, IDENTIF, 3, 5, 2, METODOM = 2);

Igual al anterior, pero generando muestras JACKKNIFE, las cuales dependen
del número de variables en que son evaluados los individuos. Con este método de muestreo, el parámetro que indica el número de muestras aleatorias no es necesario.

e. % CLASIFI (DATA2, IDENTIF, 4, 5, 1, 100, METODOM = 3, LVCUALI = X_1-X_{10}, LVCUANT = X_{11}-X_{16}, IMPR = 0);

En este ejemplo se efectúa un análisis de clasificación a los individuos en el archivo DATA2, identificados con la variable IDENTIF sobre una mezcla (4) de variables cualitativas (X_1-X_{10}) y de variables cuantitativas (X_{11}-X_{16}).

Se desea analizar 5 nodos empezando con el nodo número 1 y tomando 100 muestras SPECIAL JACKKNIFE sin imprimir las matrices de distancia y de similaridad.

Ejemplos

a. **% CLASIFI (DATA1, IDENTIF, 3);**

Realiza CLUSTER ANALYSIS mediante el método AVERAGE a los individuos identificados por la variable IDENTIF cuya información sólo sobre variables cuantitativas está almacenada en el archivo DATA1.

b. **% CLASIFI (DATA1, IDENTIF, 3, LCLUST = 4);**

Además de lo anterior, lista los individuos de cada uno de los cuatro grupos que se desea formar.

c. **% CLASIFI (DATA1, IDENTIF, 3, 5, 2, 100);**

Realiza un análisis sobre 5 nodos, empezando con el nodo número 2, para lo cual la macro genera aleatoriamente 100 muestras tipo BOOTSTRAP (Default).

C L A S I F I

DOCUMENTO DE TRABAJO

UN PROGRAMA SAS PARA ANALISIS DE CLASIFICACION

James A. García, Myriam Cristina Duque, Joe M. Tohme, Shizong Xu y Morris Levy

3.2. Modo de uso

% CLASIFI (ARCH, IDENTIF, TIPOVAR, NODOS, NODINI, NMUESTR, METODOM = *, METODOC = *, LCLUST = *, LVCUALI = *, LVCUANT = *, IMPR = *, ARCHISAL = *, STD = *, SIMILB = *, TDISTBI = *, TDISTCU = * , SALIDA = *);

*: Indica un valor adecuado entre las opciones de cada parámetro.

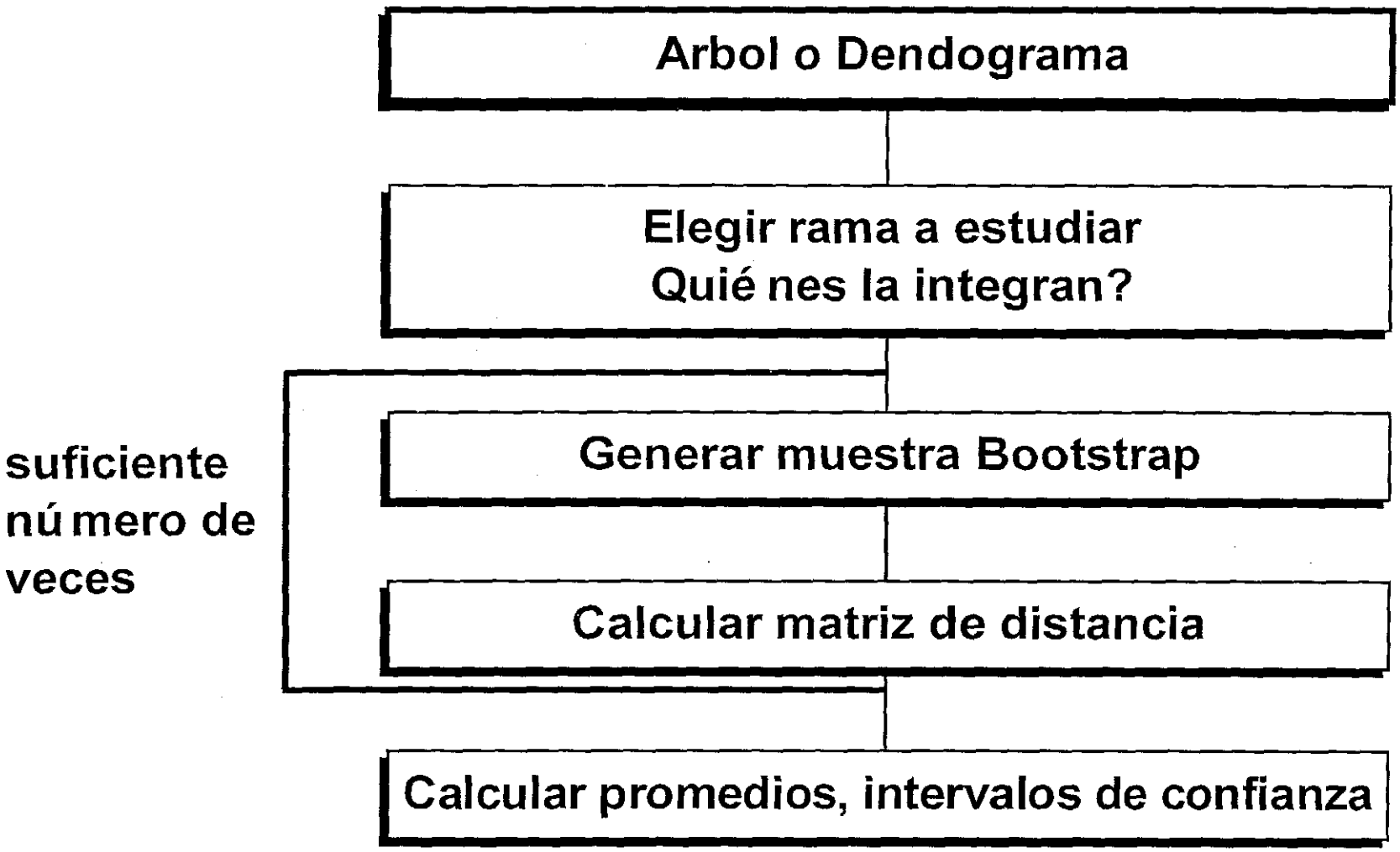
SOFTWARE:

CLASIFI

WINBOOT

M.C. Duque- CIAT

Bootstrap


Arbol o Dendograma
Elegir rama a estudiar
Quié nes la integran?
Generar muestra Bootstrap
Calcular matriz de distancia
Calcular promedios, intervalos de confianza
suficiente
nú mero de
veces

En el árbol o dendograma ya obtenido de la muestra inicial aparece el punto donde se unen el genotipo i y el j.

Ahora podemos consultar de cada muesta Bootstrap el valor de distancia entre i y j.

Con un número adecuado de muestras (1000) podemos calcular el valor promedio e intervalos de confianza para la distancia i y j.

Finalmente podemos tener una opinión objetiva sobre la significancia de las ramas.

Ejemplo:

muestra	bandas
1	1, 3, 2, 5, 1
2	4, 3, 1, 1, 5
3	1, 3, 4, 5, 2
4	4, 3, 2, 5, 5
.	

Cada muestra dará información para todos los genotipos en términos de las bandas que los definen.

Por lo tanto, cada muestra dará lugar a una matriz de similaridad y a una matriz de distancia.

Técnica del Bootstrap

Consiste en tomar muestras (con reemplazamiento) de los elementos que constituyen la muestra original.

Cada genotipo se expresa en un perfil de n bandas.

Elegiremos aleatoriamente muestras de n bandas, pero, al ser con reemplazamiento, una banda específica puede estar considerada más de una vez.

Bootstrap

Modalidad 1

Se estudiará el intervalo de confianza para cada nodo de interés

(García et al. 1995)

- por razones técnicas
- por razones económicas
- por tiempo

Qué hacer ?

Existen técnicas no paramétricas que utilizan el re-muestreo o muestreo repetido para resolver el problema.

Hasta el momento, conocemos algunos aspectos del conjunto de datos bajo estudio.

Pero, sabemos algo sobre su variabilidad ?

En ocasiones éste es el momento crítico por tener sólo una muestra :

- por disponibilidad del material

VARIABIALIDAD

M.C. Duque- CIAT

Software:

NTSYS

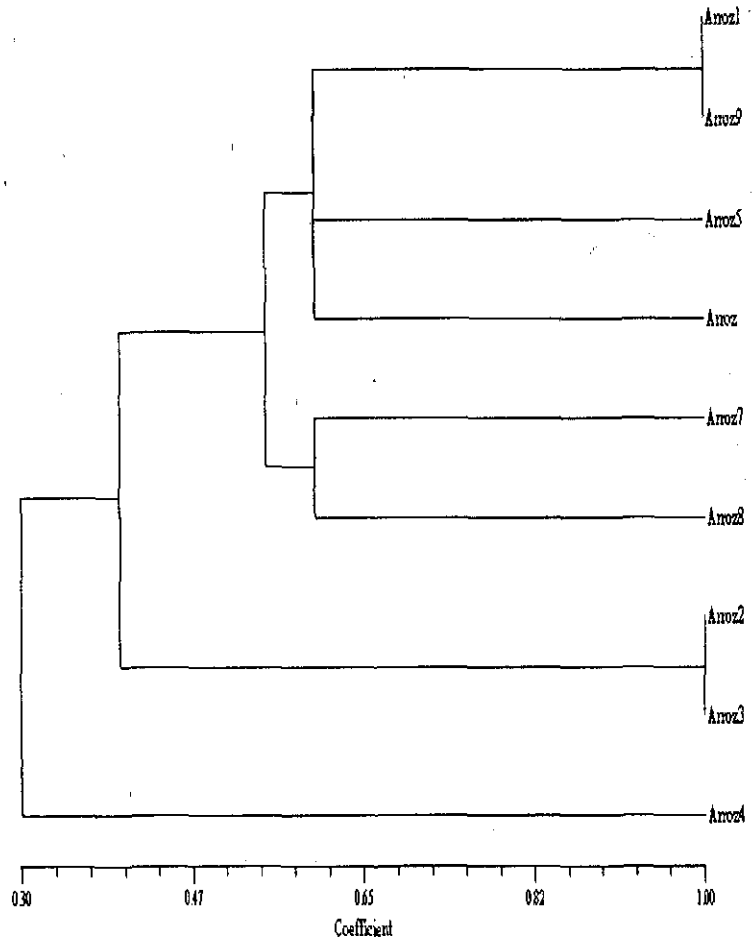
Aspectos Básicos

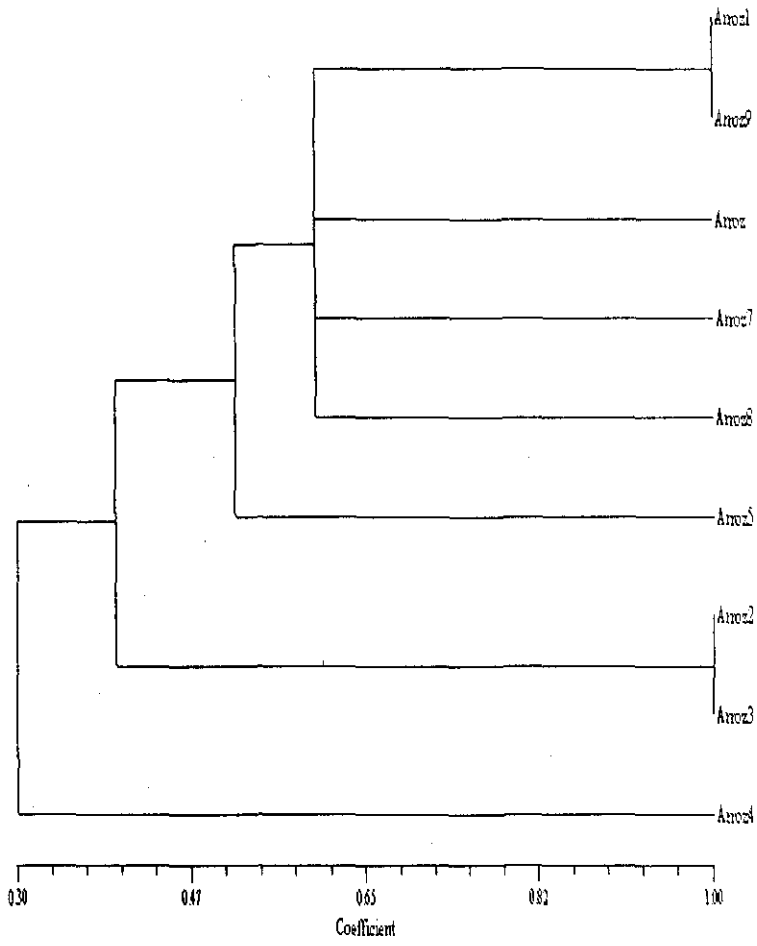
M.C. Duque- CIAT

Y en su opinión...

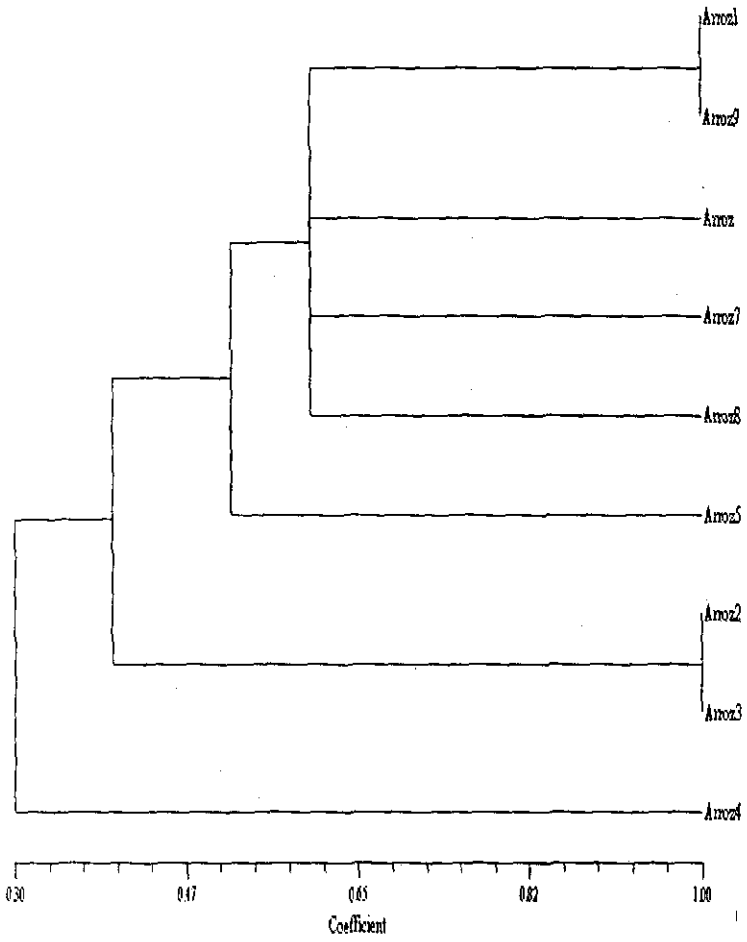
Cuál es el verdadero?

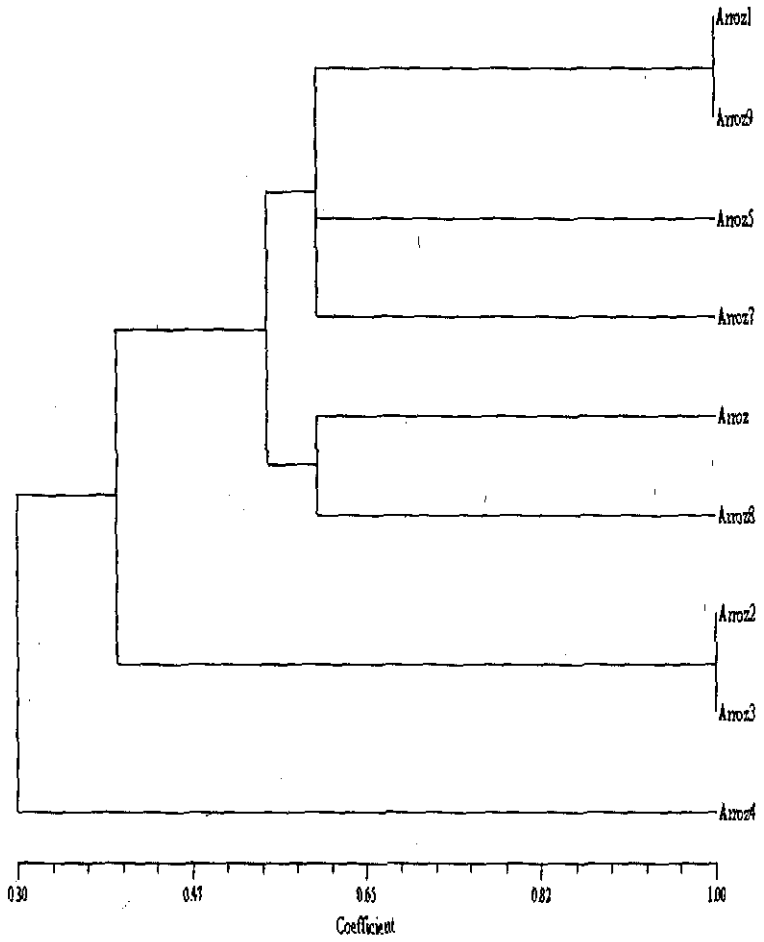
M.C. Duque- CIAT



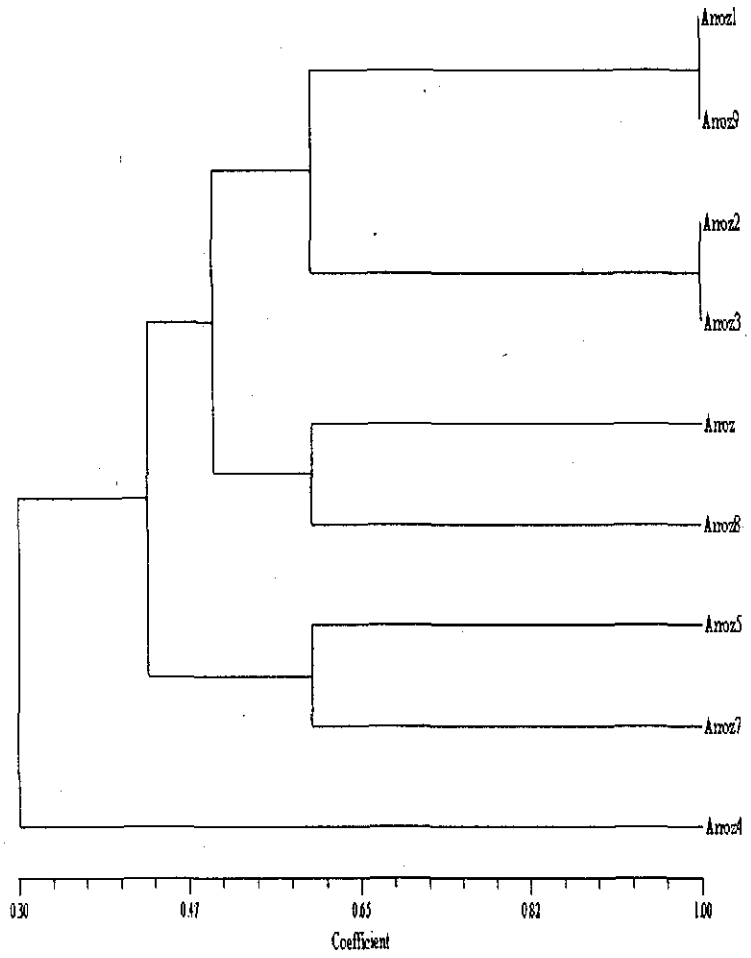


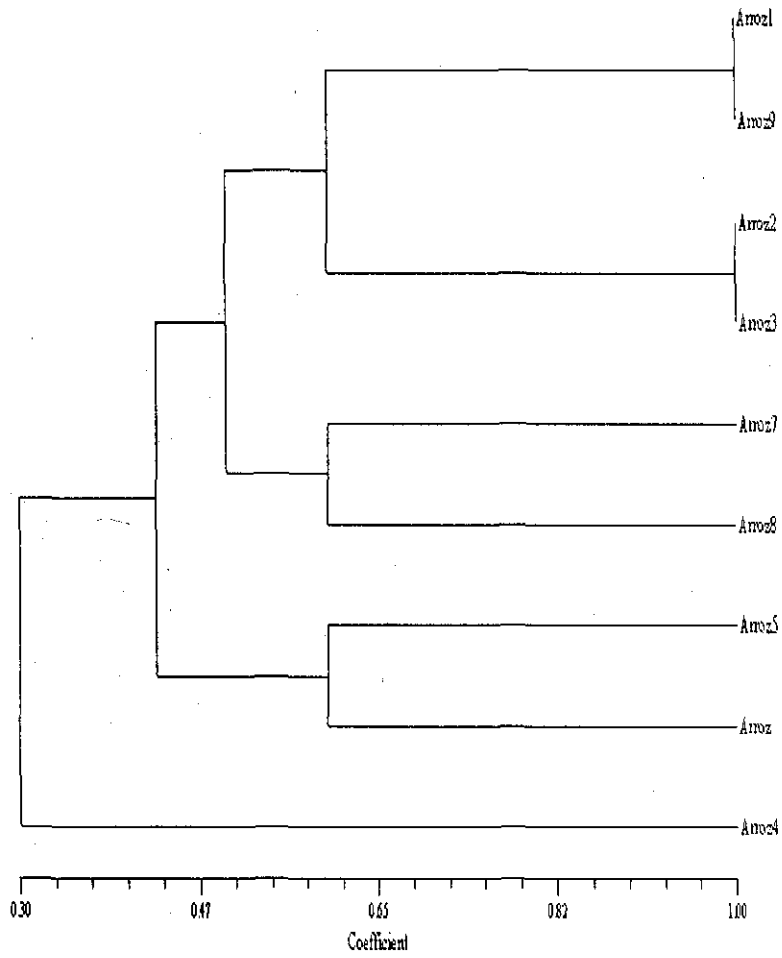
M.C. Duque- CIAT



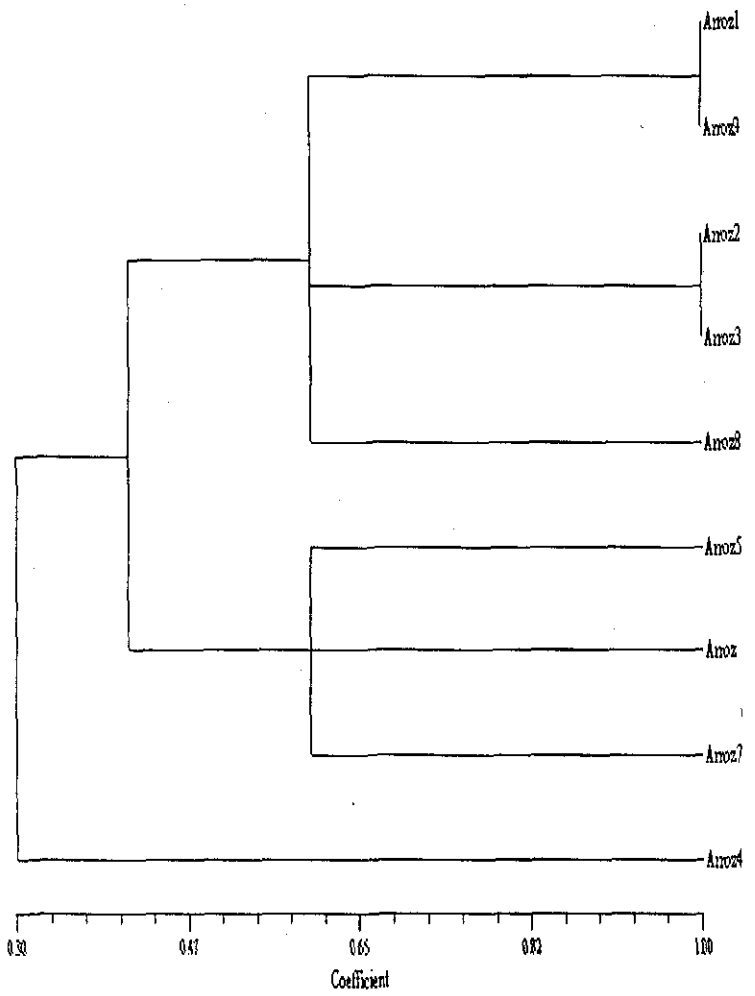


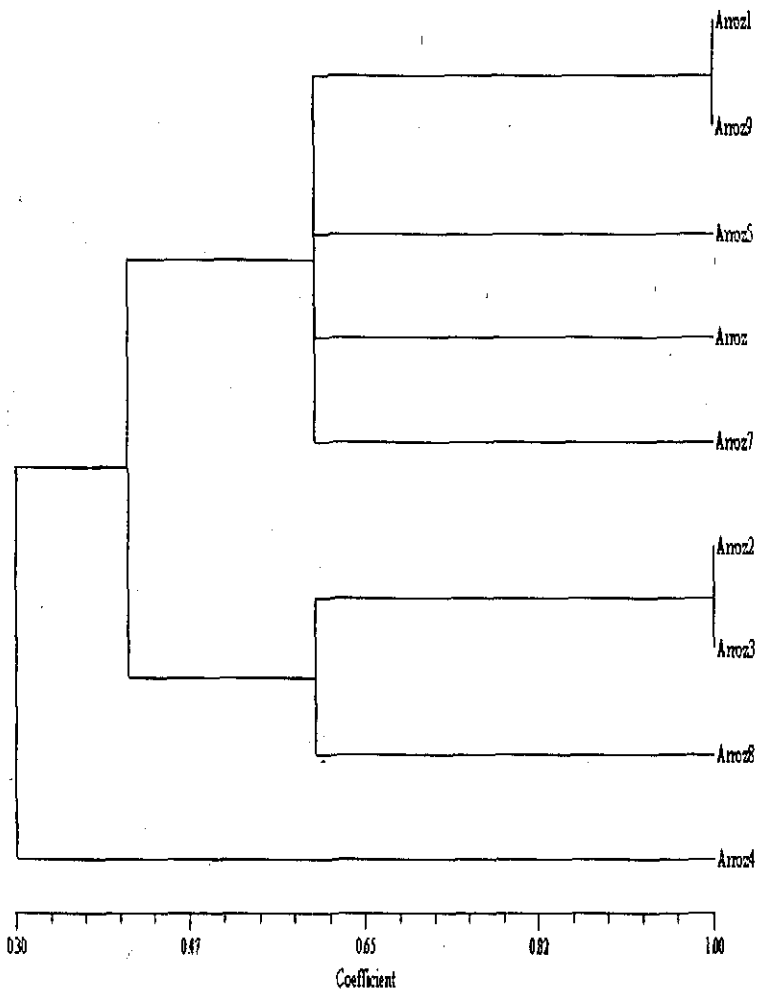
M.C. Duque- CIAT





M.C. Duque- CIAT

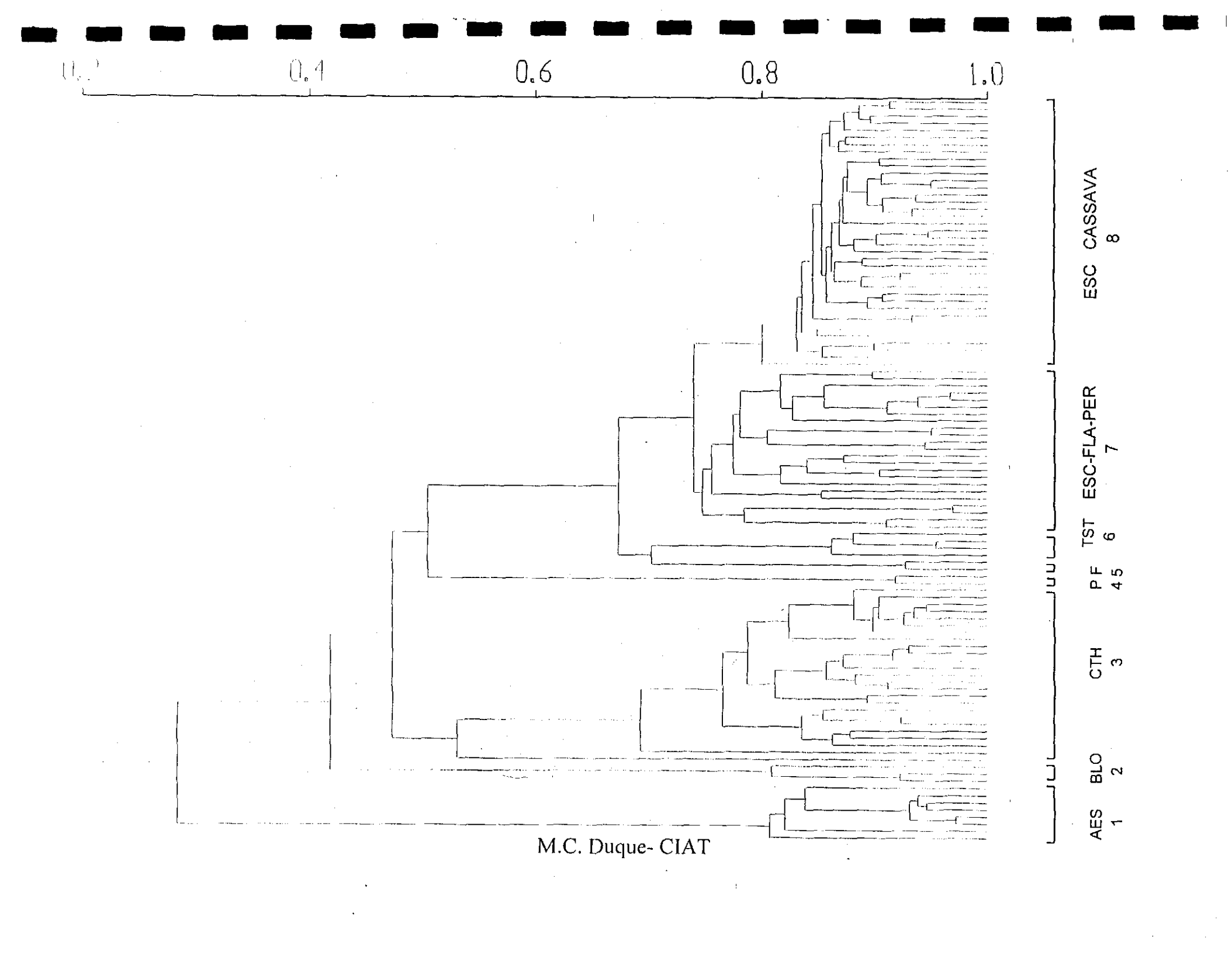




M.C. Duque- CIAT

Ejemplo:

El comportamiento de un grupo de genotipos de arroz frente a diferentes aislamientos de *P. grisea.*



M.C. Duque- CIAT

Ejemplo :

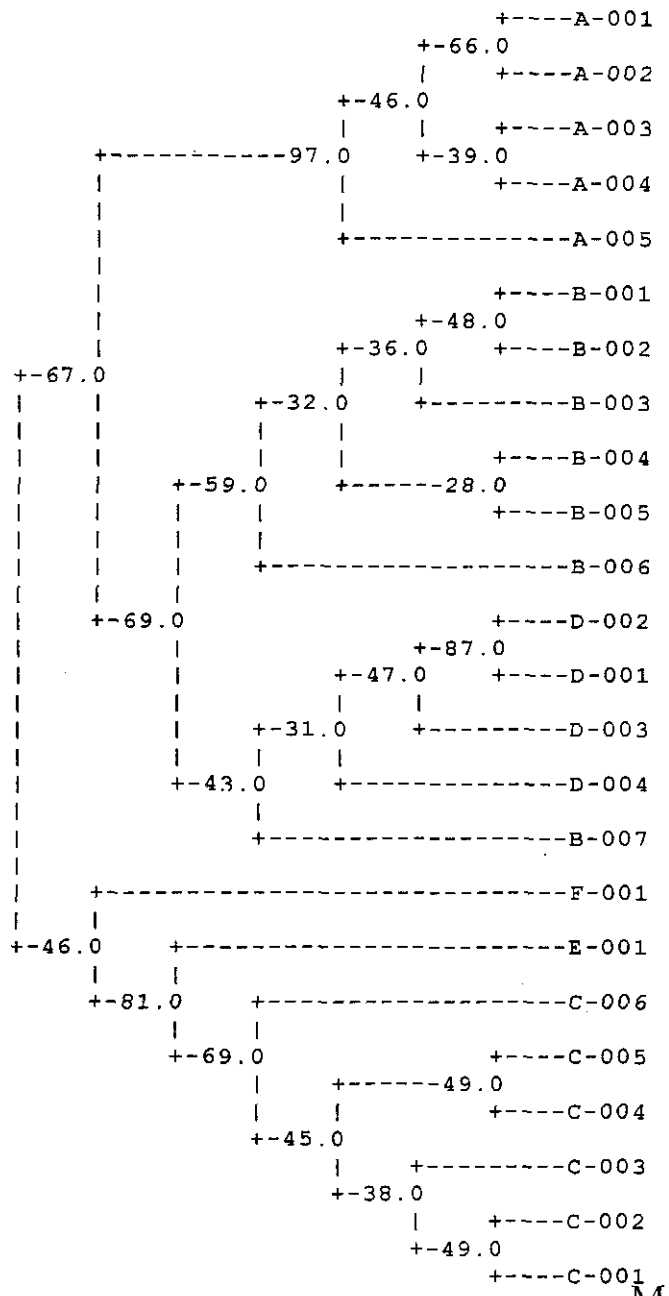
Análisis de las relaciones entre la yuca (*Manihot esculenta* Crantz) y otras especies del género *Manihot* mediante técnicas de AFLP.

(Roa et al. (1997)

Ejemplo

Aplicar el método de ligamiento medio para generar los posibles grupos, a la matriz de Distancias D, sobre datos hipotéticos en *P. vulgaris*

M.C. Duque- CIAT



DENDOGRAMA FINAL :

El número que aparece en el nodo corresponde al porcentaje de veces de la simulación en que los individuos aparecen formando un grupo.

ANALISIS DE CORRESPONDENCIA MULTIPLE

Principios generales

M.C. Duque- CIAT

ANALISIS DE CORRESPONDENCIA

. Representación multidimensional de la dependencia entre filas y columnas de una tabla de contingencia N.

. Se pretende comparar las filas (Homogeneidad).

Definición

Perfil de fila: vector de probabilidad condicional de pertenecer a la columna J dado que está en la fila i.

Fila ---> perfil ---> espacio Euclideano ---> distancias

Distancia = X^2 (métrica Euclideana Ponderada)

$$Distancia^2(k,l) = \sum_{j=1}^{J} \frac{\left(\frac{n_{kj}}{n_{k+}} - \frac{n_{lj}}{n_{l+}}\right)^2}{\left(\frac{n_{+j}}{n}\right)}$$

Matricialmente:

$$d_c(a_k - a_l) = (a_k - a_l)^T D_c^{-1} (a_k - a_l)$$

Ejemplo:

n_{f+}						
I1	1	0	1	1	0	3
I2	1	1	1	1	1	5
I3	0	1	1	1	1	4
I4	0	0	1	0	1	2
I5	0	0	1	1	0	2
n_{+c}	2	2	5	4	3	

$$D^2(1,2) = \frac{\left[\frac{1}{3} - \frac{1}{5}\right]^2}{2} + \frac{\overbrace{\left[\frac{0}{3} - \frac{1}{5}\right]^2}}{2} + \frac{\left[\frac{1}{3} - \frac{1}{5}\right]^2}{5} + \frac{\left[\frac{1}{3} - \frac{1}{5}\right]^2}{4} + \frac{\overbrace{\left[\frac{0}{3} - \frac{1}{5}\right]^2}}{3}$$

$$D^2(1,2) = \frac{\left[\frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}\right]^2}{n_{+1}} + \frac{\left[\frac{n_{12}}{n_{1+}} - \frac{n_{22}}{n_{2+}}\right]^2}{n_{+2}} + \ldots + \frac{\left[\frac{n_{15}}{n_{1+}} - \frac{n_{25}}{n_{2+}}\right]^2}{n_{+5}}$$

H_0: Independencia $\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i+} \cdot n_{+j}}{n} \right)^2}{\frac{(n_{i+} \cdot n_{+j})}{n}}$

Equivalente a : $\chi^2 = n \sum r_i (a_i - c)^T D_c^{-1} (a_i - c)$

Inercia : $\frac{\chi^2}{n}$

Promedio ponderado de las distancias cuadráticas de los perfiles de fila a su centroide.

Independencia resulta equivalente a homogeneidad.

Paso siguiente : Buscar una explicación en un espacio reducido de la falta de homogeneidad en los perfiles de filas (No independencia)

Equivalencia : Encontrar las componentes principales de los puntos pero, teniendo en cuenta la ponderación.

El mejor ajuste : Vector propio de:

$$\left(A - IC^T\right)^T D_R \left(A - IC^T\right) D_C^{-1}$$

Nota: Se procede similarmente con las columnas y se logra la misma solución.

No hay supuestos sobre distribuciones ni modelos.

La generalizción a más de 2 variables recibe el nombre de

ANALISIS DE CORRESPONDENCIA MULTIPLE

parte de la matriz Z'Z donde Z es la matriz de indicadores binarios.

Situación Hipotética

- Un grupo de investigadores haciendo una práctica de extracción de DNA en un auditorio

- Un conjunto de implementos para que el grupo trabaje

Problema

Describir el proceso que se cumple en el laboratorio

- Hay grupos de personas realizando tareas específicas? (O todas hacen de todo)

- Hay grupos de implementos asociados con esas tareas? (o los instrumentos no son específicos)

Estrategia de análisis

Busco un "mejor sitio para ver".

Me alejo del escenario, me corro a un lado, y subo algunos escalones.

Desde allí,

sin cambiar lo que está ocurriendo,

puede ser que tenga un mejor enfoque (mayor claridad), de lo que ocurre sobre la mesa de trabajo. (También es posible que no lo logre).

- Debo evaluar la calidad del nuevo sitio y ver
- que desplazamiento hice para llegar a él.

- En términos de los hechos más relevantes puedo tratar de responder las preguntas que tengo.

Situación real :

Un conjunto de genotipos del género *Manihot* caracterizado por la presencia o ausencia de bandas.

Genotipos : investigadores
Bandas : implementos

Matriz de datos: Matriz binaria

	banda 1	banda 2	...	banda n
Gen 1	1	0	...	0
Gen 2	0	1	...	1
..	.	.		.
	.	.		.
..	.	.		.
	.	.		.
Gen m	1	1	...	1

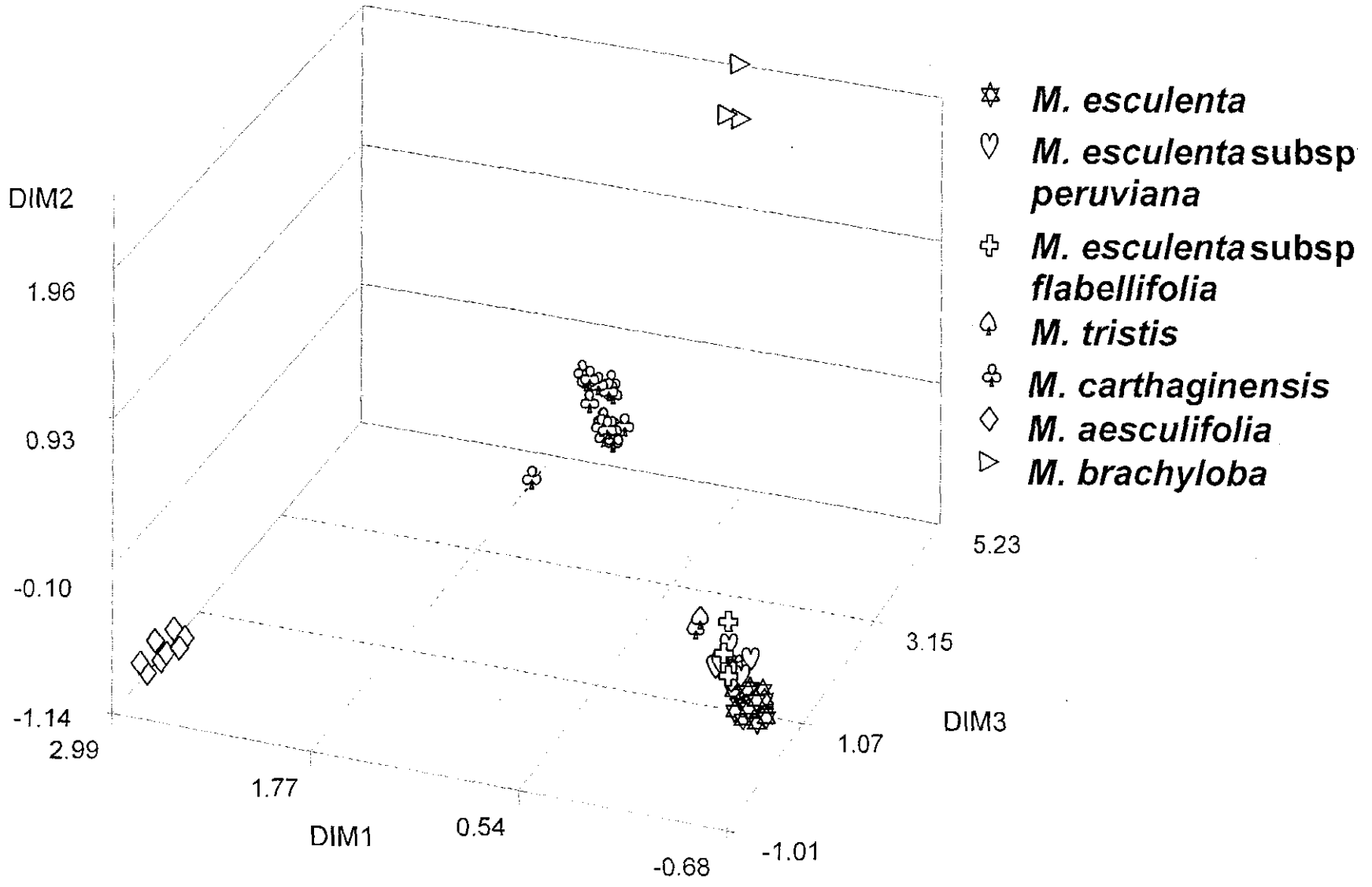
Indicadores importantes :

Contribuciones parciales:

Importancia de las variables en términos de su contribución a la componente principal
(Hecho relevante).

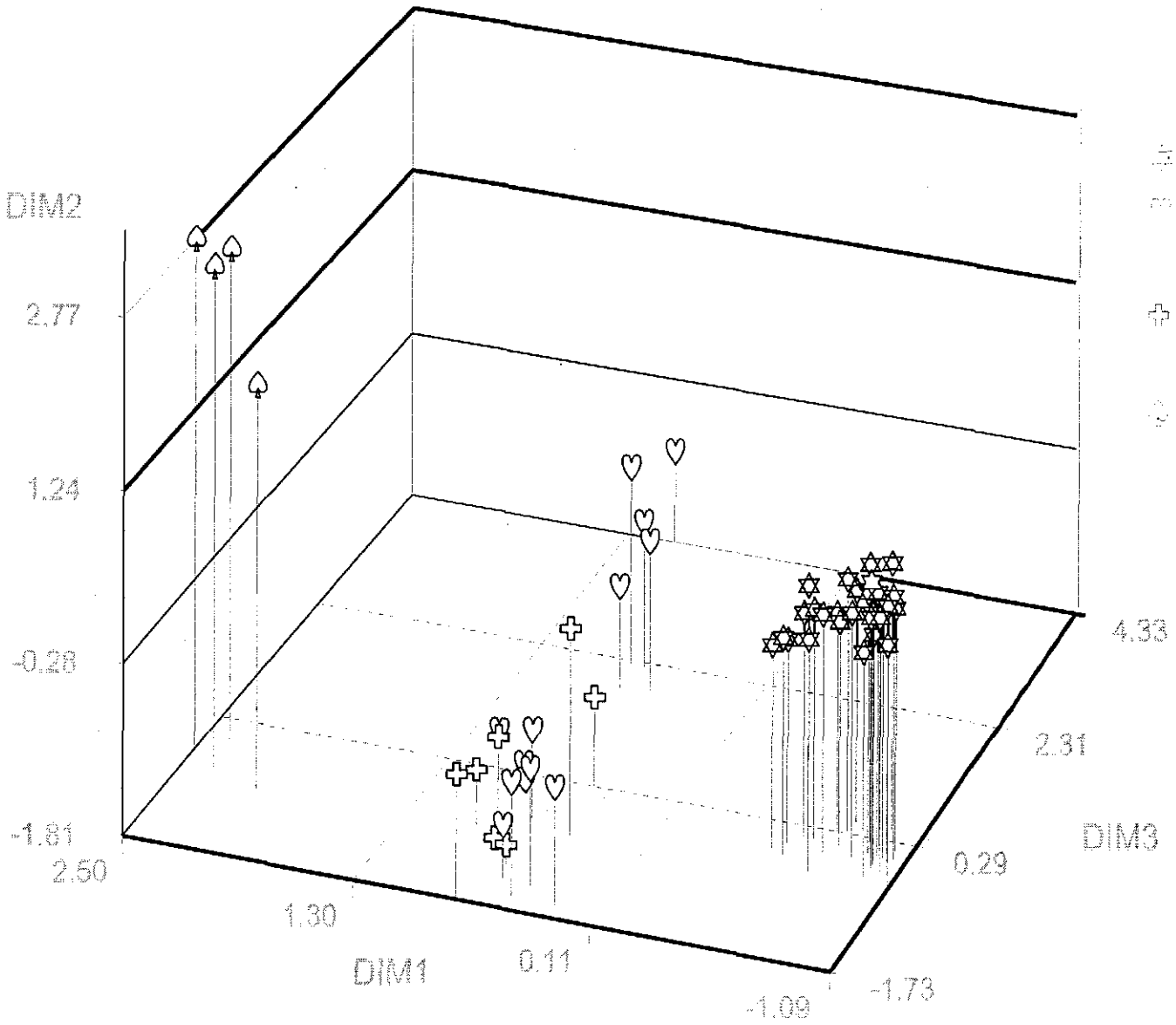
Cosenos cuadrados:

Determina en cuales componentes la variable tiene mayor influencia (Impacto).



Tridimentional representation of principal axes of variation from MCA of AFLP in 7 *Manihot* species.

M.C. Duque- CIAT



- *M. esculenta*
- *M. esculenta* subsp *peruviana*
- *M. esculenta* subsp *flabellifolia*
- *M. tristis*

Tridimentional representation of principal axes of variation from MCA of AFLP in 4 *Manihot* species

M.C. Duque- CIAT

Valores de similaridad de Nei-Li entre y *dentro* de grupos de especies de *Manihot*

GRUPO	ESC	ASC	BLO	CTH	EF-PER	TST
ESC	*0.85*	0.31	0.41	0.47	0.70	0.65
ASC		*0.86*	0.26	0.26	0.27	0.27
BLO			*0.84*	0.43	0.40	0.41
CTH				*0.77*	0.46	0.45
EF-PER					*0.70*	0.67
TST						*0.88*

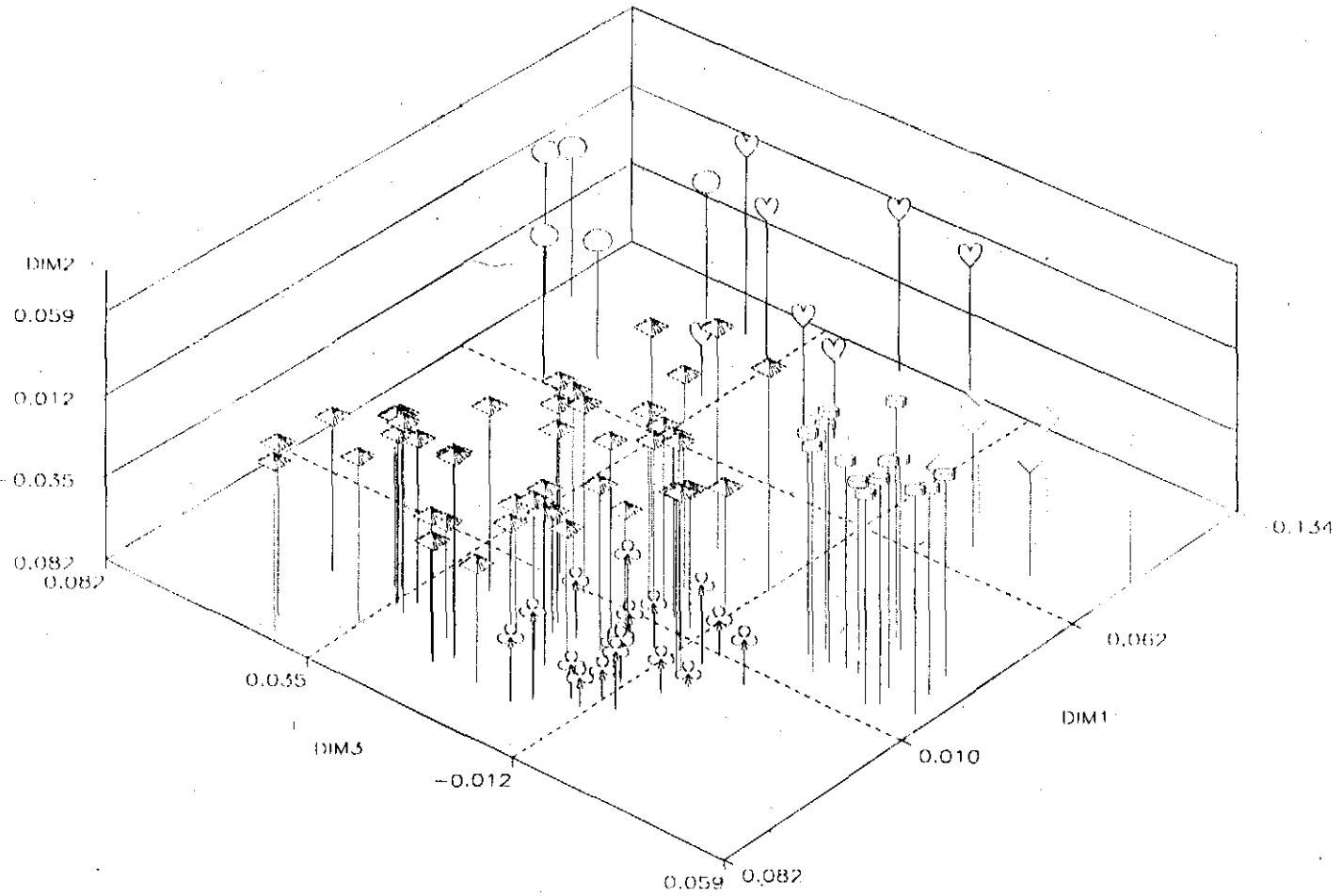
Un programa SAS para análisis de Correspondencia y Similaridad con datos binarios

Juan B. Cuasquer

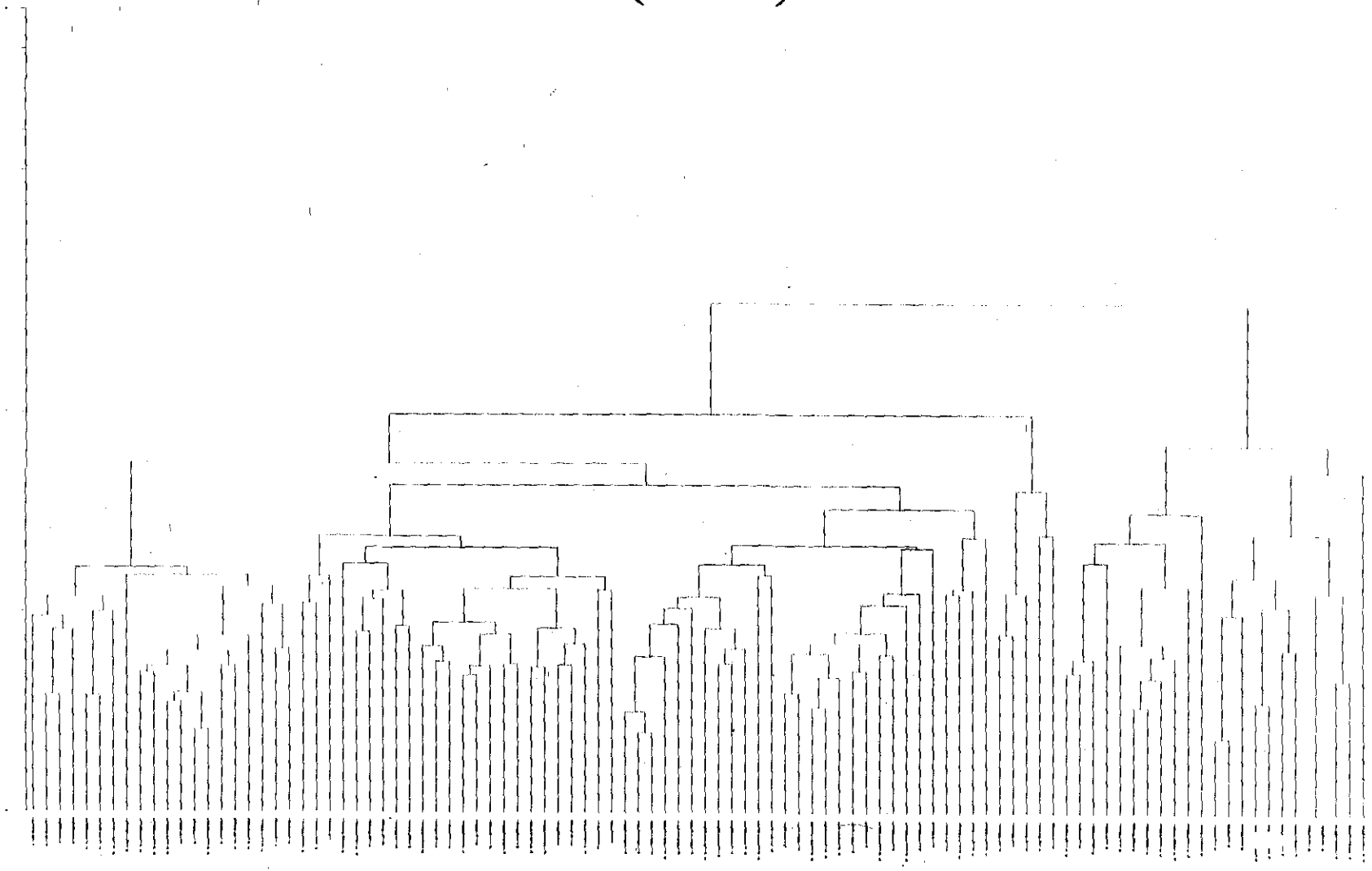
Componenetes del programa

- Sección 1: Entrada de datos (lectura de la matriz)
- Sección 2: Análisis de correspondencia (AC)
- Sección 3: Cluster sobre archivo de salida del AC
- Sección 4: Gráficas de AC
- Seccíon 5: Análisis de Similaridad
- Sección 6: Cluster sobre datos de similaridad
- Sección 7: Gráficas (Dendrogramas) con datos de similaridad
- Sección 8: Similaridad promedio entre grupos y dentro de grupos con datos del AC

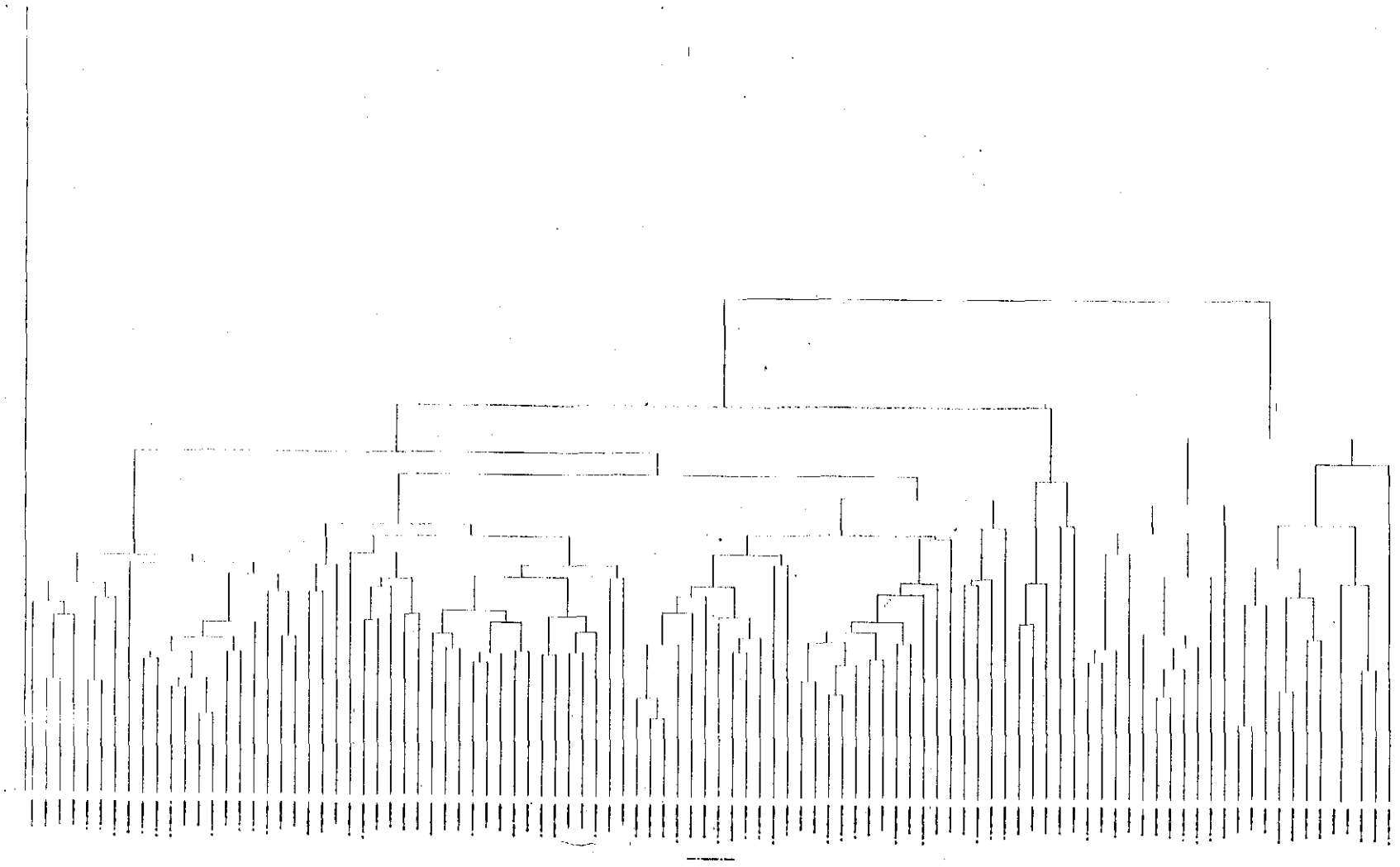
Grafica 3d Analisis de Correspondencia (SAS)



Dendrograma de Análisis de Correspondencia (SAS)



Dendrograma de similaridad (SAS)



DIVERSIDAD

M.C. Duque- CIAT

MEDIDAS DE VARIACION GENETICA

Frecuencias Génicas son las bases para medir la variación genética → **Evolución** de una población es principalmente el resultado de **cambios en las frecuencias** génicas por mutación, selección, migración y deriva.

Las frecuencias génicas se prefieren a las genotípicas debido a:

❖La frecuencias génicas permanecen **relativamente estables** en el tiempo y son **relativamente independientes** del sistema reproductivo, mientras que las genotípicas se mezclan después de cada generación reproductiva.

Nei (73), introdujo el concepto de **Diversidad Génica** (algo medible) para describir la **Variación Genética** (algo conceptual), válida para poblaciones sexuales y asexuales.

H: Diversidad Génica

Probabilidad de obtener 2 alelos diferentes en un locus cuando 2 haploides se toman de una población.

Probabilidad de que 2 individuios tomados al azar sean diferentes en un locus.

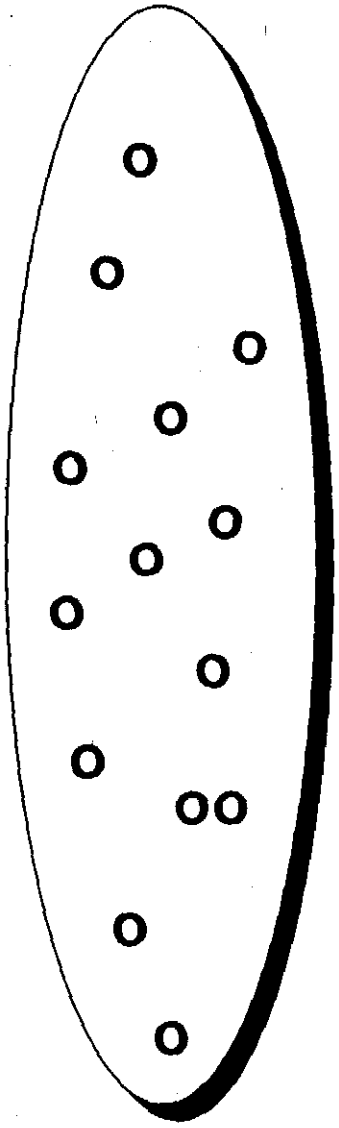
$$H = 1 - \sum_{i=1}^{k} p_i^2, \quad H \in [0,1]$$

p_i = **Frecuencia del alelo *i* en el locus**

En una población diploide, la diversidad génica, así definida, es equivalente a la Heterozigocidad esperada bajo reproducción aleatoria.

NOTA: **La definición anterior puede ampliarse a Diversidad Haplotípica o genotípica!!!**

Población total



Tomemos un locus....
Alelos : 1,2 ... n

x_1 = frecuencia alelo 1
x_2 = frecuencia alelo 2
.
.
x_n = frecuencia alelo n

Cuál es la probabilidad de que al elegir aleatoriamente dos individuos resulten idénticos en ese locus ?

Probabilidad de identidad (J) : 1 locus

- Iguales en el alelo 1 o iguales en el alelo 2 o ... iguales en el alelo n.

$$P\,(\,identidad\quad) = x_1^2 + x_2^2 + \ldots + x_n^2$$

$$P\,(\,identidad\quad) = \sum_{i=1}^{n} x_i^2$$

$$P\,(\,identidad\quad) = J$$

$$P\,(\,diferencia\quad) = H = 1 - J$$

H: DIVERSIDAD GENETICA

Ahora … hay más de 1 locus : 1, 2,…k loci

x_{11} = frec. alelo 1, locus 1 x_{21} = frec. alelo 2, locus 1 . . $x_{n\,1}$= frec. alelo n, locus 1	…	x_{1k} = frec. alelo 1, locus k x_{2k} = frec. alelo 2, locus k . . $x_{n\,k}$= frec. alelo n, locus k

Cuál es la probabilidad de que al elegir aleatoriamente dos individuos resulten idénticos ?

(recuerde que ahora hay varios loci)

$$locus \quad 1 : h_1 = 1 - \sum_{i=1}^{n_1} x_i^2$$ Heterogeneidad en el locus 1, que tiene n_1 alelos.

$$locus \quad 2 : h_2 = 1 - \sum_{i=1}^{n_2} x_i^2$$ Heterogeneidad en el locus 2, que tiene n_2 alelos.

...

...

$$locus \quad k : h_k = 1 - \sum_{i=1}^{n_k} x_i^2$$ Heterogeneidad en el locus k, que tiene n_k alelos.

$$H_t = \frac{h_1 + h_2 + ... + h_k}{k} \longleftrightarrow H_t = 1 - \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_i} x_{ij}^2}{k}$$

H_t: heterogeneidad media por locus

Nota:

en poblaciones diploides con alelos codominantes cada valor de h_i tiene un sesgo de estimación que debe ser corregido mediante el uso de un factor adecuado:

- En poblaciones de polinización abierta el factor es:

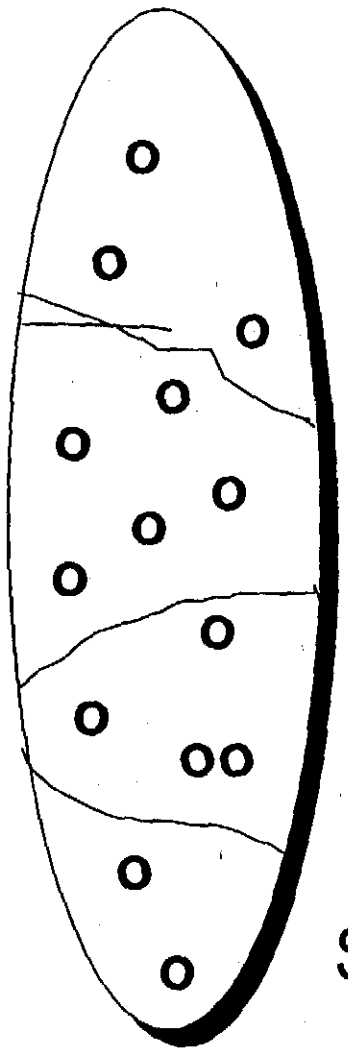
$$f = \frac{2n}{n-1}$$

-En poblaciones autofecundadas el factor es:

$$f = \frac{n}{n-1}$$

Estos factores se recomiendan para n<= 50 individuos

Población Subdividida (Partición)



S_1 , N_1 ⇒ Diversidad genética en S_1

S_2, N_2 ⇒ Diversidad genética en S_2

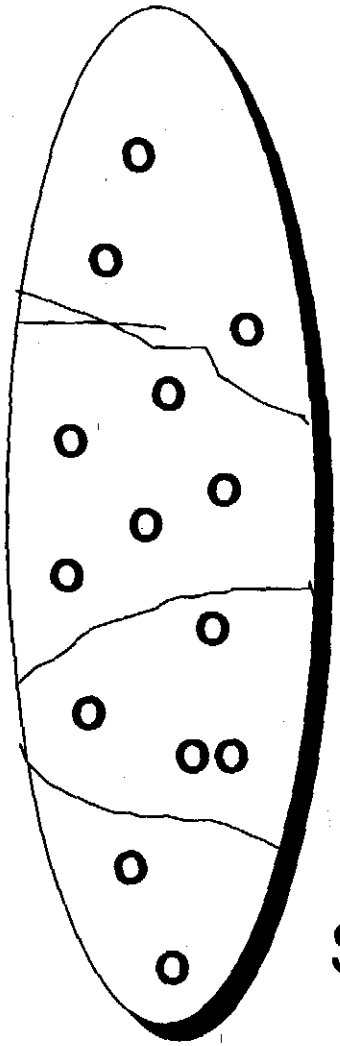
S_3, N_3 ⇒ Diversidad genética en S_3

S_4, N_4 ⇒ Diversidad genética en S_4

J_t , H_t

H_s : dentro de subpoblaciones

Población Subdividida (Partición)



S_1, N_1 ⇒ $H_1 * N_1 / N$ (En S_1)

S_2, N_2 ⇒ $H_2 * N_2 / N$ (En S_2)

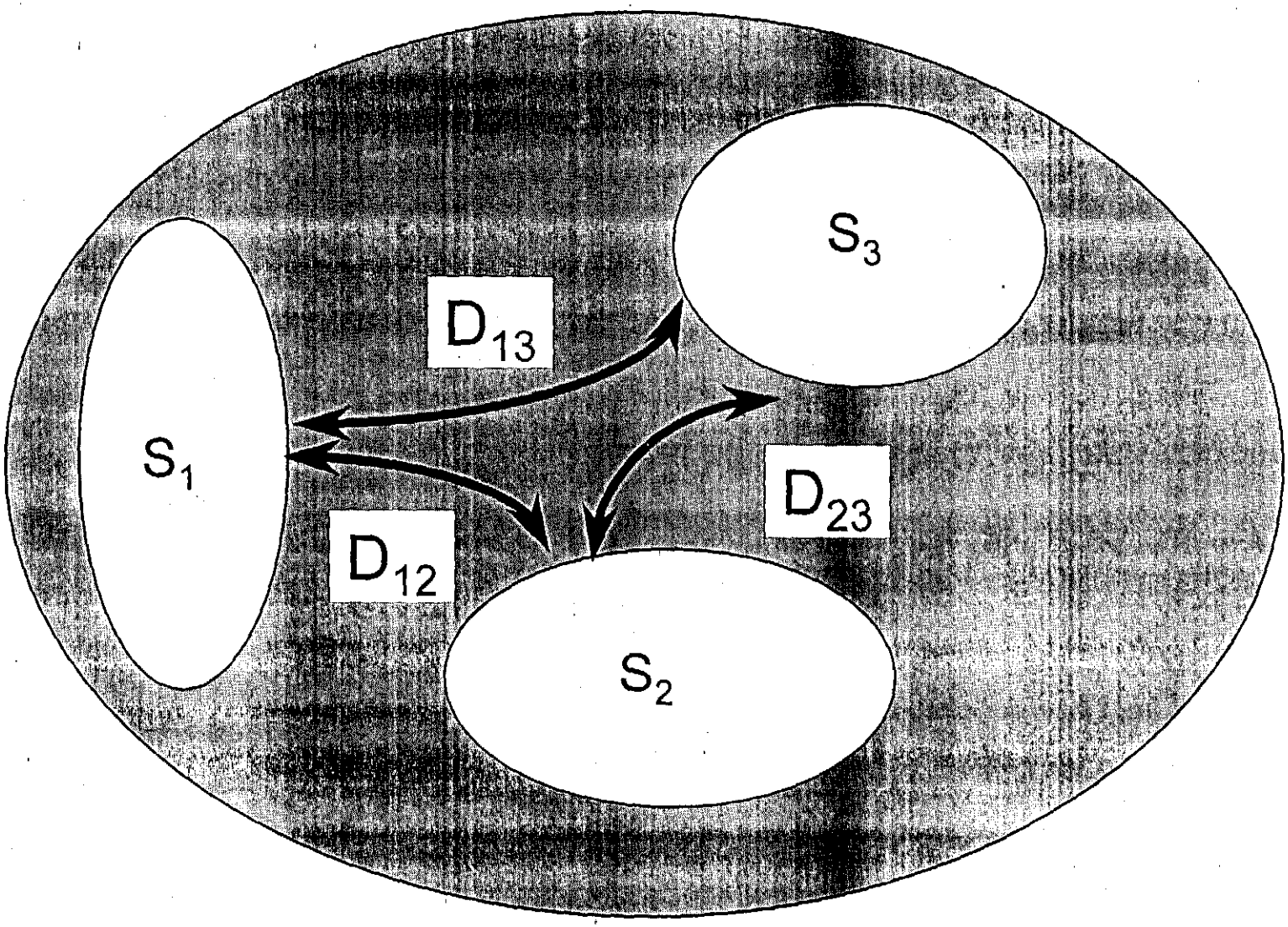
S_3, N_3 ⇒ $H_3 * N_3 / N$ (En S_3)

S_4, N_4 ⇒ $H_4 * N_4 / N$ (En S_4)

J_t , H_t

H_S **Dentro de subpoblaciones**

DIVERSIDAD GENÉTICA ENTRE SUBPOBLACIONES



M.C. Duque- CIAT

Diversidad Genética Total =

Diversidad Genética Dentro De Subpoblaciones

+

Diversidad Genética Entre Subpoblaciones

$$H_t = H_s + D_{st}$$

Relación que permite calcular el coeficiente de DIFERENCIACIÓN GENETICA (G_{st}):

$$G_{st} = D_{st} / H_t$$

PARTICION DE LA VARIACIÓN GENETICA

Gst: Describe la cantidad de variación atribuida a una subdivisión, con relación a la variación genética total

Gst no depende de la historia poblacional.

$$G_{st} \in [0,1]$$

MÉTODO NO PARAMÉTRICO PARA PROBAR DIFERENCIACIÓN GENETICA.

Frecuencia esperada : $P_{ij} = P_i \times P_j$

Para las diferentes subpoblaciones calcular la frecuencia esperada según la definición anterior, y hacer una prueba de χ^2 de bondad de ajuste.

La diferencia entre la frecuencia observada y la esperada es una medida del desequilibrio de ligamiento o del desequilibrio gamético y se usa como una medida de intercambio genético y de la recombinación.

NOTA:

Debido a que la estructura poblacional frecuentemente es determinada por la partición de la variación genética, se tiene que la inferencia realizada es altamente dependiente de la diversidad genética ....es decir del marcador.

Las comparaciones deben hacerse sobre datos con herramientas genéticas iguales y dicho marcador debe haber sido elegido dependiendo de los objetivos experimentales específicos.

Valores de Diversidad dentro y entre grupos calculados según expresiones de Nei en una colección de *Manihot.*

GRUPO	N	H_i	Hs
ESC	38	0.077	0.093
ASC	8	0.069	
BLO	3	0.091	
CTH	25	0.099	
EF-PER	27	0.123	
TST	4	0.052	
TOTAL=H_t	105	0.198	0.198

G_{st} = 53.08 %

Ejemplo :

Estudio de la base genética de arroz por medio de microsatélites

(Gallego G. et al. En preparación)

Base genetica de Arroz- Microsatelites

OBS	ID ENT	ALELO 5	MS 5	ALELO 6	MS 6	ALELO 1	MS 1	ALELO 16	MS 16
1	?-CO39	1	1000000	3	00100000	4	00010000000	1	100000
2	I-Amistad82	1	1000000	2	01000000	3	00100000000	5	000010
3	I-Anayansi	1	1000000	3	00100000	1	10000000000	3	001000
4	I-Araure4	1	1000000	3	00100000	2	01000000000	3	001000
5	I-C46-15	1	1000000	2	01000000	3	00100000000	3	001000
6	I-CentaA-1	1	1000000	3	00100000	1	10000000000	3	001000
7	I-Cica8	2	0100000	3	00100000	1	10000000000	3	001000
8	I-DGWG	1	1000000	3	00100000	3	00100000000	1	100000
9	I-Huarangopampa	1	1000000	3	00100000	8	00000001000	3	001000
10	I-IAC1278	1	1000000	3	00100000	1	10000000000	3	001000
11	I-ICTAPolochi	1	1000000	3	00100000	2	01000000000	3	001000
12	I-ICTAQuirigua	3	0010000	1	10000000	8	00000001000	3	001000
13	I-ICTAVirginia	2	0100000	3	00100000	1	10000000000	3	001000
14	I-IR84-63-5-18	2	0100000	2	01000000	1	10000000000	3	001000

OBS	IDENT	ALELO5	MS5	ALELO6	MS6	ALELO11	MS11	ALELO16	MS16
15	I-KhaoDawkMali105	3	0010000	6	00000100	8	00000001000	5	000010
16	I-Metical	3	0010000	3	00100000	1	10000000000	3	001000
17	I-P3055F4-3-4P-1P-1BWC-106	2	0100000	2	01000000	4	00010000000	3	001000
18	I-PalizadaA-86	1	1000000	3	00100000	6	00000100000	3	001000
19	I-Saavedra	3	0010000	1	10000000	2	01000000000	2	010000
20	I-Tanioka	6	0000010	1	10000000	3	00100000000	3	001000
21	J-ElPasoL-94	1	1000000	1	10000000	2	01000000000	1	100000
22	J-Fanny	1	1000000	1	10000000	8	00000001000	2	010000
23	J-Irat146Acc406	3	0010000	1	10000000	9	00000000100	1	100000
24	J-MarogParocRexoro	5	0000100	1	10000000	5	00001000000	1	100000
25	J-Monolaya	3	0010000	1	10000000	2	01000000000	1	100000
26	J-Oryzica-Sabana6	3	0010000	1	10000000	5	00001000000	1	100000
27	J-Tox1859-102-GM-3WC5036	3	0010000	3	00100000	1	10000000000	3	001000
28	J-Tox340-1-7-3(ITA133)-Ace4	4	0001000	1	10000000	5	00001000000	1	100000
29	S-Ciwini	2	0100000	1	10000000	3	00100000000	1	100000
30	S-P5589-1-1-3P-4-MPPatselrec	4	0001000	1	10000000	4	00010000000	6	000001

MS5	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0000010	1	3.3	1	3.3
0000100	1	3.3	2	6.7
0001000	2	6.7	4	13.3
0010000	8	26.7	12	40.0
0100000	5	16.7	17	56.7
1000000	13	43.3	30	100.0

MS6	Frequency	Percent	Cumulative Frequency	Cumulative Percent
00000100	1	3.3	1	3.3
00100000	13	43.3	14	46.7
01000000	4	13.3	18	60.0
10000000	12	40.0	30	100.0

Es como el diagrama anteror PERO no se restringe a una rama.

Cada vez mira el árbol completo y cuenta la frecuencia con que ocurre la clasificación inicial

SOFTWARE:

WinBoot:

A program for performing bootstrap analysis of binary data to determine the confidence limits of upgma-based dendrograms

Immanuel V. Yap and Rebecca J. Nelson

IRRI

Formatos para entrada de datos: Se sugiere que la extension del archivo sea ".dat"

a. Formato Phylip

linea 1: numero de individuos y nújmero de bandas (separados por espacio al menos)

linea 2 en adelante: datos para individuos(por filas).

Inicia la identificación del individuo(columnas 1 a 10 : caracteres sin blancos) seguida por un espacio en blanco en la columna 11 y por los datos, a partir de la columna 12.

Entre los datos no son necesarios espacios en blanco, pero pueden existir para facilitar correcciones.

Los datos de un individuo pueden ubicarse en varias líneas consecutivas.

Un nuevo individuo, debe iniciar una fila.

b. Delimitado por tabs.

Sigue las mismas normas.

Species in order:

A-001,..., F-001 (24 OTU's)

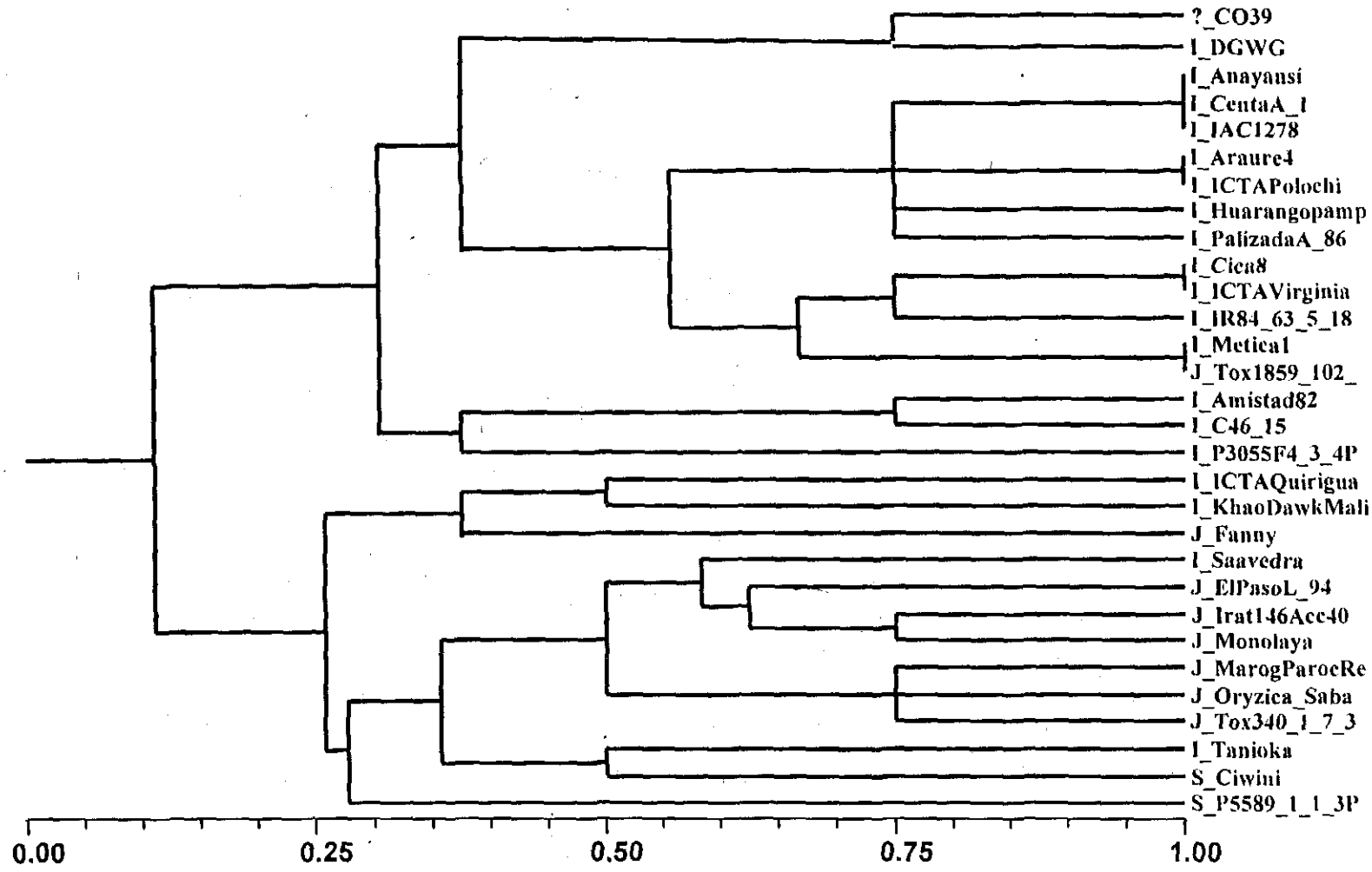
Sets included in the consensus tree:

Set (species in order)	How many times out of 100
*****.....	97
....**	87
....******.. ..*.	81
.....***** **......** **..	69
....******..	69
********** **......** **..	67
...	
********** **........ ***.	1
********** **........ .*..	1
********** ********.. ..*.	1

MS11	Frequency	Percent	Cumulative Frequency	Cumulative Percent
00000000100	1	3.3	1	3.3
00000001000	4	13.3	5	16.7
00000100000	1	3.3	6	20.0
00001000000	3	10.0	9	30.0
00010000000	3	10.0	12	40.0
00100000000	5	16.7	17	56.7
01000000000	5	16.7	22	73.3
10000000000	8	26.7	30	100.0

MS168	Frequency	Percent	Cumulative Frequency	Cumulative Percent
000001	1	3.3	1	3.3
000010	2	6.7	3	10.0
001000	16	53.3	19	63.3
010000	2	6.7	21	70.0
100000	9	30.0	30	100.0

Base Genética de Arroz Microsatélites



Coeficiente de Dice- Similaridad- Método UPGMA

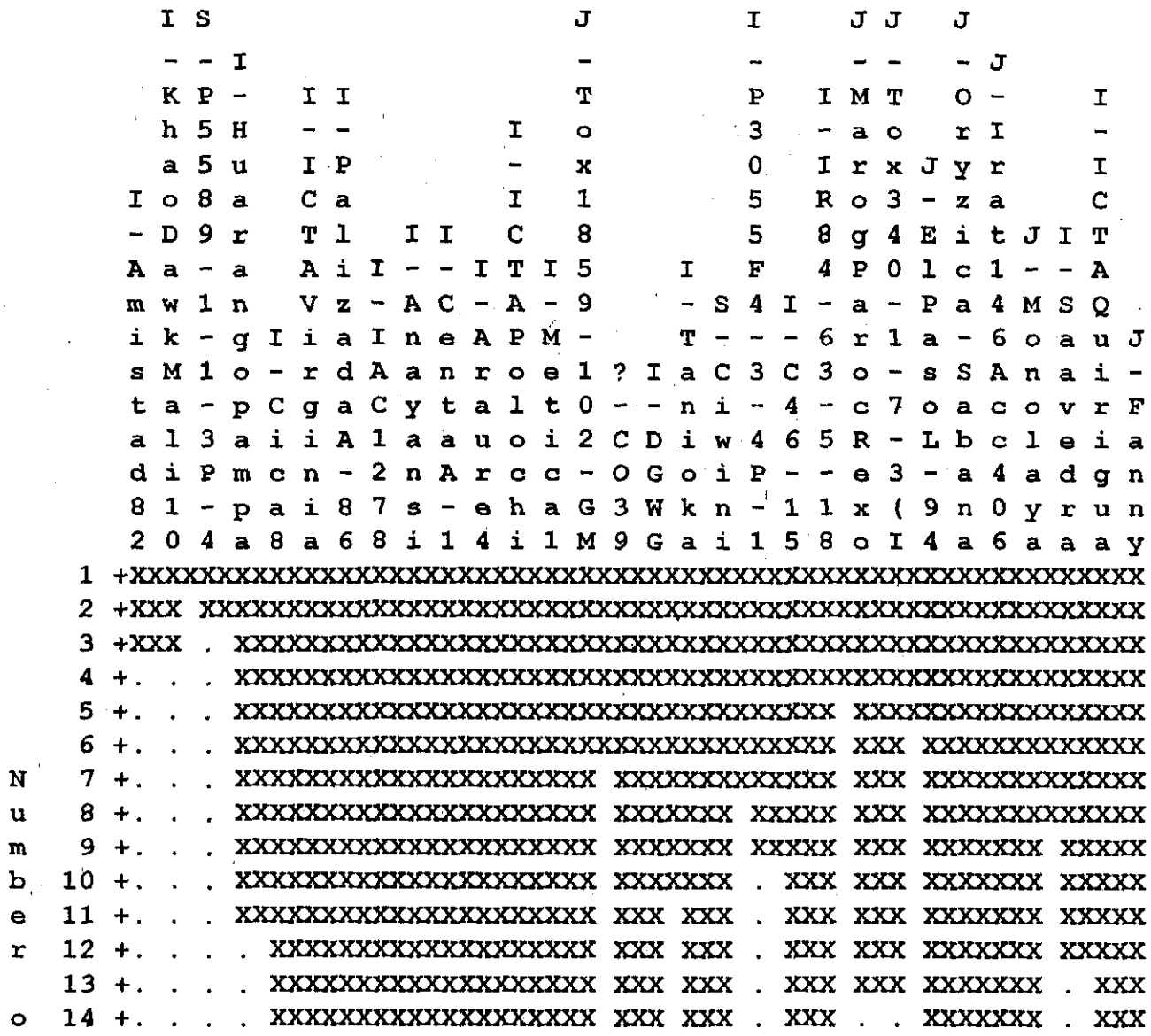
M.C. Duque- CIAT

ANALISIS DE CORRESPONDENCIA MULTIPLE

M.C. Duque- CIAT

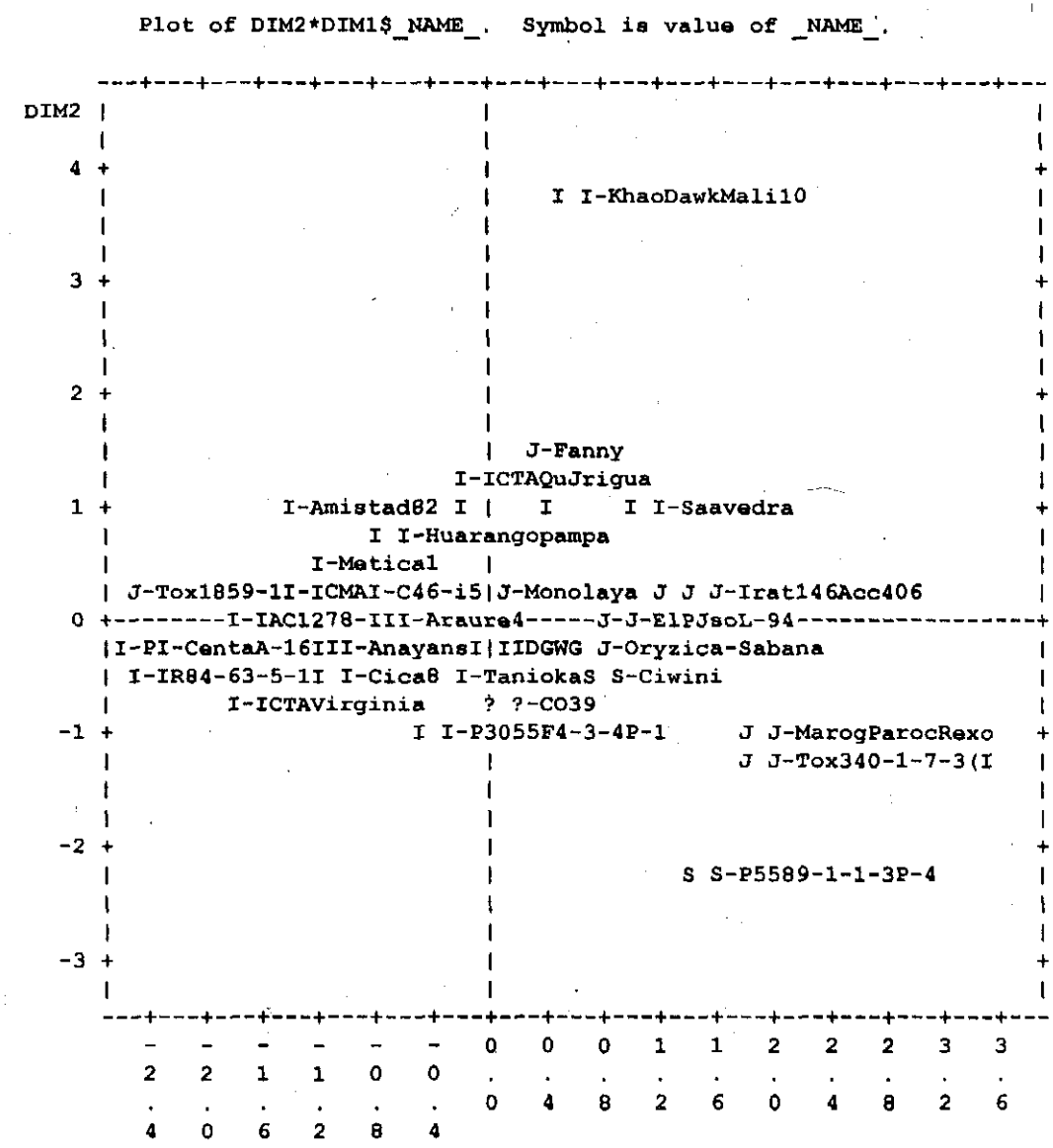
Average Linkage Cluster Analysis

NCL	Clusters Joined		FREQ	RMS STD	SPRSQ	RSQ	Norm RMS Dist	Tie
29	I-Anayansi	I-CentaA-1	2	0.0000	0.00000	1.000	0.00000	T
28	CL29	I-IAC1278	3	0.0000	0.00000	1.000	0.00000	T
	...							
13	J-MarogParocRexo	J-Tox340-1-7-3(I	2	0.3845	0.00510	0.962	0.38451	
12	CL22	I-Saavedra	3	0.3243	0.00680	0.955	0.38887	
11	CL14	I-Huarangopampa	11	0.2825	0.00831	0.947	0.40058	
10	CL18	CL16	4	0.3690	0.00969	0.937	0.41516	
9	CL17	I-P3055F4-3-4P-1	3	0.3838	0.00785	0.929	0.43300	
8	CL12	CL15	7	0.4403	0.02645	0.903	0.53103	
7	CL10	CL9	7	0.5455	0.03733	0.866	0.64490	
6	CL7	CL11	18	0.6129	0.13110	0.735	0.77988	
5	CL8	CL13	9	0.6190	0.06050	0.674	0.82703	
4	CL6	CL5	27	0.8120	0.26518	0.409	0.99426	
3	I-Amistad82	I-KhaoDawkMali10	2	1.1378	0.04464	0.364	1.13780	
2	CL4	S-P5589-1-1-3P-4	28	0.8802	0.13020	0.234	1.50840	
1	CL2	CL3	30	1.0000	0.23406	0.000	1.58598	



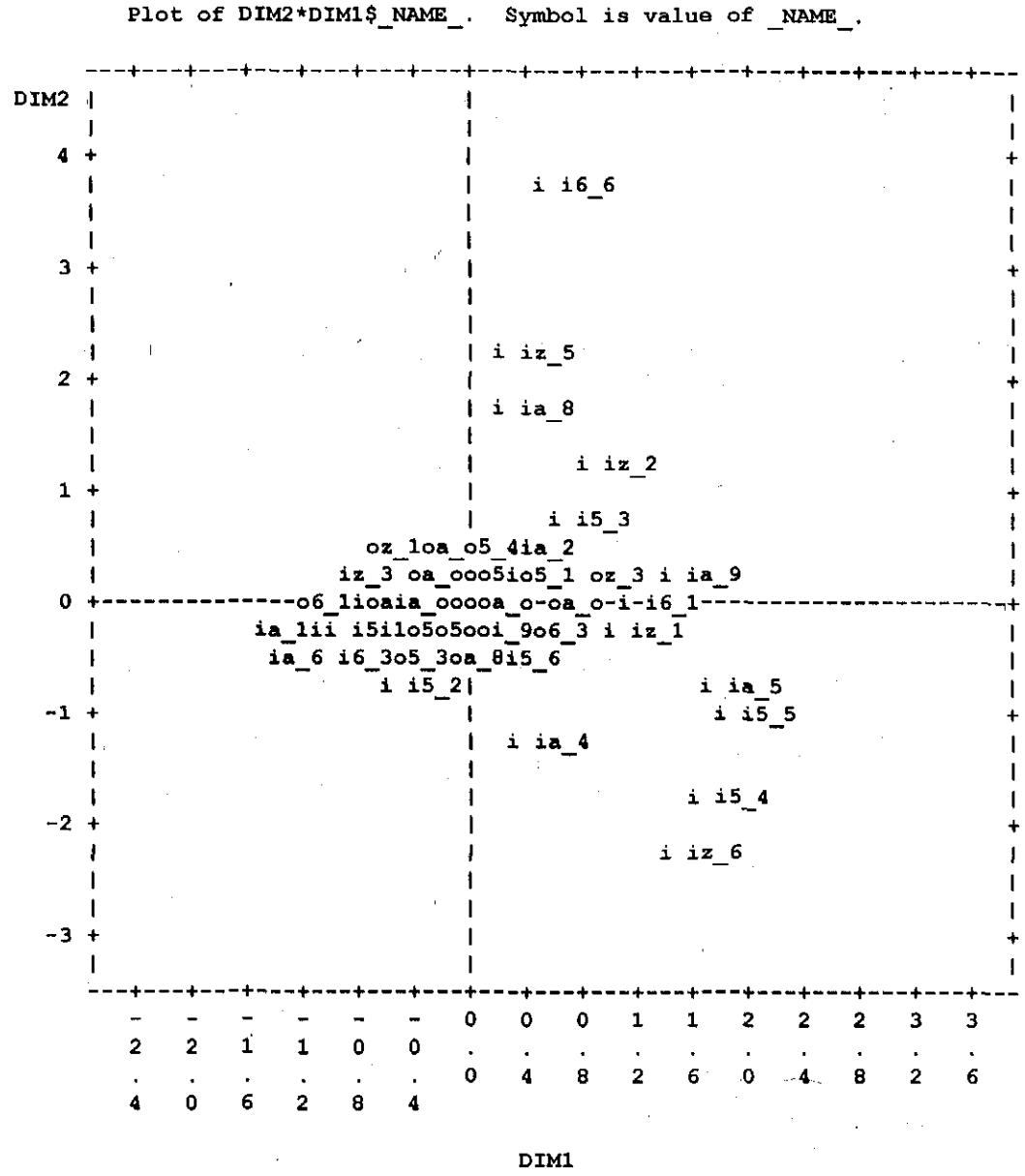
M.C. Duque- CIAT

Microsatelites, Base genetica de Arroz

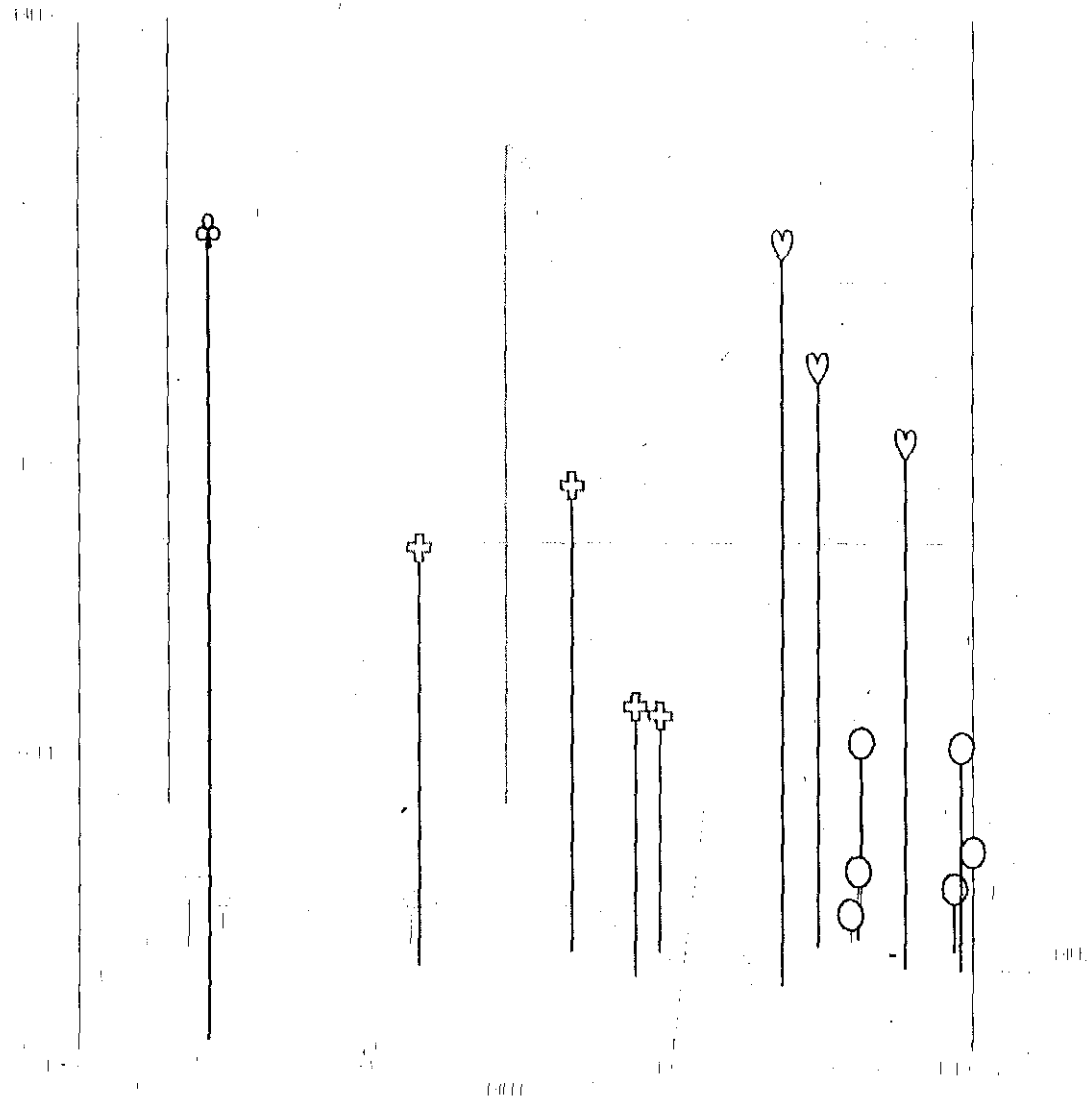


M.C. Duque- CIAT

Microsatelites, Base genetica de Arroz



M.C. Duque- CIAT



♣ S-P5589
▷ I-Khao Dawk Mali

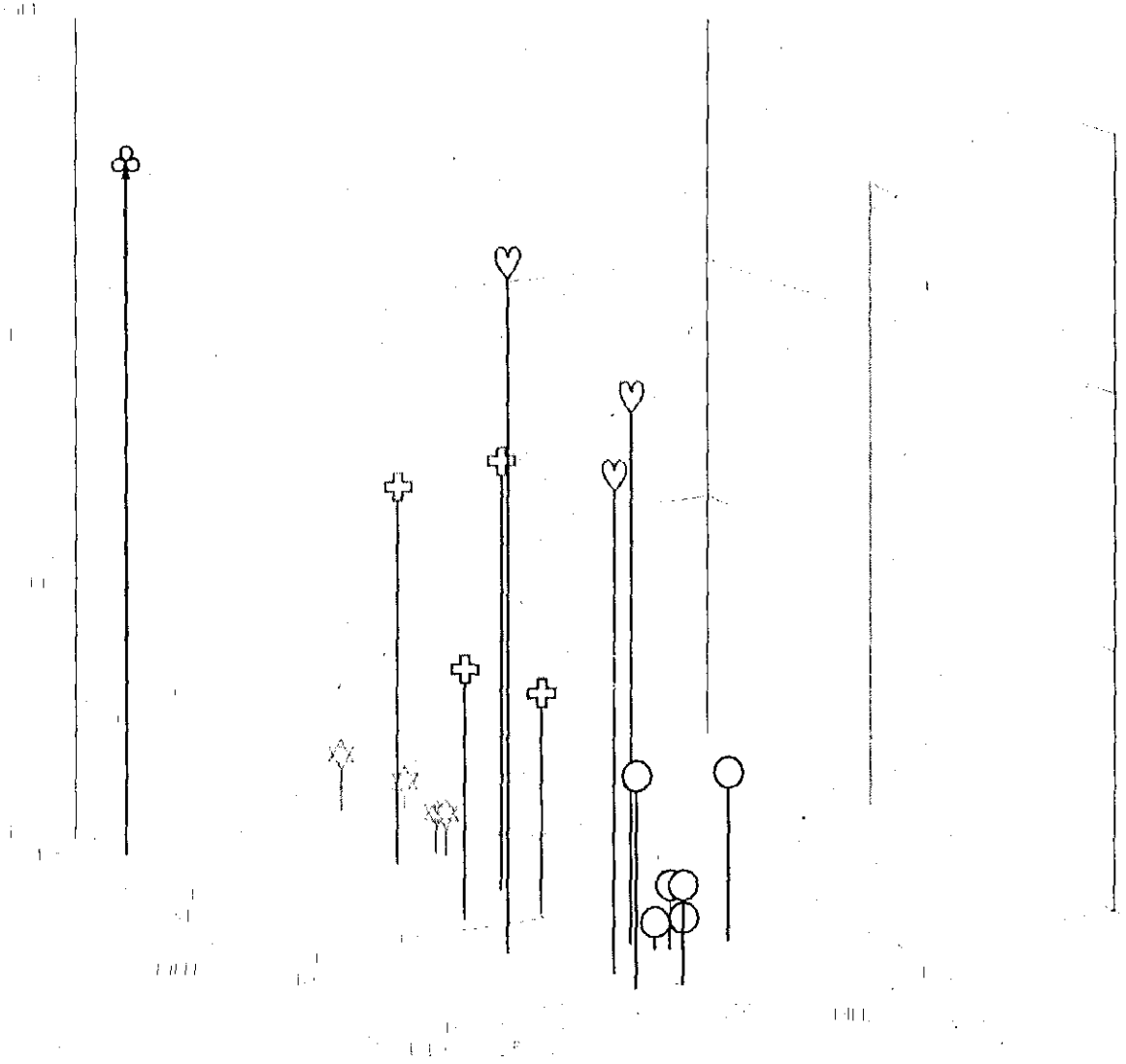
○ I-Huarangopampa, I-Cica8,I-Icta-Virginia, I-Palizada, I-Iac1278, I-Anayansi, I-CentaA I-Araure4, I-Ictapolochi I-Metica1, J-Tox1859

✚ I-DGWG, I-Tanioka ?-CO39, S-Ciwini

♡ I-P3055F4, I-C46, I-IR84 J-Marog Paroc Rex, J-TOX340

o J-El Paso, J-Oryzica-Sabana, J-Irat146, J-Monolaya J-Fanny, I-Saavedra, I-ICTA Quirigua

M.C. Duque- CIAT



♣ S-P5589
▷ I-Khao Dawk Mali

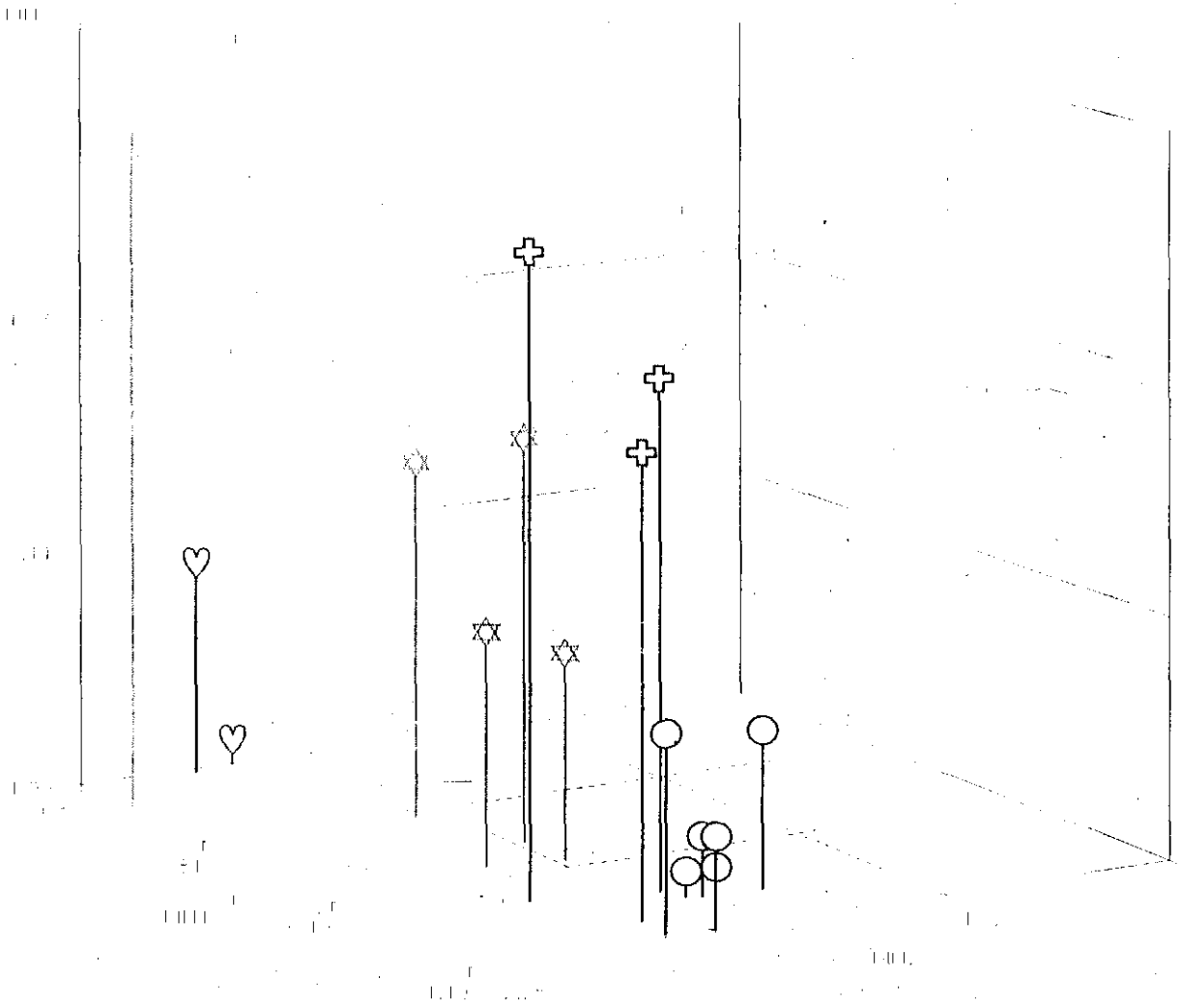
○ I-Huarangopampa, I-Cica8,I-Icta-Virginia, I-Palizada, I-Iac1278, I-Anayansi, I-CentaA I-Araure4, I-Ictapolochi I-Metica1, J-Tox1859

✣ I-DGWG, I-Tanioka ?-CO39, S-Ciwini

♡ I-P3055F4, I-C46, I-IR84 J-Marog Paroc Rex, J-TOX340

o J-El Paso, J-Oryzica-Sabana, J-Irat146, J-Monolaya J-Fanny, I-Saavedra, I-ICTA Quirigua

M.C. Duque- CIAT



S-P5589

I-Amistad 82

I-Huarangopampa, I-Cica8,I-Icta-Virginia, I-Palizada, I-Iac1278, I-Anayansi, I-CentaA I-Araure4, I-Ictapolochi I-Metica1, J-Tox1859

I-DGWG, I-Tanioka, ?-CO39, S-Ciwini

I-P3055F4, I-C46, I-IR84

J-Marog Paroc Rex, J-TOX340 J-El Paso, J-Oryzica-Sabana, J-Irat146, J-Monolaya, J-Fanny, I-Saavedra, I-ICTA Quirigua

M.C. Duque- CIAT

Valores de similaridad de Nei-Li entre y *dentro* de grupos de arroz.

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
(n)	(11)	(3)	(4)	(4)	(3)	(2)	(1)	(1)	(1)
Grupo 1	**0.61**	0.32	0.25	0.35	0.41	0.18	0.45	0.31	0.19
Grupo 2		**0.61**	0.52	0.34	0.22	0.40	0.23	0.40	0.33
Grupo 3			**0.65**	0.35	0.16	0.64	0.23	0.25	0.38
Grupo 4				**0.40**	0.32	0.25	0.40	0.33	0.18
Grupo 5					**0.40**	0.10	0.50	0.30	0.10
Grupo 6						**0.70**	0.15	0.10	0.55
Grupo 7							.	0.40	0.10
Grupo 8								.	0.10
Grupo 9									.

A partir de las gráficas anteriores, cómo podria recomendarse la selección de progenitores en un programa de mejoramiento, de tal manera que haya opciones de lograr mayor variabilidad genética?

Sería importante respetar las barreras genéticas para los cruces, por ejemplo, la clasificación como Indica y Japónica?

La presencia de introgresión de alelos puede visualizarse en una presentación de este tipo?

Cómo ?

Qué recomendaciones haría para éste caso en particular?

Qtl's
Conceptos básicos

Myriam Cristina Duque E.
CIAT

Definiciones básicas

Genoma:

- Conjunto de material hereditario transmitido de padres a hijos

- Moléculas de ADN arregladas en cromosomas

-El ADN se caracteriza por la secuencia de nucleótidos cuya longitud se expresa en pares de bases

Mapas genéticos:

-En los mapas genéticos hay un ordenamiento de zonas de interés a lo largo del cromosoma, pero la métrica no es física y además está bajo control genético.

-En los mapas genéticos las distancias dependen de la probabilidad de recombinación esperada entre dos puntos

Definiciones básicas (continuación)

Recombinación y mapas genéticos:

- La probabilidad de recombinación entre dos puntos del mismo cromosoma es más alta que en otra parte (el límite superior es 50%) .

- La distancia a la cual se espera un evento de recombinación depende de la región del genoma.

-Construir un mapa genético implica ordenar los loci y suministrar una medida de distancia entre ellos.

-Cuando un nuevo locus se agrega a un mapa hay que reestimar sus distancias debido a la no aditividad de sus fracciones de recombinación.

Definiciones básicas (continuación)

Recombinación y mapas genéticos:

-Los mapas consisten en la sucesión de eventos identificables o de marcadores que están en posiciones conocidas.

-Una distancia *m* en un mapa intenta medir todos los sobrecruzamientos entre dos sitios

-No todos los eventos de recombinación son observables

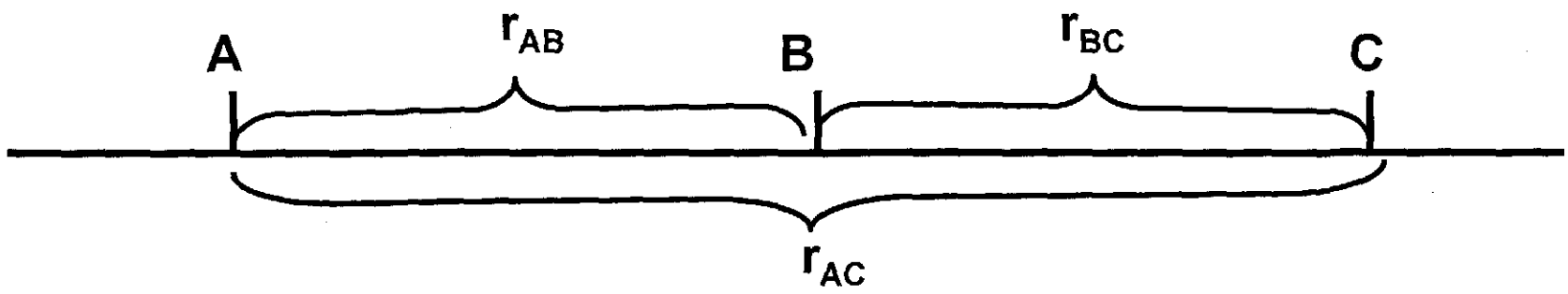
-El sobrecruzamiento o recombinación es un hecho biológico

-La frecuencia de recombinación es lo que podemos observar acerca del evento.

-*r* : *frecuencia de recombinación*, sólo mide aquella parte de los eventos recombinantes que terminan en un número impar de sobrecruzamientos .

Definiciones básicas (continuación)

Sobrecruzamientos:



r_{ij} : probabilidad de un número impar de sobrecruzamientos entre I-J

$1-r_{ij}$: probabilidad de un número par de sobrecruzamientos (incluye 0)

Un número impar de sobrecruzamientos entre AC ocurre por:

- número impar en AB y par en BC

o

- número par en AB e impar en BC

Definiciones básicas (continuación)

Frecuencias de recombinación (r):

Si no hay interferencia, lo que pasa entre AB no afecta lo que pase entre BC.

$$r_{AC} = r_{AB} \times (1 - r_{BC}) + (1 - r_{AB}) \times r_{BC}$$

$$r_{AC} = r_{AB} + r_{BC} - 2\, r_{AB}\, r_{BC}$$

Si hay interferencia, entra el factor Φ (0<=Φ<=1) de la siguiente manera:

$$r_{AC} = r_{AB} + r_{AC} - 2(1 - \varphi)\, r_{AB}\, r_{AC}$$

$$\Phi \begin{cases} 0: & \text{no interferencia} \\ 1: & \text{interferencia total} \end{cases}$$

Definiciones básicas (continuación)

-Sobrecruzamientos aleatorios e independientes (no interferencia) -

Haldane(1919) $\longrightarrow$ $m = \frac{-\ln(1-2r)}{r}$

-Modelo de Interferencia parcial :

Kosambi(1944) $\longrightarrow$ $m = \frac{1}{4}\ln\left(\frac{1+2r}{1-2r}\right)$

Definiciones básicas (continuación)

QTL's:

-Muchas características de utilidad para mejoramiento tienen variación contínua y reciben en inglés el nombre de "quantitative trait loci". Se les conoce por su abreviatura QTL's.

-Los principios son similares a los mendelianos o cualitativos, pero debido a la segregación compleja de genes, que no puede seguirse individualmente, se han desarrollado nuevos métodos y conceptos para aproximarse a su comprensión.

-La complejidad se debe a la incorporación de muchos genes y al efecto de cada uno de ellos, pequeño, si se compara con efectos ambientales.

-En QTL's el fenotipo suministra poca información acerca del genotipo.

Definiciones básicas (continuación)

QTL's:

-El interés principal : identificar porciones específicas del genoma involucradas en la variación de QTL's para :orientar programas de mejoramiento, caracterizar y/o manipular el genoma según la necesidad.

- Una forma, es que los marcadores genéticos o moleculares asociados con loci que afectan QT de interés, sean estudiados en su segregación y permitan hacer una selección indirecta llamada selección asistida por marcadores para mejorar las características más eficientemente .

-Las isoenzimas, los RFLP, os RAPD's, los microsatélites y otros marcadores a nivel de ADN han sido herramientas importantes para la detección de QTL's.

Definiciones básicas (continuación)

QTL's:

-La disponibilidad de mapas detallados de ligamiento de marcadores moleculares, hace posible la disección de QT dentro de unidades o factores llamados QTL's (Gelderman, 1975) y cuando una cantidad suficientemente grande de marcadores ha sido registrada en una familia segregante para una característica poligénica, pueden obtenerse estimaciones más precisas de efectos genéticos y ubicaciones de QTL's

Diseño experimental:

En poblaciones derivadas de 2 líneas autofecundadas se usan princincipalmente 4 tipos de diseño experimental:

1. F2
2. Retrocruces

} Reducción de tiempo

3. Líneas autofecundadas recombinantes RIL
4. Dobles Haploides

} Alto número de repeticiones

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

-Dependen del número de marcadores involucrados en el análisis

-Las más sencillas buscan: asociación entre 1 marcador y 1 QTL.

No requieren mucho conocimiento del genoma.

-Después buscan asociación : Marcador 1 ⟶ QTL ⟵ Marcador 2
(o sea la asociación entre el QTL y un intervalo que lo contenga).

Requieren conocer que los marcadores son contiguos

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

-Posteriormente buscan asociación: Grupo de ligamiento- QTL

-Finalmente, combinaciones de los dos últimos.

Requieren una clara definición de los grupos de ligamiento y un mapa confiable.

Nota: Aún son métodos:

Aproximativos, incompletos y sesgados

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Un marcador - un QTL - Esquema general

-Clasificación de individuos según el genotipo del marcador

-Cálculo de las medias de las características en los grupos definidos por los genotipos del marcador

-Si la diferencia entre los grupos, en términos de promedios es "no significativa" entonces la agrupación es arbitraria en lo que a esa característica se refiere. El marcador y la característica son independientes (no ligamiento).

-Si los promedios son diferentes es una prueba indirecta de asociación entre el marcador genético y la característica.

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Un marcador - un QTL- Objetivos

-Control de calidad de los datos

-Identificación de un modelo genético (patrón ed segregación)

-Selecćon de marcadores asociados con características de interés

Supuestos

-No hay segregación distorsionada

-No hay error en la determinación del genotipo del marcador

-Se conocen las fases de ligamiento (acoplamiento o repulsión)

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Un marcador - un QTL : Formas generales de evaluarlo:

-Pruebas de "t"

-ANOVA

-Regresión lineal

-Máxima verosimilitud (Maximum Likelihood : MLE)

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Un marcador - un QTL - Formas de evaluación en los diferentes diseños experimentales:

-En Dobles Haploides (DH) una prueba de " t " para $H_0 : \mu_{11} = \mu_{22}$ es una prueba indirecta de que r=0.

Cumple con supuestos de homogeneidad de varianzas.

Es imposible estimar sólo el efecto aditivo ya que está confundido con la fracción de recombinación.

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Un marcador - un QTL

-En F2, por tener más de dos genotipos debe pensarse en ANOVA,

pero

F2 no cumple con los supuestos debido a la dominancia y/o al ligamiento entre el marcador y el QTL,

por lo tanto

debe usarse Anova con la matriz de varianza covarianza o si es posible, prueba de "t" con varianzas no homogéneas.

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Un marcador - un QTL

En términos generales para F2:

Pruebe la homogeneidad de varianza.

Si son diferentes: ligamiento o dominancia

Si son iguales y los promedios son iguales: No hay asociación

Si son iguales y los promedios son diferentes: prueba indirecta de no dominancia y asociación.

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Un marcador - un QTL

En retrocruces con sólo un tipo (al padre 1 o al padre 2) no hay problemas con ANOVA,

pero

si se unen los dos tipos de retrocruces, se introduce de nuevo el problema de heterogeneidad de varianza.

Los valores de "a" (aditividad), "d" (dominancia) y "r" (fracción de recombinación) están confundidos.

En RIL también se dá la confusión de efectos.

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Un marcador - un QTL: Pros y contras:

Ventajas:

No requieren ordenamineto de genes ni mapas de ligamiento completos

Desventajas:

Magnitud y posiciones confundidas, posiciones no determinables y efectos no cuantificables.

A no ser que haya dominancia completa a=d , no se puede diferenciar un QTL de bajo efecto y alta asociación de otro de gran efecto y baja asociación.

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Un marcador - un QTL

Recomendación general:

Uso solamente para detección de QTL sin concluir nada acerca de su posición o magnitud.

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Un marcador - un QTL : Cómo se analiza mediante regresión lineal.

- Sea M el marcador, Q el locus asociado a una característica cuantitativa, evaluada en F2. Sean 2 alelos codominantes del locus marcador, ligados con una fracción de recombinación de r entonces 2r es la probabilidad de sobrecruzamiento en la meiosis.

Se cumple:

$$\left.\begin{aligned} \mu_{Q1,Q1} &= \mu + a \\ \mu_{Q1,Q2} &= \mu + d \\ \mu_{Q2,Q2} &= \mu - a \end{aligned}\right\} \quad y_i = \mu + a\,x_i + d\,(1-x_i^2) + e_i \qquad i = 1, 2 \ldots n$$

con x= 1 si el marcador M1 M1,

x= 0 si el marcador es M1M2 y

x= -1 si el marcador es M2M2

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Un marcador - un QTL

Modelo de regresión lineal en BC :

$$y_j = \mu + f(M_j) + \varepsilon_j \quad y_j = \mu + a_j M_j + \varepsilon_j$$

y_j : valor de la característica en el individuo j

μ : Valor medio poblacional de la característica evaluada

f (M_j): función del genotipo del marcador en términos de:

el valor genético del genotipo de Q (M_j) y

de la relación de ligamiento entre M y Q (b_j)

ε_j: residuo del individuo j

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Un marcador - un QTL

Modelo de regresión lineal general **en BC:**

$$y_j = \mu + a_j M_j + \varepsilon_j$$

$$y_j = \beta_0 + \beta_1 M_j$$

Marcador	M_1M_1	M_1M_2
códigos	1	-1
	1	0

El intercepto $ß_0$ da una estimación de : $0.5 (\mu_1 + \mu_2)$

La pendiente $ß_1$ da una estimación de : $0.5 (1 - 2 r) (\mu_1 - \mu_2)$

Ejemplo

Se tiene el cruce de Lemont x Barthii en un diseño esperimental correspondiente a

Se ha evaluado para cada línea resultante : su genotipo con respecto a un conjunto de microsatélites probados (laboratorio molecular) y una serie de variables cuantitativas relacionadas con rendimiento, morfología, fisiología, calidad y comportamiento frente a enfermedades (campo).

El interés está en tratar de encontrar si hay alguna asociación entre algún microsatélite y alguna(s) característica(s) de las evaluadas.

Ejemplo:

Marcador RM22 = B

Variable=RDTO Rendimiento (Kg/ha)

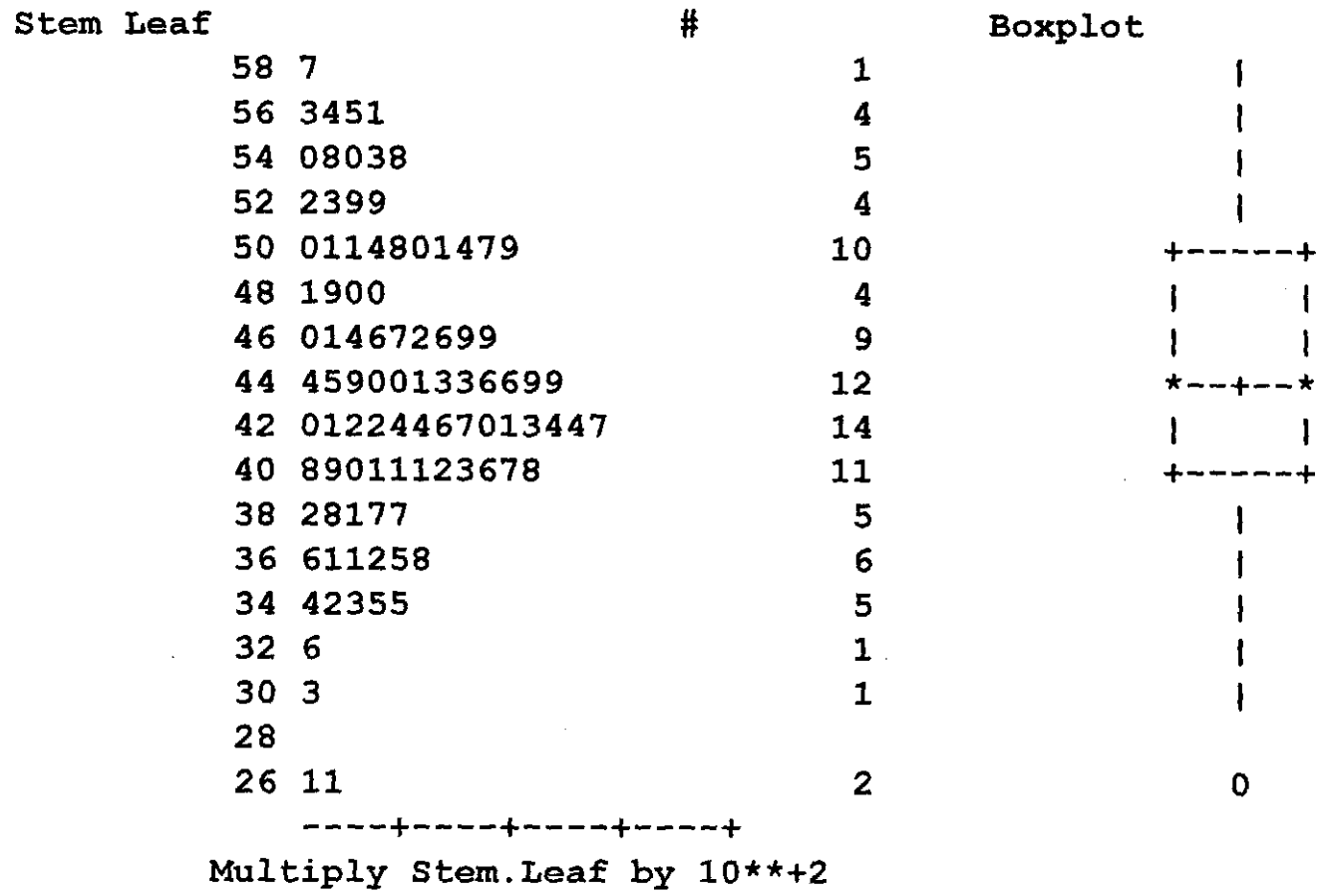
Moments

N	94	Sum Wgts	94
Mean	4477.261	Sum	420862.6
Std Dev	669.2797	Variance	447935.3
Skewness	-0.12475	Kurtosis	-0.10559
USS	1.926E9	CSS	41657984
CV	14.94842	Std Mean	69.03093
T:Mean=0	64.85877	Pr>|T|	0.0001
Num ^= 0	94	Num > 0	94
M(Sign)	47	Pr>=|M|	0.0001
Sgn Rank	2232.5	Pr>=|S|	0.0001

Quantiles(Def=5)

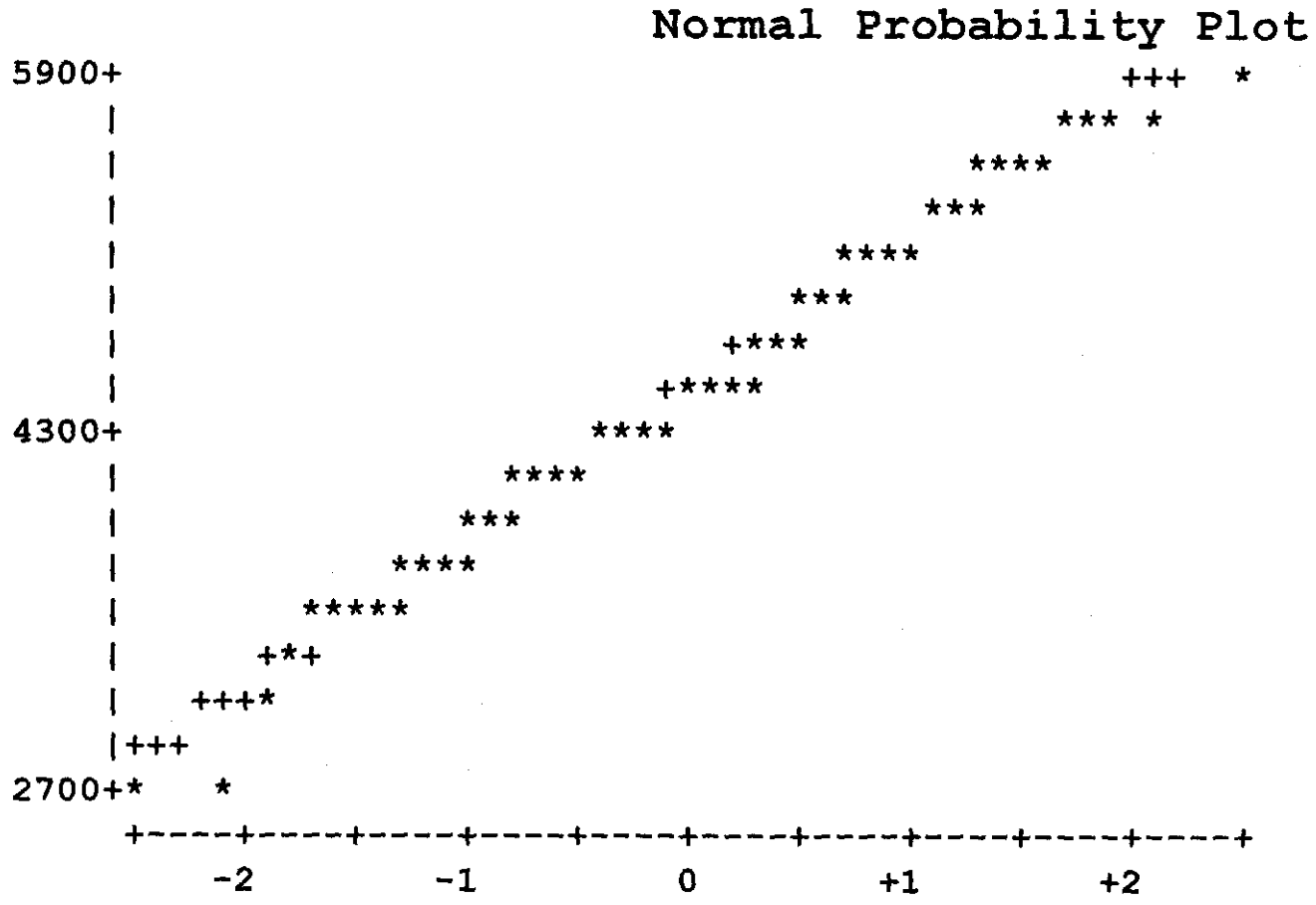
100% Max	5873.909	99%	5873.909
75% Q3	4997.957	95%	5625.906
50% Med	4466.717	90%	5395.434
25% Q1	4105.834	10%	3659.267
0% Min	2709.511	5%	3438.403
		1%	2709.511
Range	3164.398		
Q3-Q1	892.1233		
Mode	2709.511		

Marcador Rm22 =B



marcador RM22 =B

Variable=RDTO Rendimiento (Kg/ha)



marcador RM22 =L

Variable=RDTO Rendimiento (Kg/ha)

Moments

N	203	Sum Wgts	203
Mean	4534.978	Sum	920600.6
Std Dev	707.1834	Variance	500108.3
Skewness	-0.56551	Kurtosis	0.337574
USS	4.2759E9	CSS	1.0102E8
CV	15.59397	Std Mean	49.63454
T:Mean=0	91.36739	Pr>|T|	0.0001
Num ^= 0	203	Num > 0	203
M(Sign)	101.5	Pr>=|M|	0.0001
Sgn Rank	10353	Pr>=|S|	0.0001

Quantiles(Def=5)

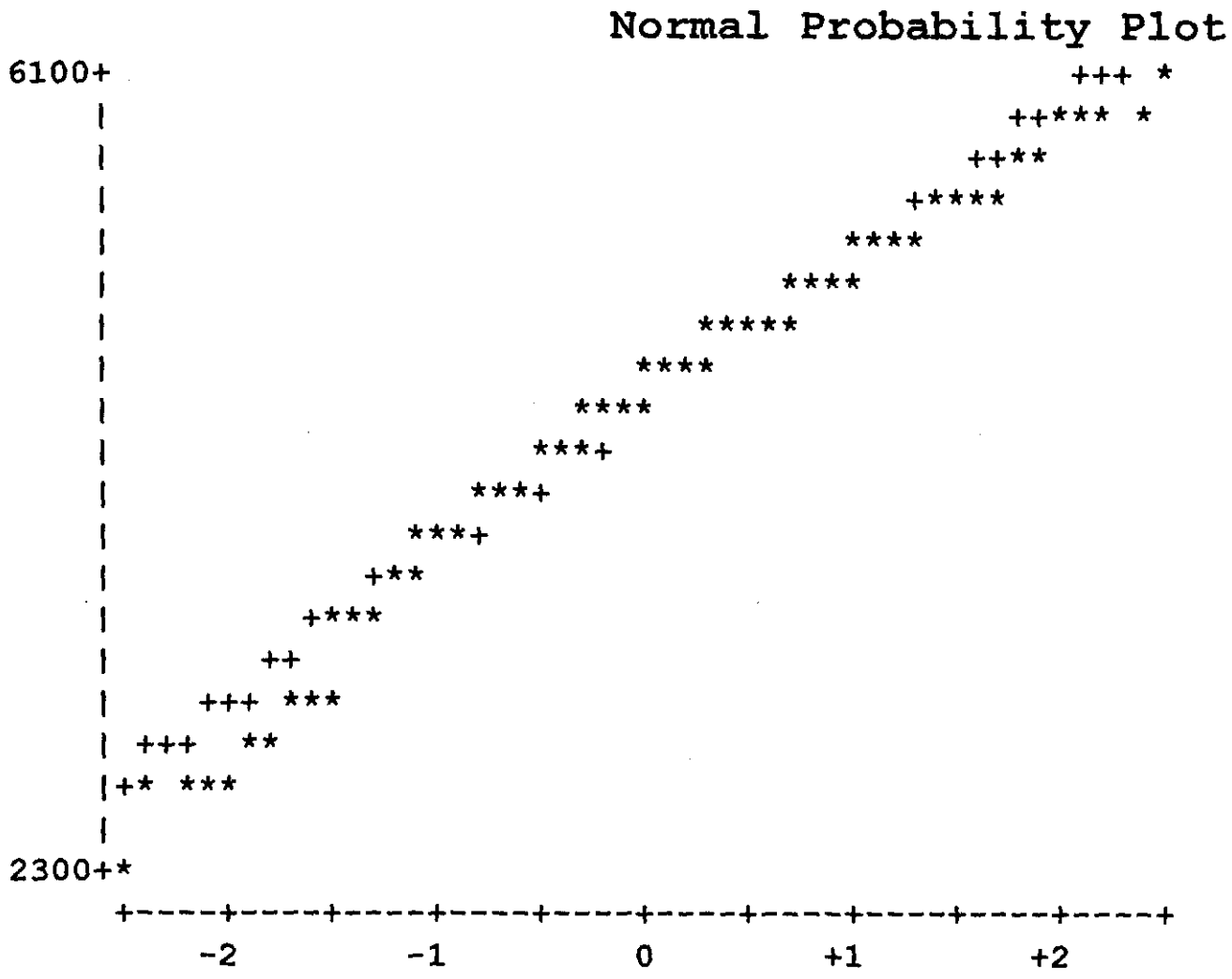
100% Max	6193.512	99%	5899.237
75% Q3	4988.447	95%	5514.249
50% Med	4632.479	90%	5369.835
25% Q1	4121.304	10%	3596.454
0% Min	2389.108	5%	3041.202
		1%	2678.925
Range	3804.404		
Q3-Q1	867.143		
Mode	2389.108		

marcador RM22 =L
Variable=RDTO Rendimiento (Kg/ha)

```
Stem Leaf                                  #        Boxplot
  60 9                                     1           |
  58 06905                                 5           |
  56 23                                    2           |
  54 3555501137                           10           |
  52 145588013357789                      15           |
  50 012333778800155 67                   17           |
  48 033455770011333445566888899          27          +-----+
  46 012333344566778889023345789 99       29          *-----*
  44 335779122223334445 88                20          |  +  |
  42 001234682245567888 99                20          |     |
  40 23467790022344789                    17          +-----+
  38 24455880579                          11           |
  36 057923478                             9           |
  34 2678227                               7           |
  32                                                   |
  30 01467                                 5           |
  28 799                                   3           |
  26 2809                                  4           0
  24                                                   |
  22 9                                     1           0
     ----+----+----+----+----+----
Multiply Stem.Leaf by 10**+2
```

marcador RM22 =L

Variable=RDTO Rendimiento (Kg/ha)

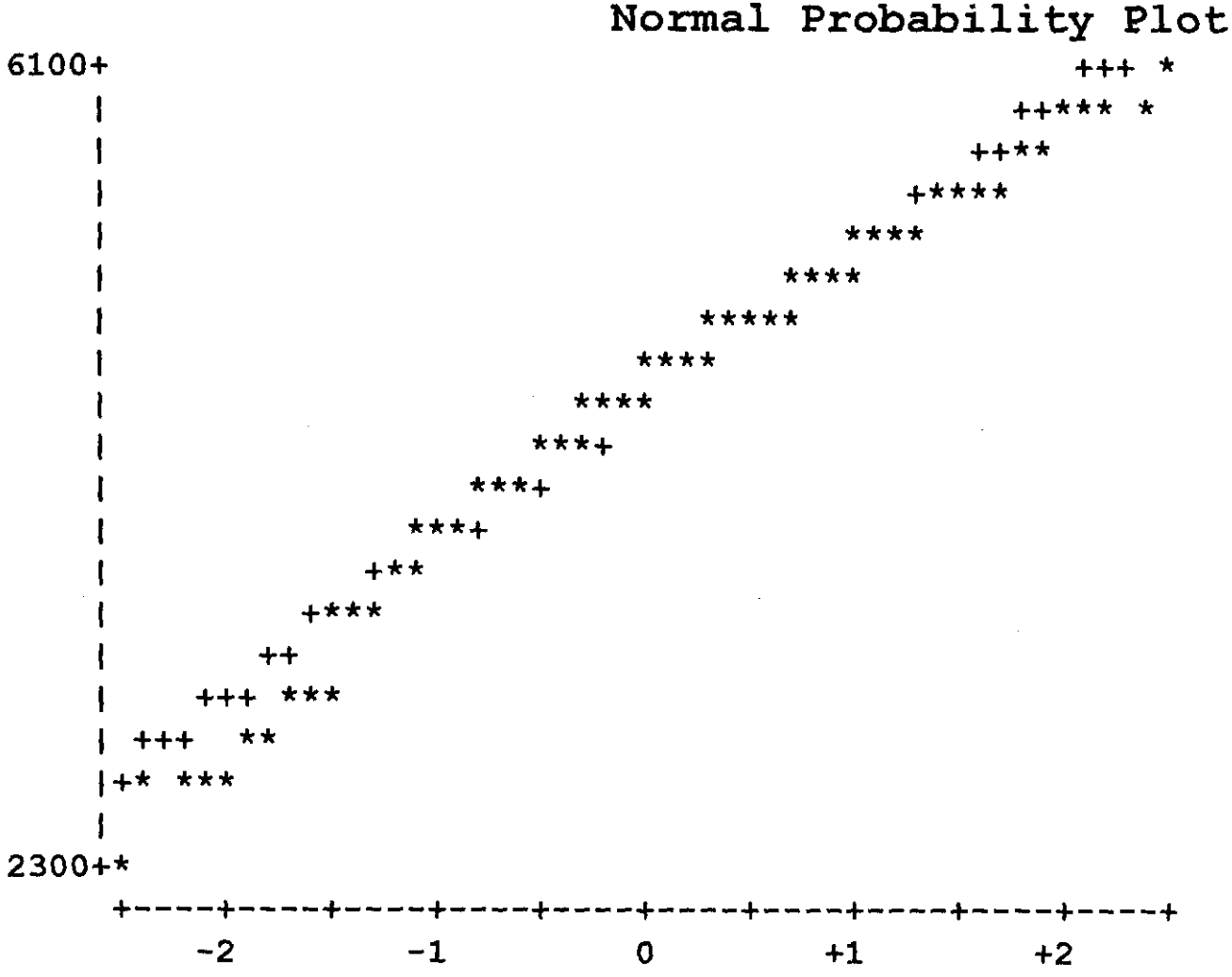


marcador RM22 =L
Variable=RDTO Rendimiento (Kg/ha)

```
Stem Leaf                                  #        Boxplot
  60 9                                     1           |
  58 06905                                 5           |
  56 23                                    2           |
  54 3555501137                           10           |
  52 145588013357789                      15           |
  50 012333778800155 67                   17           |
  48 033455770011333445566888899          27          +-----+
  46 012333344566778889023345789 99       29          *-----*
  44 335779122223334445 88                20          |  +  |
  42 001234682245567888 99                20          |     |
  40 23467790022344789                    17          +-----+
  38 24455880579                          11           |
  36 057923478                             9           |
  34 2678227                               7           |
  32                                                   |
  30 01467                                 5           |
  28 799                                   3           |
  26 2809                                  4           0
  24                                                   |
  22 9                                     1           0
     ----+----+----+----+----+----
 Multiply Stem.Leaf by 10**+2
```

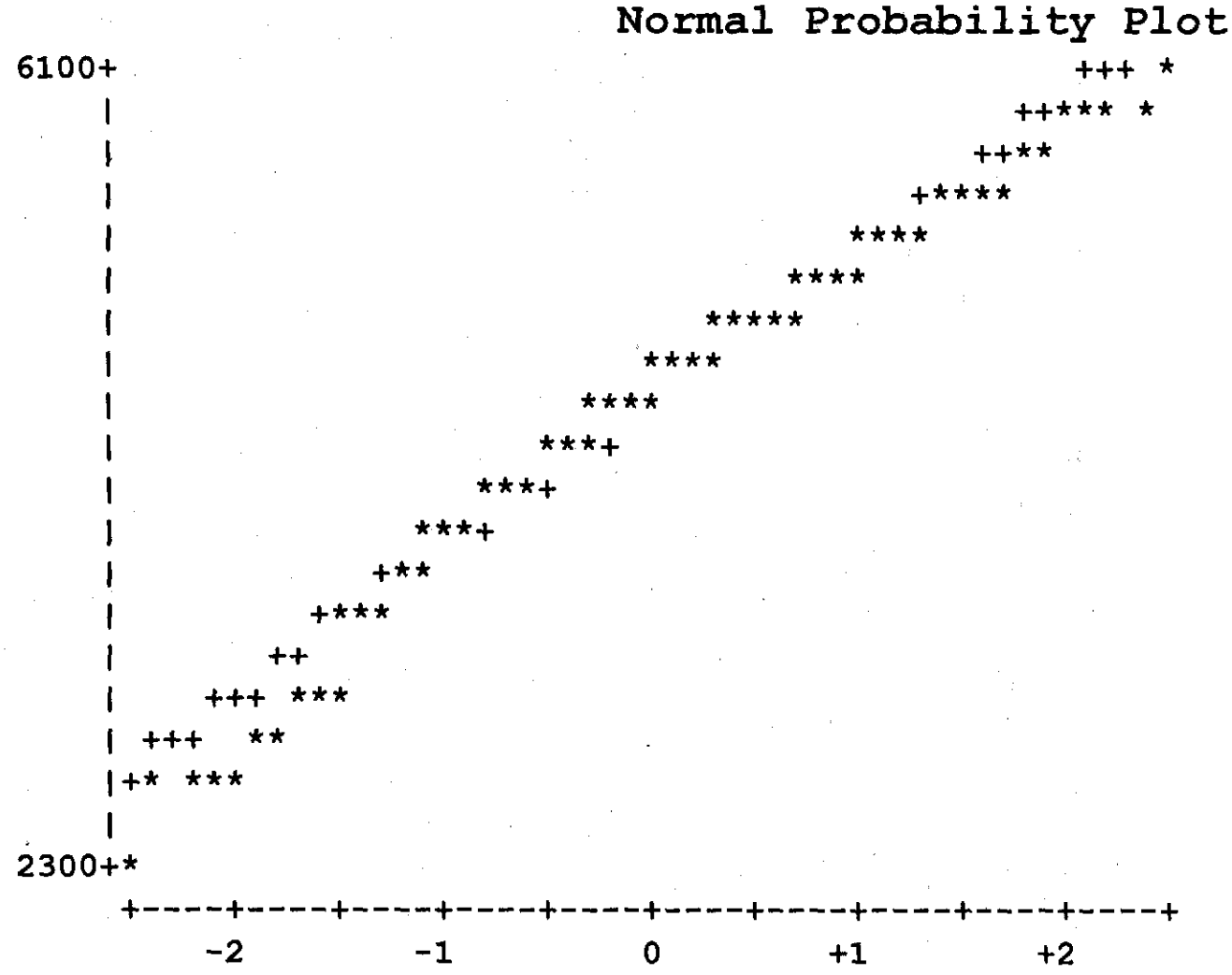
marcador RM22 =L

Variable=RDTO Rendimiento (Kg/ha)



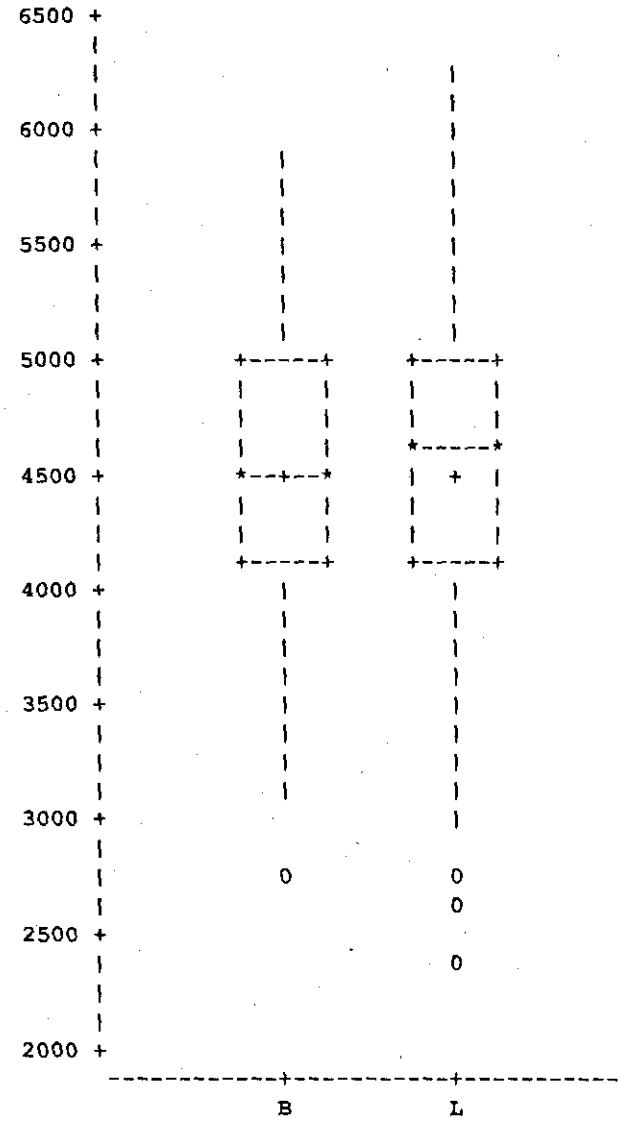
marcador RM22 =L

Variable=RDTO Rendimiento (Kg/ha)



marcador RM22 Schematic Plots

Variable=RDTO Rendimiento (Kg/ha)



Otro microsatélite....

marcador RM81b =B

Variable=RDTO Rendimiento (Kg/ha)

Moments

N	102	Sum Wgts	102
Mean	4360.704	Sum	444791.8
Std Dev	715.7787	Variance	512339.2
Skewness	-0.11466	Kurtosis	-0.28213
USS	1.9914E9	CSS	51746258
CV	16.41429	Std Mean	70.87266
T:Mean=0	61.52872	Pr>|T|	0.0001
Num ^= 0	102	Num > 0	102
M(Sign)	51	Pr>=|M|	0.0001
Sgn Rank	2626.5	Pr>=|S|	0.0001

Quantiles(Def=5)

100% Max	5873.909	99%	5709.711
75% Q3	4785.844	95%	5519.373
50% Med	4319.688	90%	5389.877
25% Q1	3880.986	10%	3518.249
0% Min	2699.564	5%	3131.858
		1%	2709.511
Range	3174.345		
Q3-Q1	904.8588		
Mode	2699.564		

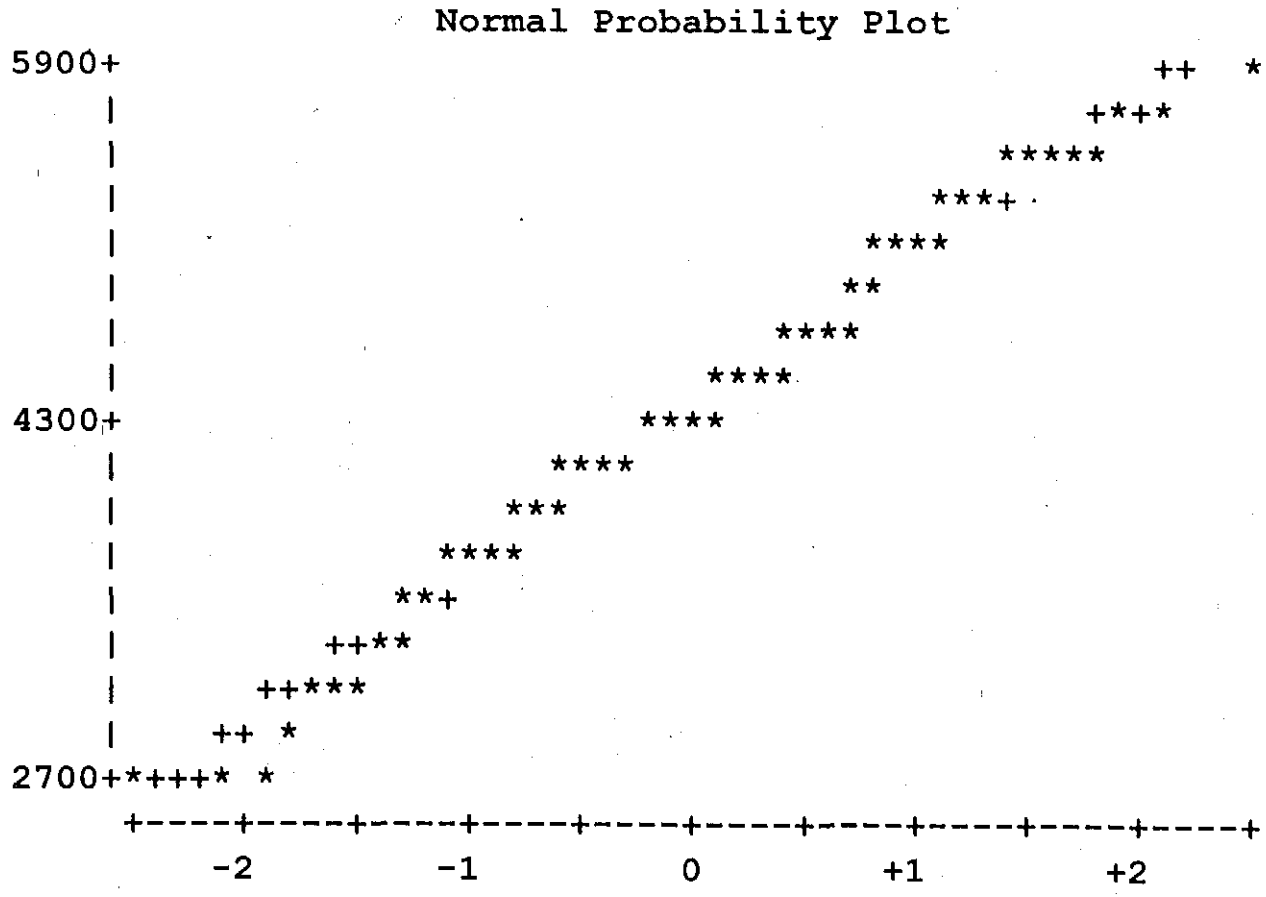
marcador RM81b =B

Variable=RDTO Rendimiento (Kg/ha)

```
Stem Leaf              #     Boxplot
  58 7                 1        |
  56 31                2        |
  54 0580238           7        |
  52 2399              4        |
  50 011801479         9        |
  48 8                 1        |
  46 0014667824699    13     +------+
  44 4599003369       10     |      |
  42 01222344712347   14     *--+--*
  40 790111223678     12     |      |
  38 22248877          8     +------+
  36 67124578          8        |
  34 255               3        |
  32 376               3        |
  30 437               3        |
  28 9                 1        |
  26 011               3        |
     ----+----+----+----+
Multiply Stem.Leaf by 10**+2
```

marcador RM81b =B

Variable=RDTO Rendimiento (Kg/ha)



marcador RM81b =L

marcador RM81b =L

Variable=RDTO Rendimiento (Kg/ha)

Moments

N	212	Sum Wgts	212
Mean	4612.27	Sum	977801.3
Std Dev	671.0654	Variance	450328.8
Skewness	-0.64253	Kurtosis	0.670216
USS	4.6049E9	CSS	95019370
CV	14.54957	Std Mean	46.08896
T:Mean=0	100.0732	Pr>|T|	0.0001
Num ^= 0	212	Num > 0	212
M(Sign)	106	Pr>=|M|	0.0001
Sgn Rank	11289	Pr>=|S|	0.0001

Quantiles(Def=5)

100% Max	6193.512	99%	5899.237
75% Q3	5059.119	95%	5573.815
50% Med	4660.899	90%	5377.384
25% Q1	4245.405	10%	3752.233
0% Min	2389.108	5%	3438.403
		1%	2678.925
Range	3804.404		
Q3-Q1	813.7148		
Mode	2389.108		

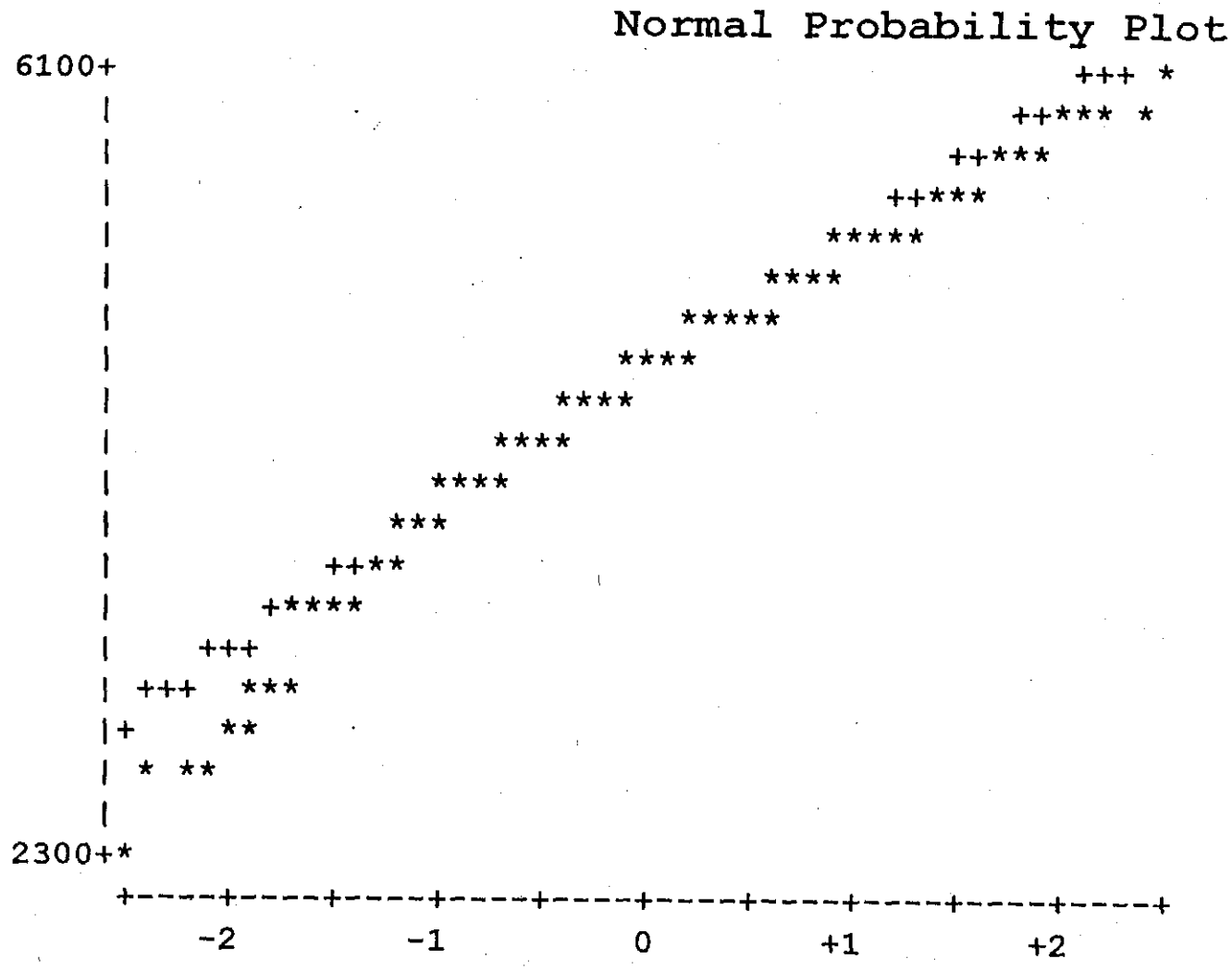
marcador RM81b =L

Variable=RDTO Rendimiento (Kg/ha)

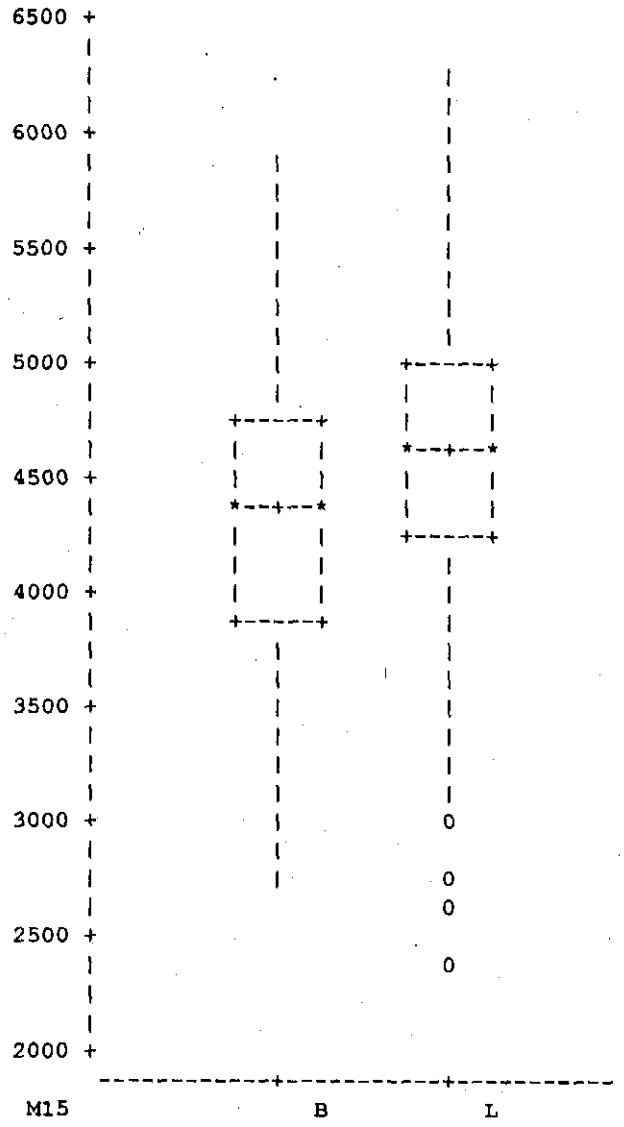
```
Stem Leaf                                  #     Boxplot
  60 9                                     1        |
  58 06905                                 5        |
  56 2335                                  4        |
  54 355501137                             9        |
  52 11455680011334577899                 20        |
  50 0123333577880013555679               22     +-----+
  48 013345577900013334445566888 99       29     |     |
  46 12333344567788902233578999           26     *--+--*
  44 33457791122223334445678 8            24     |     |
  42 00124666801224455678 8899            24     +-----+
  40 246778900234789                      15        |
  38 4555801579                           10        |
  36 09258                                 5        |
  34 246782237                             9        |
  32                                                |
  30 016                                   3        0
  28 79                                    2        0
  26 289                                   3        0
  24
  22 9                                     1        0
     ----+----+----+----+----+----
 Multiply Stem.Leaf by 10**+2
```

marcador RM81b =L

Variable=RDTO Rendimiento (Kg/ha)



Variable=RDTO Rendimiento (Kg/ha)



marcador RM81b

TTEST PROCEDURE

Variable: RDTO Rendimiento (Kg/ha)

M15	N	Mean	Std Dev	Std Error	Minimum	Maximum
B	102	4360.7040560	715.77873181	70.87265526	2699.5639535	5873.9088663
L	212	4612.2701440	671.06539877	46.08896081	2389.1079215	6193.5116279

Variances	T	DF	Prob>|T|
Unequal	-2.9757	188.4	0.0033
Equal	-3.0438	312.0	0.0025

For H0: Variances are equal, F' = 1.14 DF = (101,211) Prob>F' = 0.4364

marcador RM81b

Model: MODEL1
Dependent Variable: RDTO Rendimiento (Kg/ha)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	4358234.3272	4358234.3272	9.265	0.0025
Error	312	146765628.83	470402.65652		
C Total	313	151123863.16			

Root MSE	685.85906	R-square	0.0288
Dep Mean	4530.55122	Adj R-sq	0.0257
C.V.	15.13853		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	4360.704056	67.91016677	64.213	0.0001
MM	1	251.566088	82.64787478	3.044	0.0025

marcador RM81b

General Linear Models Procedure

Dependent Variable: RDTO Rendimiento (Kg/ha)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4358234.3272329	4358234.3272329	9.26	0.0025
Error	312	146765628.8330950	470402.6565163		
Corrected Total	313	151123863.1603280			

R-Square	C.V.	Root MSE	RDTO Mean
0.028839	15.13853	685.85906462	4530.5512237

Source	DF	Type III SS	Mean Square	F Value	Pr > F
M15	1	4358234.3272329	4358234.3272329	9.26	0.0025

Y Ahora....

Cuáles son las conclusiones ?

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Dos marcadores - un QTL

Objetivos:

Detección de ligamiento en el mismo cromosoma

Estimación de la fracción de recombinación

Requerimientos:

Mapa genético: definición muy precisa de grupos de ligamiento y del ordenamiento de locus

Supuestos:

Segregación no distorsionada

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Estimación de Máxima Verosimilitud (Maximum likelihood MLE)

Objetivo: tratar de separar la estimación de la fracción de recombinación de los parámetros a y d mediante la función de verosimilitud L.

L : Encuentra el valor de los parámetros que hacen que la muestra obtenida sea la más probable (r, μ_i, σ_i^2).

Depende del tipo de diseño experimental utilizado.

Por ejemplo, para DH: un individuo puede ser M_1M_1 or M_2M_2.

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Estimación de Máxima Verosimilitud

La función de probabilidad para la característica de interés en el grupo de los $M_i M_i$ es $f(y_i)$

$$f(y_i) = pr(no\_sobrec)f(Q_1Q_1) + pr(sobrec)f(Q_2Q_2)$$

Y para toda la población es:

$$L = f(y_1)\times\ldots\times f(y_1)\times f(y_2)\times\ldots\times f(y_2)$$

(n_1 factores $f(y_1)$; n_2 factores $f(y_2)$)

Individuos pertenecientes a la clase genotípica M_1M_1

Individuos pertenecientes a la clase genotípica M_2M_2 (n_2)

Métodos para detección de QTL's en poblaciones derivadas de líneas autofecundadas

Indicadores

Lod score (Z) $\longrightarrow$ $Z = \log \left[\frac{L(\hat{\theta})}{L(0.5)} \right]$

Valor crítico : 3

Interpretación:

El ligamiento con una fracción de recombinación estimada de

es 10^3 = 1000 veces mas probable que el no ligamiento (0.5)

Introduction

Data obtained from molecular analysis, such as DNA sequences or banding patterns (fingerprints), are often used as basis for classifying individuals in a population and for making evolutionary or phylogenetic inferences. Dendrograms are constructed based on the degrees of similarity between individuals. The confidence limits for the groupings produced by dendrograms, however, are not generally amenable to computation by the usual statistical procedures.

Felsenstein (1985) proposed using bootstrapping (Efron 1979) as a way of obtaining a nonparametric estimate of these confidence limits. Felsenstein's method involves repeated sampling with replacement (bootstrapping) of the characters in a matrix of Operational Taxonomic Units (OTUs) × characters, to create numerous bootstrap matrices of the same size as the original matrix. Dendrograms for each of these bootstrap samples can then be constructed. The frequency with which a particular group appears among all the dendrograms constructed provides an indication of the degree of support for that group. The results can also be combined into a majority-rule consensus tree to obtain an overall bootstrap estimate of the dendrogram. (For a different approach to bootstrapping, see Brown [1994], which implements hypothesis testing for dendrograms using bootstrapping.)

There are two general methods for constructing dendrograms—parsimony analysis and cluster analysis. Parsimony analysis has a theoretical phylogenetic basis but it is computation intensive; depending on the size of the data set, a single run may take several hours, even on a fast computer. Since bootstrapping requires several hundred iterations to be statistically accurate (up to 2,000 bootstrap replications at the 95% level, Hedges [1992]), it is simply not practical to perform bootstrapping for parsimony-based dendrograms at this time.

A much faster method of obtaining a dendrogram is cluster analysis. This involves first computing a matrix of similarity coefficients for all pairs of OTUs and then performing the actual cluster analysis based on the similarity matrix by the Unweighted Pair-Group Method, Arithmetic Mean (UPGMA). The resulting dendrogram provides a good estimate of the phylogeny of a particular group of organisms. Most UPGMA runs take less than a minute to complete.

Felsenstein's PHYLIP (Phylogeny Inference Package) offers a way to perform bootstrapping for parsimony-based dendrograms. Although PHYLIP can perform UPGMA, it does not perform bootstrapping for UPGMA. We have developed WinBoot therefore to perform UPGMA-based bootstrapping.

Note that you should do bootstrapping analysis in conjunction with a more traditional UPGMA cluster analysis, using the same binary data to perform UPGMA as input data for WinBoot. Most large statistical software packages such as SAS or Systat can perform some type of cluster analysis. In our laboratory, we use the companion program WinDist (included with WinBoot, see Appendix 2) to compute for similarity matrices. We use a commercial software package called NTSYS-PC (Rohlf 1994) to do the cluster analysis.

Although WinBoot was developed to analyze DNA fingerprints (restriction fragment length polymorphism [RFLP] banding patterns) of rice pathogens, any type of binary data from any source can be used as input. The only requirement is that the data be in binary form, where 1 stands for "character [band] present" and 0 stands for "character [band] absent".

Technical Information

The main algorithms for WinBoot were derived from the PHYLIP programs NEIGHBOR, which computes UPGMA dendrograms, and CONSENSE, which computes for consensus trees (Felsenstein 1985). The program was developed using Borland Pascal 7.0 running under Windows 3.1. The source code takes advantage of Borland's object-oriented extensions to the Pascal language and uses the Object Windows Library (OWL). The interface was designed according to the specifications of the Borland Windows Custom Controls (BWCC). It has been run under Windows 3.1, Windows for Workgroups 3.11, and Windows 95.

Installation

WinBoot files
The WinBoot distribution diskette contains at least four files:

WINBOOT.EXE
WINDIST.EXE
BANDS.DLL
BWCC.DLL

The diskette you receive may also contain some additional files. Consult the README.TXT file on the diskette (if present) for updates and more information.

WinBoot does not have an automatic installation program, so you must manually copy the WinBoot files from the distribution diskette to your hard disk. Use the following procedure (consult an experienced user if you have trouble):

1. Create a subdirectory on your hard disk called WinBoot.
2. Insert the diskette into the disk drive.
3. Copy all the files from the diskette to the newly created subdirectory.
4. In Program Manager, create a new program group called "WinBoot" (optional).
5. In the new program group (or a pre-existing group), create two new program items for the program files WINBOOT.EXE and WINDIST.EXE.

Input File Formats

Data for WinBoot must be in the form of a matrix of OTUs × characters. Each OTU could represent a species, race, strain, etc. The characters must be encoded in binary form (i.e., 1 or 0). For DNA fingerprint data, each character represents a band position; a "1" indicates that a band is present at that position and a "0" indicates that the band is absent. The data file may be either in the format described for PHYLIP (Felsenstein 1985) or in tab-delimited format. (Additional formats may be added in the future.) Although not mandatory, it is suggested that all input files be given the extension ".DAT" since this is default extension for input files in the **Input File Name** dialog box (described in the next section).

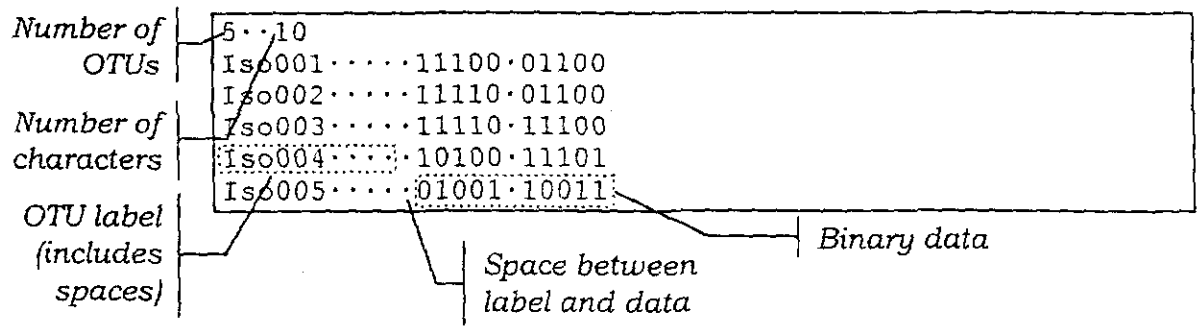
PHYLIP format

PHYLIP uses a stylized text file format, as shown in Listing 1. The first line contains two numbers indicating the number of OTUs and the number of binary characters (e.g., band positions), respectively, separated by at least one blank space. Subsequent lines contain the binary data for each OTU. Each of these lines starts with a label identifying the OTU, followed by the binary characters for that OTU.

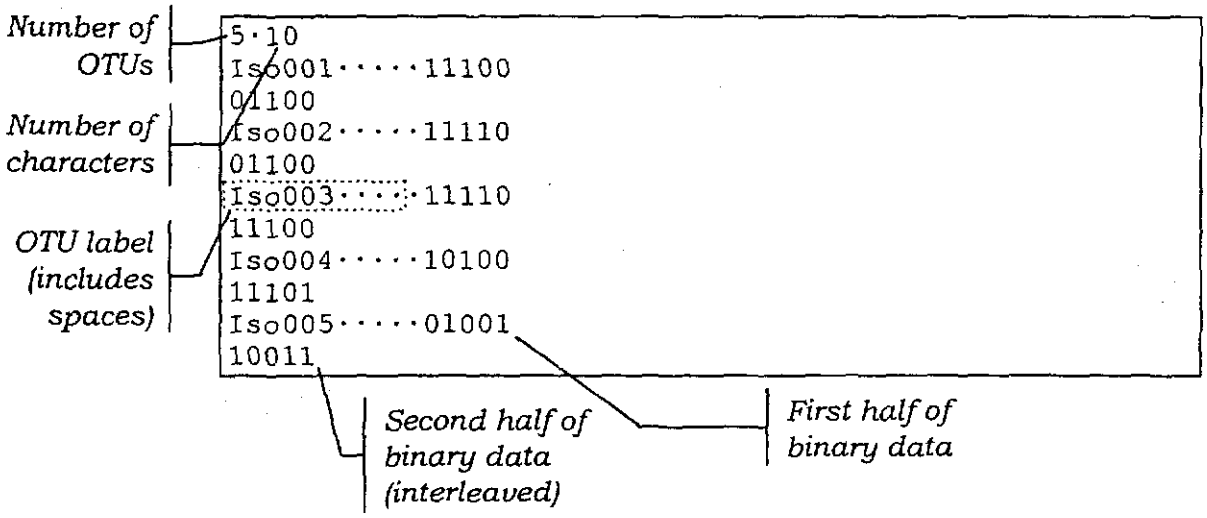
The OTU label consists of ten printable symbols and can include any combination of letters, numbers, punctuation marks, and special symbols. For maximum compatibility with other programs, however, you should only use letters and numbers. Although the program allows blank spaces in the OTU label, it is better to use an underscore "_" character instead. If the label is fewer than ten symbols long, you should fill in the remainder with blank spaces. Note that each label in Listing 1 consists of six printing characters plus four blank spaces. The label must be separated from the binary data with at least one space.

The binary data are composed of a string of 1's and 0's, indicating character present and character absent, respectively. These are the *numerals* "1" and "0" and *not* the *letters* "l" (lower-case L) and "O" (upper-case O). Spaces may be inserted between characters

Listing 1. PHYLIP file format. For clarity, spaces are represented by a "·" symbol; these will be printed as white space.



Listing 2. Interleaved PHYLIP format. For clarity, spaces are represented by a "·" symbol; these will be printed as white space.



for clarity. In Listing 1, for example, a space has been inserted between the fifth and the sixth characters. The label and data for the next OTU should begin at the start of a new line.

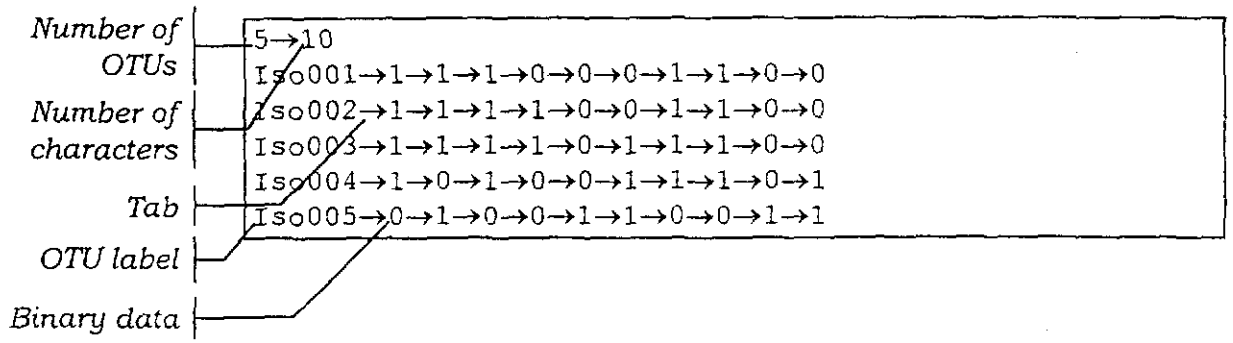
Another way to format a PHYLIP file is to interleave the binary data with the labels. In Listing 2, instead of a single line containing both the OTU label and all of the binary characters, the sixth to tenth characters have been entered onto a different line.

Tab-delimited format

The tab-delimited format has been included to make it easier to transfer data from spreadsheet programs such as Microsoft Excel, Quattro Pro, or Lotus 1-2-3. Different elements (i.e., OTU labels, binary characters) are separated from each other by tabs (i.e., the character produced when you press the Tab key on the keyboard).

The first line consists of two numbers—the number of OTUs and the number of characters—separated by a "tab" (not a space as in

Listing 3. Tab-delimited file format. For clarity, tabs are represented by a "→" symbol; these will be printed as white space.



the PHYLIP format). Subsequent lines contain the binary data for each OTU, starting with a label identifying the OTU and followed by the binary characters for that OTU.

The OTU label consists of up to ten printable characters and can include any combination of letters, numbers, punctuation marks, and special symbols. If the label contains more than ten characters, then the excess will be truncated. The program allows blank spaces, but their use is strongly discouraged for reasons of compatibility with other programs; if you want to put a space in a label, use an underscore "_" character instead. The label must be separated from the binary data with a tab.

The binary data is composed of a string of 1's and 0's, indicating character present and character absent, respectively. These are the *numerals* "1" and "0" and *not* the *letters* "l" (lower-case L) and "O" (upper-case O). Characters must be separated from each other by tabs.

Using Excel to create a tab-delimited file

The following procedure describes how to create a tab-delimited file using Microsoft Excel 5.0:

1. Place the number of OTUs in cell A1 and the number of characters in cell B1.
2. Beginning with row 2, enter the label and data for the OTUs, one row for each. In column A, enter the label. Enter the binary characters for the respective OTU in succeeding columns, one character per cell.
3. From the **File** menu, choose **Save As**.
4. In the **Save File as Type** combo box, choose **Text (Tab delimited)**.
5. In the **File Name** edit box, type in a file name. Change the ".TXT" extension (which is entered by default) to ".DAT".
6. Click the OK button.
7. A warning box appears telling you that the "Selected file type will save only the active sheet." Click OK.

If you use a spreadsheet program other than Excel, consult your user manual for the proper procedure.

Using WinBoot

WinBoot is intended to be user-friendly enough to be used with little difficulty by anyone familiar with the conventions of Windows programs. Hence it is assumed that the user is already familiar with the DOS/Windows interface.

Running WinBoot

Like other Windows programs, you can run WinBoot by double-clicking on its icon in the Program Manager (Fig. 1).

Figure 1. The WinBoot icon.

The Main Program Window

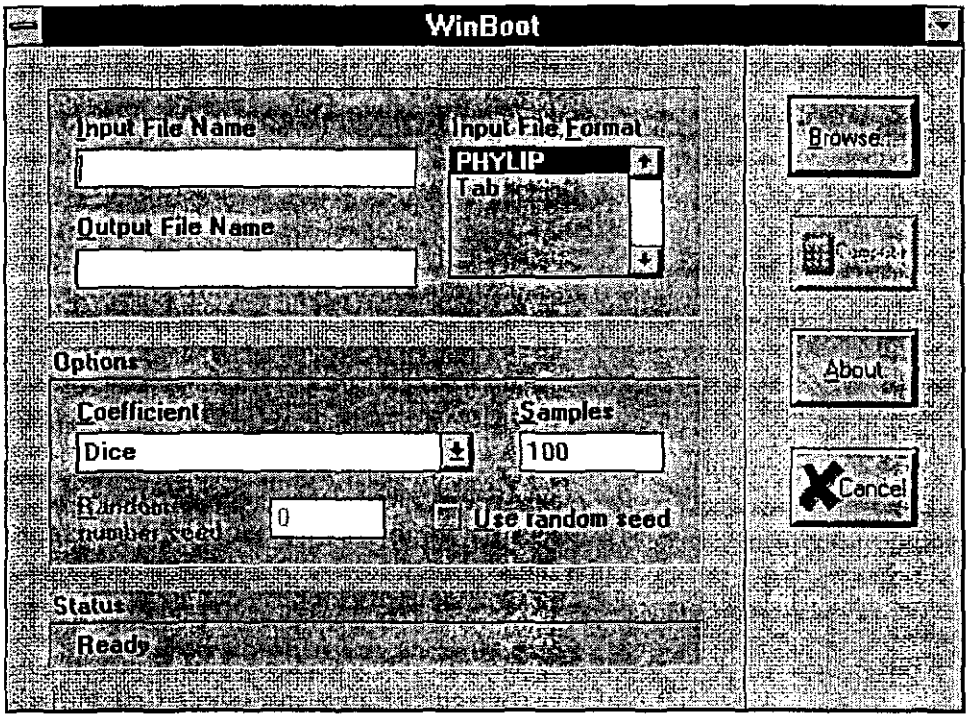


Figure 2. The WinBoot main program window.

The main program window appears when you run the program. It consists of several panels (Fig. 2). In the topmost panel, you can specify the names of the input and output files as well as the format of the input file. In the middle **Options** panel you can specify the coefficient the program will use to construct similarity matrices, how many bootstrap samples the program will perform, and whether or not the program should use a specific random number seed. The bottom **Status** panel indicates what stage the program is in; the initial status is **Ready**, indicating that the program is ready for your input. On the right side is an action panel with several buttons that allow you to perform the various actions of the program. Each of these panels is explained in more detail below.

Specifying the Input File

To specify an input file, you can type directly in the **Input File Name** edit box on the main program window. If the file is not in the current directory, you must give a complete path specification for the file.

An easier way to specify the input file is to click on the Browse button. This brings up the **Input File Name** dialog box, which is a standard dialog box where you can specify the name and location of your input file (Fig. 3).

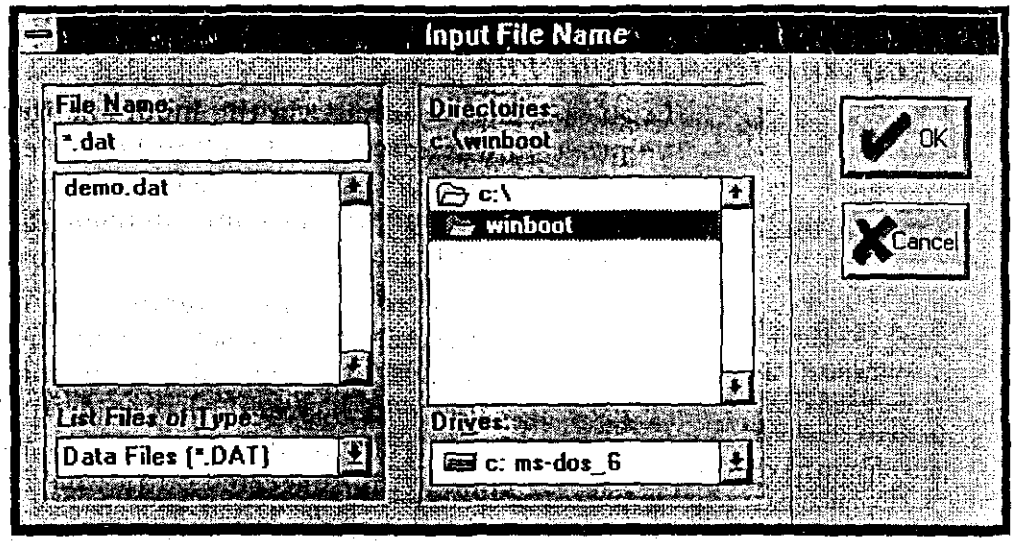
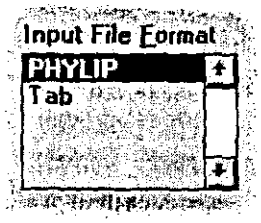


Figure 3. The Input File Name dialog box.

You can type in the name of the input file in the **File Name** edit box, or simply select it by clicking on the name in the corresponding list box. If the file you want is not in the current drive or directory, you can select these from the **Drives** combo box and the **Directories** list box, respectively. By default, only files with the extension .DAT will be listed, but you can change this by selecting from the **List Files of Type** combo box. When you have selected the correct file, click on the OK button. The program will return to the main program window with the **Input File Name** edit box automatically filled in with the name you specified.

Choosing an Input File Format



You can select the format of the input file from the **Input File Format** list box (Fig. 4). At present, you only have a choice of either **PHYLIP** or **Tab** for the PHYLIP file format or tab-delimited file format, respectively, but support for additional file formats may be added in the future. The section entitled "Input File Formats" explains these formats in more detail.

Figure 4. The Input File Format list box.

Specifying the Output File

If you specified an input via the **Input File Name** dialog box, then the program will automatically fill in both the **Input File Name** and **Output File Name** edit boxes based on your selection. By default, the output file name will have an extension of ".TXT", but you can edit this to your liking.

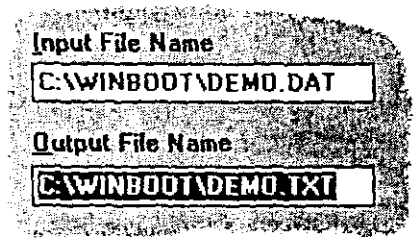


Figure 5. The Input and Output File Name edit boxes.

In Figure 5, for example, a file with the name "DEMO.DAT" was selected in the previous step; the program automatically filled this in the **Input File Name** edit box and also entered "DEMO.TXT" in the **Output File Name** edit box. At this point, you may change the suggested output file name if you wish.

On the other hand, if you specified the input file by typing directly into the **Input File Name** edit box, then you will need to do the same thing to specify the output file—type directly into the **Output File Name** edit box.

Similarity Coefficient

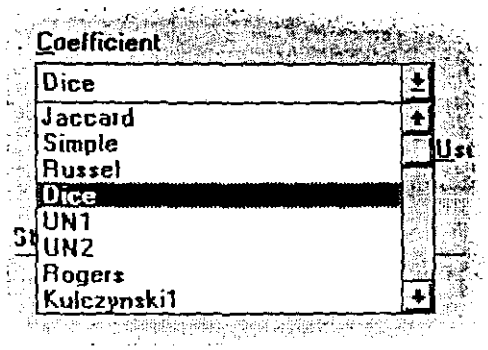


Figure 6. The Coefficient combo box with the Dice coefficient selected.

WinBoot gives you a choice of coefficients to use to compute for the similarity matrix. By default, WinBoot uses the Dice coefficient. To change this, select the **Coefficient** combo box to choose the similarity coefficient you wish to use (Fig. 6). The coefficients listed are those defined by Sokal and Sneath (1963). (See Appendix 1 for a complete description of each coefficient.) Note that you should use the same similarity coefficient for bootstrapping as the one you used to perform UPGMA cluster analysis. The WinDist program (Appendix 2) can be used to produce a similarity matrix from the same input file and with the same coefficient as WinBoot.

Number of Bootstrap Samples

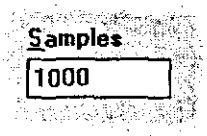


Figure 7. The Samples edit box.

The accuracy of a bootstrap estimate is a function of both the value of the bootstrap and the number of bootstrap samples or replications (Hedges 1992). To ensure that the accuracy of the bootstrap estimate is ±1%, you must do 400 replications if the bootstrap estimate is 99% or 2,000 replications if the bootstrap estimate is 95%. Lower bootstrap estimates at lower replications have correspondingly lower accuracy. By default, WinBoot will perform 100 samples, which is sufficient for a quick estimate of the bootstrap value. For more rigorous testing, increase the number of replications in the **Samples** edit box (Fig. 7).

Random Number Seed

At the heart of the bootstrap procedure is the random sampling of the data matrix. To simulate this, the program generates a series of pseudo-random numbers. The series starts from a specific number called a "seed." Each time the program runs with a specific seed value, the same sequence of numbers is produced; running the program with a different seed value produces a different series of numbers.

The **Use random seed** check box controls the usage of the random number seed. If you check this, the **Random number seed** edit box is activated, and you can explicitly state the random number seed you want to use (Fig. 8). If the check box is off, or the value of the seed you entered is 0, the program will get its seed value from the system clock, effectively producing random results.

Figure 8. The Random number seed edit box and Use random number check box.

In most cases, you can leave the check box unselected, and let the program choose its own seed value. Sometimes, though, you may get an error (Run-time error 207) which prematurely terminates the program. This is caused by the program attempting to do something illegal like dividing by zero. In cases like this, you may find it useful to explicitly specify the random number seed. Keep changing the seed value until you find one that works.

Performing the Bootstrapping

Click the Compute button to start the bootstrapping process.

Replacing an Existing File

If the output file you specified is already present, you will get the **Output File Name** dialog box, asking you whether you want to replace the existing file (Fig. 9).

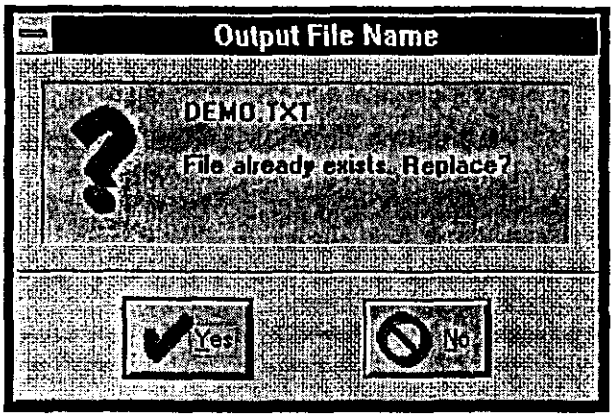


Figure 9. The Output File Name dialog box.

Click on the Yes button to replace the file—any previous contents the file may have will be deleted. The program will go on computing. Click on the No button to go back to the main program window, where you can change the output file name.

Program Status

Figure 10. The program Status indicator.

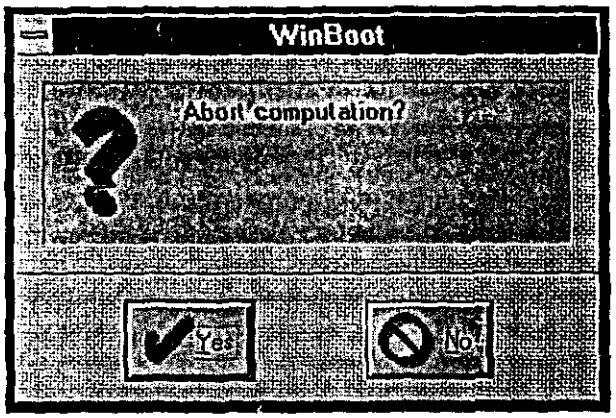
Once the program begins computing, the cursor will turn into an hourglass and the **Status** indicator at the bottom of the main program window will change from **Ready** to **Computing...** (Fig. 10). The number after the ellipses indicates the number of bootstrap samples the program has processed; this will increase from 1 to the number of samples you specified in the **Samples** edit box. When the program has finished, it will beep, and the status indicator will change back to **Ready**.

Switching to Other Programs

Note that the hourglass cursor changes back to an arrow when you move it out of the WinBoot window area. This indicates that you can minimize or switch away from WinBoot even while the program is computing. WinBoot can continue to compute in the background while other applications are being used.

Canceling the Computation

You can abort computation at any time by clicking the Cancel button. A dialog box will appear asking you to confirm the termination (Fig. 11).



You can click Yes to stop computing and return to the main program screen. If you click No, the program will continue computing.

Figure 11. The Abort Computation dialog box.

Exiting the program

Click the Cancel button to exit the program.

WinBoot Output

Listing 4 shows a printout of the output file produced by WinBoot. You can examine and print the file using a text editor such as the Windows Notepad. This is a standard ASCII text file which by default has the extension ".TXT". The output file consists of four sections, explained below. Note that the word "species" in the printout actually refers to the OTUs in the input file.

The first section lists the OTUs in the order that they were given in the input data file. The second section lists the subsets (groups) of OTUs that appear in the consensus tree. This is the same information that appears in the consensus tree. The third section lists the subsets that appeared in one or more of the individual (bootstrap) trees but did not appear frequently enough to be included in the consensus tree. If a particular set appears in the UPGMA dendrogram but does not appear in the bootstrap consensus tree, you may find it here.

Both of the lists of subsets consist of a row of symbols; a "." (period) symbol indicates that the species is not present in the set while a "*" (asterisk) symbol indicates that the species is present in the set. The order of symbols, from left to right, corresponds to the order of species in the species list, from top to bottom. Printed to the right is a column indicating the number of times that a particular set appeared among the bootstrap replications.

The fourth section is a diagram of the consensus tree itself. Printed at each internal node of the tree is the percentage of times

the group above and to the right of that node is found. Thus the tree shown in Listing 4 shows that the group (Iso002, Iso003) occurred in 55.3% of the bootstrap trees, the group (Iso001, Iso002, Iso003) occurred 68.3% of the time, and the group (Iso001, Iso002, Iso003, Iso004) occurred 89.3% of the time. (Note that *percentages* are printed in the consensus tree while the *actual numbers* are printed in the two lists of sets.) The percentages can be considered to be statistical tests (confidence limits) on the validity of the various groups. The higher the percentage, the greater the confidence that a particular group is true, rather than being an artifact of the clustering process.

Listing 4. Sample WinBoot output.

```
Species in order:

Iso001
Iso002
Iso003
Iso004
Iso005


Sets included in the consensus tree:

Set (species in order)    How many times out of 2000

*****                     1785
***..                     1367
.**..                     1107


Sets NOT included in the consensus tree:

Set (species in order)    How many times out of 2000

**...                     892
..**.                     334
*..*.                     290
...**                     139
*...*                     44
.***.                     37
.*.*.                     2
.****                     2
.***.                     1


CONSENSUS TREE:
the numbers at the forks show the percentage
of times the group consisting of the species
which are to the right of that fork occurred

          +---------Iso001
     +-68.3
     |    |    +----Iso002
+-89.3    +-55.3
|    |         +----Iso003
|    |
|    +--------------Iso004
|
+-------------------Iso005

WinBoot computational run time: 0:01:42.095
```

The second species (OTU) in the list...

...indicates the second species (OTU) in a set

The actual number of times that this set appeared among the bootstrap replications

These three OTUs are part of this set...

...while these two are not

The set with these four OTUs appeared in 89.3% of the bootstrap replications

Analysis Using WinBoot: Case studies

Bacterial leaf blight

Nelson et al (1994) used RFLP fingerprinting to analyze the population structure of the bacterial leaf blight pathogen of rice (*Xanthomonas oryzae* pv. *oryzae*). Using the probe *IS1113*, 24 distinct banding patterns (haplotypes) were distinguished from a set of 155 Philippine strains. UPGMA cluster analysis with Dice's coefficient resolves six groups, designated A to F, at the 0.75 similarity level (Fig. 12).

Felsenstein (1985) suggested that only groups with bootstrap *P* values of 95% or greater be considered significant (but see Brown 1994). Following this rule, only group A is truly robust, with a *P*

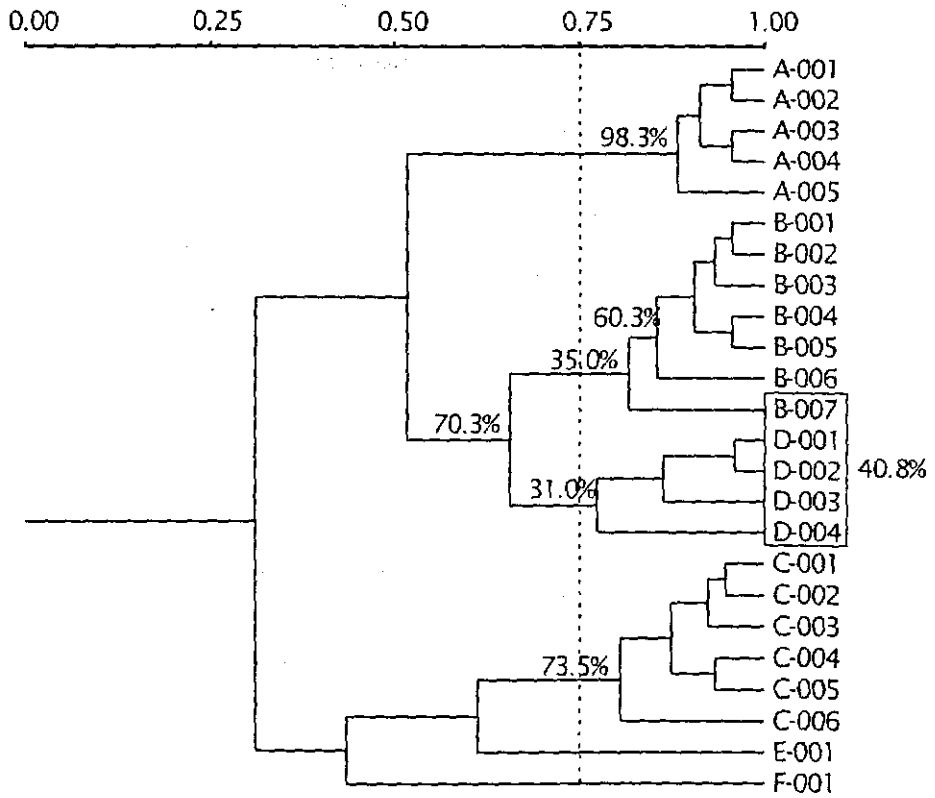


Figure 12. UPGMA dendrogram of unique haplotypes of *X. oryzae* pv. *oryzae* based on DNA banding patterns using the probe *IS1113*. Six clusters, A to F, are resolved at the 0.75 similarity level. Bootstrap *P* values are indicated at the corresponding node for each cluster. Note that group B does not actually appear on the bootstrap consensus tree; B-007 clusters together with group D at the 40.8% bootstrap level (Nelson et al 1994).

value of 98.3%*. Although group C has a relatively low *P* value, 73.5%, it is still quite distinct from the other groups. Groups E and F consist of only one haplotype each, so bootstrap values for these groups are not applicable. Groups B and D, however, are not so robust. Haplotype B-007, which by UPGMA cluster analysis falls in group B, in fact appears on the bootstrap consensus tree clustered with group D at $P = 40.8\%$. Group D itself appears on the consensus tree with a *P* value of 31.0%. In order to determine the bootstrap values for group B as it appears in the UPGMA dendrogram, you must examine WinBoot output file for the list of sets that do not appear in the consensus tree. Doing this shows that the *P* value for this group is 35.0%. The larger cluster consisting of both groups B and D has a *P* value of 70.3%, however, indicating that, though the boundary between B and D may not be so distinct, they may actually form a larger, more robust cluster. This observation is supported by other types of analysis such as parsimony analysis, and by inoculation experiments on differential rice lines.

Rice blast

Chen et al (1995) used UPGMA cluster analysis with Dice's coefficient to determine the population structure of the rice blast pathogen (*Pyricularia grisea*). At the 0.80 similarity level, ten clusters were resolved out of 1,516 strains. All the clusters showed very high bootstrap *P* values; six clusters appeared up in 100% of the bootstrap trees, two clusters had 99.7% and 99.6% *P* values, respectively. *P* values are not applicable for the other two clusters since these consisted of a single haplotype each (Figure 13). These results indicate that the population structure of *P. grisea* in the Philippines is very well defined and may consist of several distinct, clonal lineages.

* A different random number seed was used for this analysis, hence the bootstrap values given here are slightly different from the published results.

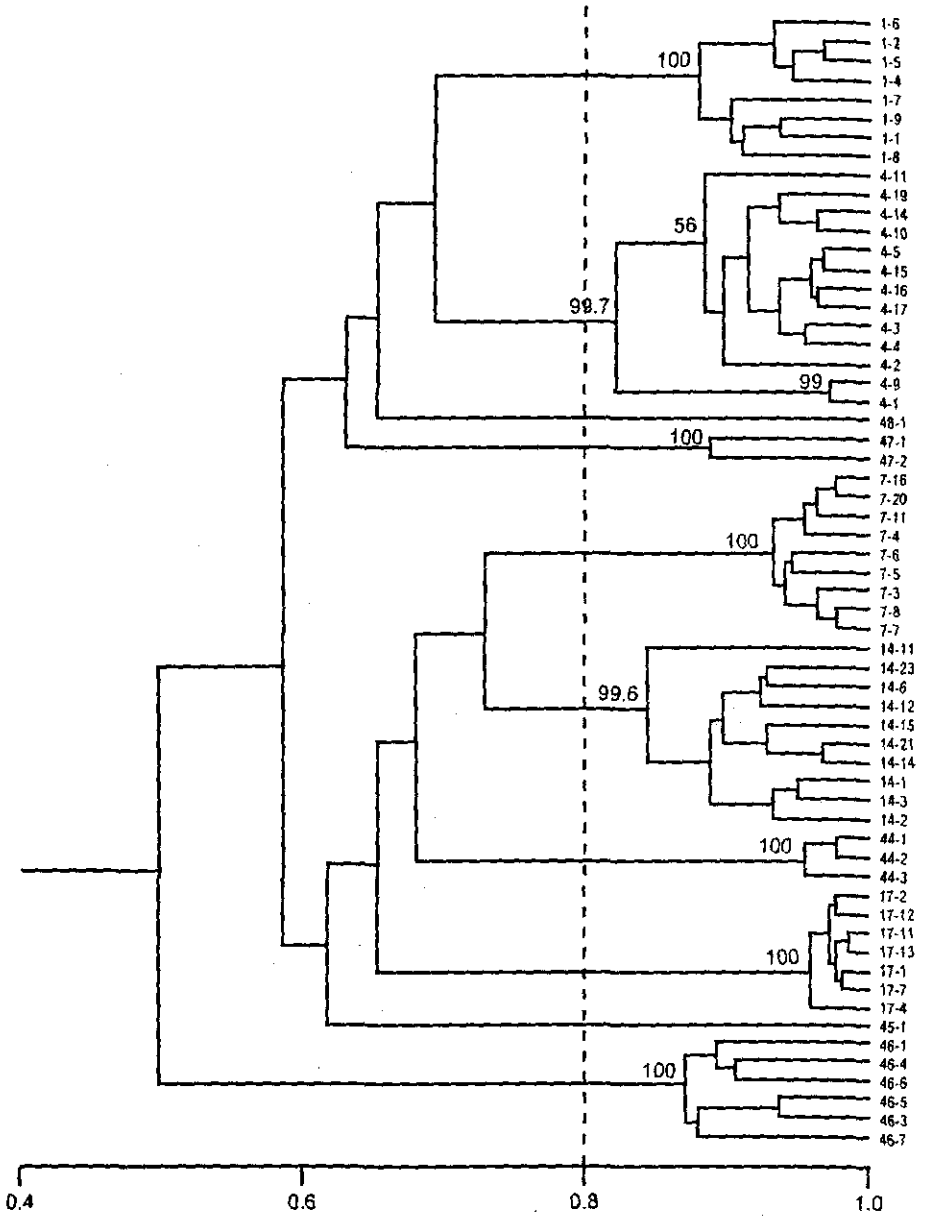


Figure 13. UPGMA dendrogram of unique haplotypes of *P. grisea* based on DNA banding patterns using the probe MGR586. Ten clusters were resolved at the 0.80 similarity level. Bootstrap *P* values are indicated at the corresponding node for each cluster (Chen et al, 1995).

References

Brown JKM. 1994. Bootstrap hypothesis tests for evolutionary trees and other dendrograms. Proc. Natl. Acad. Sci. USA 91:12293-12297.

Chen D, Zeigler RS, Leung H, Nelson RJ. 1995. Population structure of *Pyricularia grisea* at two screening sites in the Philippines. Phytopathology 85:1011-1020.

Efron B. 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7:1-26.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39: 783-791.

Hedges SB. 1992. The number of replications needed for accurate estimation of the bootstrap *P* value in phylogenetic studies. Mol. Biol. Evol. 9:366-369.

Nelson RJ, Baraoidan MR, Vera Cruz CM, Yap IV, Leach JE, Mew TW, Leung H. 1994. Relationship between phylogeny and pathotype for the bacterial blight pathogen of rice. Appl. Environ. Microbiol. 60:3275-3283.

Rohlf FJ. 1989. NTSYS-pc Numerical taxonomy and multivariate analysis system. Setauket, NY: Exeter Publishing Co, Ltd.

Sokal RR, Sneath PHA. 1963. Principles of numerical taxonomy. San Francisco: W.H. Freeman and Company. 359 pp.

Appendix 1. Similarity Coefficients

For binary data such as DNA fingerprints (banding patterns), in a data set with n band positions, the data for each pair of OTUs i and j can be arranged as in this 2 × 2 table:

OTU j \ OTU i	1	0	
1	a	b	n_{j1}
0	c	d	n_{j0}
	n_{i1}	n_{i0}	n

where:

a = the number of positions in which both i and j have a band (positive matches)

b = the number of positions in which j has a band but not i

c = the number of positions in which i has a band but not j

d = the number of positions in which neither i nor j have a band (negative matches)

n_{i1} = $a + c$; the number of bands present in i

n_{i0} = $b + d$; the number of bands absent in i

n_{j1} = $a + b$; the number of bands present in j

n_{j0} = $c + d$; the number of bands absent in j

n = $a + b + c + d$; the total number of band positions

Two additional variables can also be defined:

m = $a + d$; the number of matching bands

u = $b + c$; the number of "unmatches"

Using this arrangement, Sokal and Sneath (1963) define several coefficients. WinBoot can compute for all of these similarity coefficients; these are listed in Table 1. Note that not all of these coefficients are appropriate for binary data derived from banding patterns; specifically, those coefficients that include the "negative matches" d. On a technical note, not all of these coefficients have

been fully tested and you may get unpredictable results when using them. The most thoroughly tested coefficients are Simple and Dice.

Table 1. WinBoot and WinDist similarity coefficients.

Coefficient name		Coefficient formula
as used in WinBoot	**as defined by Sokal and Sneath**	
Jaccard	Jaccard	$\frac{a}{a+u}$
Simple	Simple Matching	$\frac{m}{n}$
Russell	Russell and Rao	$\frac{a}{n}$
Dice	Dice	$\frac{2a}{2a+u}$
UN1	'Unnamed' 1	$\frac{2m}{2m+u}$
UN2	'Unnamed' 2	$\frac{a}{a+2u}$
Rogers	Rogers and Tanimoto	$\frac{m}{m+2u}$
Kulczynski1	Kulczynski 1	$\frac{a}{u}$
UN3	'Unnamed' 3	$\frac{m}{u}$
Kulczynski2	Kulczynski 2	$\frac{1}{2}\left(\frac{a}{n_{i1}}+\frac{a}{n_{j1}}\right)$
UN4	'Unnamed' 4	$\frac{1}{4}\left(\frac{a}{n_{i1}}+\frac{a}{n_{j1}}+\frac{d}{n_{i0}}+\frac{d}{j_{j0}}\right)$
Ochiai	Ochiai	$\frac{a}{\sqrt{n_{i1}n_{j1}}}$
UN5	'Unnamed' 5	$\frac{ad}{\sqrt{n_{i1}n_{j1}n_{i0}n_{j0}}}$
Hamann	Hamann	$\frac{m-u}{n}$
Yule	Yule	$\frac{ad-bc}{ad+bc}$
Phi	Phi	$\frac{ad-bc}{\sqrt{n_{i1}n_{j1}n_{i0}n_{j0}}}$

Appendix 2. WinDist

WinDist is an additional program included with WinBoot. It uses the same input file (in the same format) as WinBoot to produce an output file with a similarity matrix. Both programs use the same code (in the form of the BANDS.DLL dynamic-link library) to compute for the similarity matrix. The file containing the similarity matrix can then be used to perform cluster analysis. The interface of WinDist is similar to that of WinBoot. WinDist can produce two types of output file, one formatted for use with PHYLIP and the other for NTSYS. For NTSYS-format files, you have the option of specifying whether to produce a similarity or a distance matrix. For PHYLIP-format files, you can only produce a distance matrix, since this is the only type that PHYLIP will accept. The distance coefficient d is computed by subtracting the similarity coefficient S from 1:

$$d = 1 - S$$

Appendix 3. Troubleshooting

The most common type of error you will encounter when running WinBoot will probably be due to improper formatting of the input file. The program will stop and show you when this occurs and allow you to correct the problem. More rarely, you may encounter run-time errors, which occur while the program is computing. These are more severe errors that immediately halt the program.

Formatting errors

The following messages indicate a problem in the formatting of the input file:

Invalid parameter line.

Line #*x*. Name is invalid.

Illegal character in line #*x*: "*c*".

Unexpected end of file.

These occur after you have pressed the Compute button, while the program is reading the input file. WinBoot will immediately halt and show a dialog box with one of the above messages and the approximate location of the error. You can then edit the file using a text editor such as Notepad. WinBoot is still running, with all its settings intact, allowing you to simply click "Compute" again after you have corrected the problem.

If the error is not immediately obvious, consult the following checklist to try and correct the error:

- *Count your OTUs and characters carefully.* Check the parameter line, the first line of the input file. It should have two numbers. The first number should indicate the number of OTUs (e.g., isolates), while the second number should indicate the number of characters (e.g., bands).
- Did you save the file in tab-delimited from Excel (or any other spreadsheet program)? If so, did you select the **Tab** option from the **Input file format** list box?
- If the **Input file format** is **PHYLIP**, then there should be at least one space (not a *tab*) between the two parameters on the first line.
- If the **Input file format** is **Tab**, then there should be a tab character (not a space) between the two parameters on the first line.
- Check to see if there are any blank cells. To indicate missing values, use a question mark (?).

- The only valid characters, aside from the OTU labels, are a '1' for character [band] present, '0' for character absent, and '?' for missing data.
- For files in the **PHYLIP** format, the OTU label must be ten, and only ten, characters long. If the label is too short, fill in the remainder with spaces. If the label is too long, you must truncate it.

Run-time errors

These are more severe errors that occur while WinBoot is doing its computations and cause an abnormal program termination. The most commonly encountered is:

Run-time error 207

As explained in the section "Random Number Seed", this is caused by the program executing an undefined mathematical operation, such as division by zero. Running the program with a different random number seed or a different similarity coefficient may remedy the problem.

Other errors

WinBoot has not been thoroughly tested. If you should encounter other errors not mentioned here that you cannot resolve, please report to us the exact circumstances under which the error occurred: CPU type, DOS version number, Windows version number, amount of RAM, and the data file. Also indicate the similarity coefficient, the number of samples (bootstrap iterations), and the random number seed used.

PROGRAMA *SAS* PARA ANÁLISIS DE CORRESPONDENCIA Y DE SIMILARIDAD CON DATOS BINARIOS

```
option ls=130 ps=50;  /* FORMATO PARA IMPRESION DE RESULTADOS */

 * SECCIÓN 1;
 *-------------------------------------------------------------------;
 *                          ENTRADA DE DATOS                         ;
 *-------------------------------------------------------------------;

data a ;
infile '~/cluster/antonia/ejemplo.dat' lrecl=165; /* UBICACIÓN DEL ARCHIVO */
input isolate $ 1-18  @19(a1-a142)(1.); /* DEFINICIÓN DE VARIABLES */
proc sort; by isolate;

 * SECCIÓN 2;
 *-------------------------------------------------------------------;
 *               ANÁLISIS DE CORRESPONDENCIA MULTIPLE                ;
 *-------------------------------------------------------------------;

proc iml ;
use a ;
read all var _num_ into m ;
read all var _char_ into isolate ;
colnm={counter};
isos=nrow(m);
rownum=j(isos,1,0);
r=1;
do while (r<=isos);rownum[r,1]=r;r=r+1;end;
m=t(m) ;    s=m[+,] ;    v=ncol(m) ;
vv=compress(char(ncol(m)));
call symput ('nvar',vv);

/* COEFICIENTES DE  SIMILARIAD Y DISTANCIA */
   simple    =j(v;v,0) ;
   nei       =j(v,v,0) ;
   jacc      =j(v,v,0) ;
   dist      =j(v,v,0) ;

do i = 1 to v ;
  do j =i+1 to v ;
     aux=m[,i] + m[,j] ;
     a=ncol(loc(aux=2)); b=s[i]-a; c=s[j]-a; d=ncol(loc(aux=0));
     b2=s[i]; c2=s[j];
     simple [i,j]=(a+d)    /(a+b+c+d) ;
     nei    [i,j]=2*a      /(b2+c2)   ;
     jacc   [i,j]=1-(a/(a+b+c))       ;
     dist   [i,j]=1-(2*a/(b2+c2))   ;
  end ;
end;
do j=1 to v ;
   do i=j+1 to v ;
   simple[i,j] = simple[j,i] ;
   nei[i,j]    = nei[j,i] ;
   jacc[i,j]   = jacc[j,i] ;
   dist[i,j]   = dist[j,i] ;
```

```
 end;
end;
do j=1 to v ;
   simple[j,j] =1  ;
   nei[j,j]    =1  ;
   jacc[j,j]   =0  ;
   dist[j,j]   =0  ;
 end;
create out1 from nei     [rowname=isolate];
append from nei     [rowname=isolate];

/* nei PUEDE SER CAMBIADO EN LAS DOS LINEAS DE ARRIBA ^ POR OTRO
  COEFICIENTE */

create rows from rownum [colname=colnm];
append from rownum [colname=colnm];

/* CALCULO DE LA DISTANCIA MEDIA POR  CADA AISLAMIENTO */

av=j(nrow(jacc),1,0);
do i=1 to nrow(jacc);
       subm1=jacc[i,];
       av[i,1]=sum(subm1)/ncol(jacc);
end;
create meandis from av [rowname=isolate];
append from av [rowname=isolate];
proc sort ;by descending col1;
proc print; title'Mean distances of each isolate';

data out;merge rows out1;
drop counter;
proc corresp noprint dim=3 out=mcaout;
var col1- col&nvar;
id isolate;

data outx (type=distance); set out;

data mca1;set mcaout;keep isolate dim1 dim2 dim3;
```

```
 * SECCIÓN 3;
 *-------------------------------------------------------------------;
 *          CLUSTER CON DATOS DE CORRESPONDENCIA MULTIPLE             ;
 *-------------------------------------------------------------------;

proc cluster data=mca1  method=ave  outtree=tree std
pseudo ccc noeigen;
var dim1- dim3;
id isolate;
title 'Cluster analysis clustering statistics MCA';

 * SECCIÓN 4;
 *-------------------------------------------------------------------;
 *          GRÁFICAS DEL ANALISIS DE CORRESPONDENCIA                  ;
 *-------------------------------------------------------------------;

goptions reset=global
ftext=simplex  htitle=1 htext=.3
device=cgmhgwa gaccess=sasgastd gsfmode=replace gsfname=gsasfile;

 filename gsasfile '~jcuasque/cluster/antonia/dendro1.cgm';
/* UBICACIÓN DEL ARCHIVO PARA CARGAR LA GRÁFICA */

proc tree  data=tree   n= 7    out=groups graphics;
/* graphics OPCION PARA DIBUJAR  DENDROGRAMA */
copy isolate;
run;
/* GENERA ARCHIVO DE GRUPOS EL "n" SE PUEDE AJUSTAR */


filename gsasfile '~jcuasque/cluster/antonia/dendro2.cgm';
/* UBICACIÓN DEL ARCHIVO PARA CARGAR LA GRÁFICA */
```

```
PROC TREE data = tree PAGES=1 HORIZONTAL graphics;
```

```
/* GRÁFICA DEL DENDROGRAMA  ORIENTACIÓN VERTICAL */
copy isolate;
run;

data out2; set groups;
drop _NAME_ clusname;
proc sort data=out ;by isolate;
proc sort data=out2;by isolate;
proc sort data=mca1;by isolate;
data clst;merge out  out2;by isolate;

data tojmp;merge mca1 clst ; by isolate;
keep isolate dim1-dim3 cluster ;

data tojmp1;set tojmp;if cluster<>.;
proc sort data=tojmp1;by cluster;
 proc print uniform noobs;

goptions reset=global gunit=pct noborder
         ftext=simplex htitle=4 htext=2
 device=cgmhgwa gaccess=sasgastd gsfmode=replace gsfname=gsasfile;


data mc; set tojmp1;
   dim1=-1*dim1; dim2=dim2*-1;

     if cluster = 1   then do; shapev='balloon';   color='magenta'; end;
else if cluster = 2   then do; shapev='pyramid';   color='olive';  end;
else if cluster = 3   then do; shapev='heart';     color='red';    end;
else if cluster = 4   then do; shapev='club';      color='blue';   end;
else if cluster = 5   then do; shapev='cross';     color='green';  end;
else if cluster = 6   then do; shapev='diamont';   color='gold';  end;
else if cluster = 7   then do; shapev='cylinder';  color='sto';  end;
/*
else if cluster = 8   then do; shapev='balloon';   color='purple'; end;
*/

proc print ; var isolate cluster  shapev color; /*LISTADO DE CONVENCIONES */

filename gsasfile '~jcuasque/cluster/antonia/mca3d1.cgm';
/* UBICACIÓN DEL ARCHIVO PARA CARGAR LA GRÁFICA */

proc g3d data= mc;
scatter dim1*dim3=dim2 / shape=shapev  color=color caxis=black
ctext=black rotate=140 tilt=30    grid size=.7;
/* OPCIONES PARA GIRAR E INCLINAR EL GRAFICO; SE PUEDEN CAMBIAR ! */
run;
```

```
 * SECCIÓN 5;
 *-------------------------------------------------------------------;
 *                      ANALISIS DE SIMILARIDAD                      ;
 *-------------------------------------------------------------------;
data out;merge rows out1;
drop counter;
* proc print;
/* SI UD. DESEA IMPRIMIR LA MATRIZ DE SIMILARIDAD BORRAR EL "*"  */

 * SECCIÓN 6;
 *-------------------------------------------------------------------;
 *              CLUSTER CON DATOS DE SIMILARIDAD                     ;
 *-------------------------------------------------------------------;

proc cluster data=out   method=ave  outtree=tree std
pseudo ccc noeigen;
var col1- col&nvar;
id isolate;
title 'Cluster analysis clustering statistics similarity';

 * SECCIÓN 7;
 *-------------------------------------------------------------------;
 *              GRÁFICA ANÁLISIS DE SIMILARIDAD (DENDROGRAMA)        ;
 *-------------------------------------------------------------------;
goptions reset=global
ftext=simplex  htitle=1 htext=.3
device=cgmhgwa gaccess=sasgastd gsfmode=replace gsfname=gsasfile;

 filename gsasfile '~jcuasque/cluster/antonia/dendro3.cgm';
/* UBICACIÓN DEL ARCHIVO PARA CARGAR LA GRÁFICA */

proc tree  data=tree   n= 7    out=groups graphics;
/* graphics OPCION PARA DIBUJAR  DENDROGRAMA */

copy isolate;
run;

/* GENERA ARCHIVO DE GRUPOS EL "n" SE PUEDE AJUZTAR */
```

```
 * SECCIÓN 8;
 *-------------------------------------------------------------------;
 *      SIMILARIDAD PROMEDIO ENTRE GRUPOS Y DENTRO DE GRUPOS        ;
 *-------------------------------------------------------------------;

proc iml;
title 'Distances or similarities within and between clusters';
use clst;
read all var _num_ into mat;
clsct1=1;clsct2=1;clsct3=0;clsct4=0;row1=1;row2=2;cluster1=1;cluster2=1;
sumr=0;totr=0;alltot=0;allcnt=0;
isos=(ncol(mat)-1);print 'Numbber of isolates' isos;
clscnt=mat[,(isos+1)];
nclus=max(clscnt);print 'Number of clusters' nclus;
do while (clsct2<=nclus);
   do while (clsct1<=isos);
     clsct4=clscnt[clsct1,1];
     if clsct4=clsct2 then clsct3=clsct3+1;
     clsct1=clsct1+1;
   end;
clust_no=clsct2;isolates=clsct3;
 print 'Isolates per cluster' clust_no isolates;
 clsct1=1;clsct3=0;clsct2=clsct2+1;
end;
pnt1=isos-1;clsp=isos+1;
do while (cluster1<=nclus);
     do while (cluster2<=nclus);
          do while (row1<=pnt1);
               do while (row2<=isos);
                    if mat[row1,clsp]=cluster1 then do;
                         if mat[row2,clsp]=cluster2 then do;
                              sumr=sumr+mat[row1,row2];
                              totr=totr+1;
                    end;
               end;
                     row2=row2+1;
               end;
               row1=row1+1; row2=row1+1;
          end;
      meandist=(sumr/totr);
         print  cluster1 cluster2 meandist;

sumr=0; totr=0; row1=1; row2=2;cluster2=cluster2+1;
end;
cluster1=cluster1+1; cluster2=cluster1; isolates=0;
end;
subclus=mat[,ncol(mat)];sumclus=sum(subclus);
allsum=sum(mat)-sumclus;allcnt=isos**2-isos+(isos*mat[1,1]);
allmn=allsum/allcnt;
print 'Overall mean distance = ' allmn;
```

Muestreo de Poblaciones

M.C. Duque- CIAT

Estudio de recursos genéticos:

(Marshall DR and Brown AHD, 1975)

Exploración

↓

Clasificación → Evaluación → Utilización

↓

Conservación

Limitantes en procesos de muestreo:

- Varía de especie a especie y según el proceso que se cumpla:
  - Vía de extinción: tiempo
  - Reproducción vegetativa-amplia diversidad-dificil conservación: procesos de conservación
  - Amplia diversidad: capacidad de evaluación y usos.

Requerimiento:

Definición objetiva de estrategias de muestreo óptimas requiere conocer :

- Cuánta variación genética tiene la especie

-Distribución de la variación

-entre plantas dentro de poblaciones?

-entre poblaciones dentro de regiones?

-entre regiones dentro de áreas mayores?

Realidad: falta de información al respecto y baja capacidad de extrapolación desde otras especies (intervención humana)

Objetivo

Definir una estrategia de muestreo que permita la colección de la máxima cantidad de variabilidad genética útil para la especie de interés.

Dentro de condiciones prácticas, necesita precisar:

-Diversidad genética (parámetro a maximizar)

-Variabilidad genética útil.

Qué significa eso ???

DIVERSIDAD GENÉTICA

1. Medidas basadas en varianza genética de características cuantitativas.

2. Medidas basadas en diversidad alélica en locus de efecto cualitativo

F(riqueza alélica) ← 2.1. Número total de alelos en la población

F(riqueza alélica, frecuencia alélica y uniformidad alélica) ←

2.2. Proporción de heterozigotos (Si hubiese un mecanismo de reproducción aleatoria)

$$H = 1 - \sum_{i=1}^{k} p_i^2$$

k: número de alelos, p_i=frecuencia del alelo i

2.3. Indice de diversidad general de Shannon-Weaver

$$H' = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$H' = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Nota: Calculado sobre base e (natural) o base 2 con el fin de dar información por individuo.

Requisitos para medidas de Diversidad Genética:

•Medir realmente la Diversidad Genética (`2.1 :?, sólo mide la variabilidad de lo expresado fenotípicamente)

•Debe ser función del número de alelos en la población.

Problema: "Conservacionistas" desean tener al menos una copia de cada alelo, pero no muestras que hablen de las frecuencias alélicas, por lo tanto la Riqueza Alélica es mas importante que la uniformidad alélica.

Conclusión: 2.1. en procesos de Exploración y Conservación.

VARIABILIDAD GENETICA UTIL

Potencialmente, el número de estados alélicos a nivel de un locus puede ser infinito. A nivel de varios loci, buscarlos puede ser tarea irrealizable!

Kimura & Crow (Alelos neutrales):

En poblaciones grandes hay muchos alelos en bajas frecuencias y pocos (2-4) en frecuencias intermedias o altas.

Alelos sobredominantes: incrementa mucho el número efectivo de alelos con relación a los neutrales.

Con sobredominancia moderada hay un mayor número de alelos con frecuencia media.

Bajo condiciones de endogamia, es dificil mantener muchos alelos a bajas frecuencias. Se encuentran mas o menos 2 alelos a frecuencias medias.

En términos prácticos:

Alelos comunes: frecuencia > 0.05 Usualmente ≅ 4

Alelos raros : frecuencia 0.05 Muchos

Además:

•Con alelos comunes, ampliamente distribuidos ..."También caerán". No se necesita estrategia especial.

•Con alelos raros, ampliamente distribuídos, la probabilidad de captura es función del tamaño de la muestra, no de la estrategia de muestreo.

•Con alelos comunes, localmente restringidos se requiere estrategia de muestreo clara.

•Con alelos raros, localmente restringidos... Vale la pena preguntarse si son deletéreos....e invertir esfuerzos en su detección.

Se sugiere:

Redefinir la búsqueda de recursos genéticos a tomar al menos una copia de cada alelo que ocurra en la población al menos con una frecuencia superior al 5%

Queda pendiente tratar el tema de **Estrategias de Muestreo**

- 1. Plantas Anuales cultivadas:

  Si la semilla viene de orígenes diferentes tomar muestras de campos individuales y hacer un agregado.

  Si la semilla se origina en un único proveedor, se toma su área de influencia como zona de muestreo y se procede de igual manera.

2. Plantas silvestres o herbáceas:

Determinar el área posible de muestreo teniendo en cuenta que tan frecuentemente se muestra marcada diferenciación genética, aún en áreas reducidas.

El investigador decide entre:

Usar Muestreo Aleatorio Simple

Definir subpoblaciones y muestrearlas separadamente.

Importan !!!:

- La densidad de plantas
- Mecanismos reproductivos
- Dispersión de polen y semillas
- Heterogeneidad ambiental

El tamaño óptimo de muestra se define como el número de individuos (plantas, semillas, muestras en general ...), para obtener con una confianza de (95% ??), los alelos (con frecuencia superior al 5% ??) presentes en la población objetivo.

Ejemplo:

Hay en una población dos alelos, A_1 y A_2, con frecuencia p_1 y p_2 respectivamente.

La probabilidad de que una muestra aleatoria de n gametos contenga al menos una copia de cada alelo es:

$$P(A_1, A_2) = 1 - (1-p_1)^n - (1-p_2)^n + (1-p_1-p_2)^n$$

Con p_1=0.95 y p_2=0.05 se necesitan 59 gametos para tener al menos una copia de cada alelo, con confianza del 95%

En términos generales se recomienda tomar, en forma aleatoria, entre 50 y 100 gametos aleatoriamente.

Igualmente, es aconsejable tomar un mayor número de muestras aleatorias en sitios altamente polimórficos por variante observada pues incrementa en mayor forma la variabilidad total que colectar variantes morfológicas escasas en cada sitio.

Se ha encontrado que el patrón de diferenciación genética dentro de especies está frecuentemente relacionado con la heterogeneidad ambiental, por lo tanto debe cubrirse un amplio rango ambiental.

Con poca o ninguna información sobre la variación en la naturaleza

- Se recomienda tomar entre 50 y 100 individuos por sitio
- Tomar el mayor número de sitios posibles
- Buscar que los sitios cubran el mayor rango de ambientes posible

Información preliminar podría sugerir la reducción de número de individuos por sitio y aumentar el número de sitios.

ESTUDIOS DE SISTEMÁTICA MOLECULAR

-Costosos

-Muestreo destructivo

NECESIDAD: maximizar la información obtenida por especimen

PLANIFICAR:

- Estrategia de muestreo
- Unidad muestral y de conservación(tipo de tejido adecuado)
- Almacenamiento
- Técnica de análisis molecular

Con relación a la estrategia de muestreo la reducción de errores depende de la varianza muestral.

Los datos genéticos, tales como las frecuencias alélicas determinadas por electroforesis de alozimas o RFLP siguen una distribución binomial donde:

Promedio = p

Varianza = p*(1-p)/n

Siendo p : frecuencia alélica y n: tamaño de la muestra

Para loci nuclear en poblaciones diploides, esta aproximación es adecuada si las frecuencias alélicas siguen HWE.

Las principales aplicaciones de la sistemática molecular son:

- **Estudios de estructura poblacional**
  - **Variación geográfica**
  - **Sistemas reproductivos**
  - **Heterozigocidad**
  - **Parentesco**
- Identificación de fronteras de la especie (hibridización)
- Estimación de filogenia

Cada una de ellas tiene diferentes manejos para cada fase del estudio y por esa razón la planificación es de primera importancia.

ESTUDIOS DE ESTRUCTURA POBLACIONAL

Estudios piloto para identificar marcadores genéticos (loci polimórficos)

Evaluación de la capacidad práctica y técnica del estudio

Consideraciones sobre la factibilidad del muestreo principalmente en términos de la magnitud en las frecuencias alélicas que se desea estimar y de la magnitud de los errores que se está dispuesto a aceptar.

ESTUDIOS DE FILOGENIA

Pruebas piloto para evaluar el alcance de la divergencia en el grupo de estudio y para identificar el método para análisis. Se sugiere tomar dos grupos cercanos y dos contrastantes

El número de individuos por grupo puede ser muy pequeños menos que se encuentre polimorfismo compartido entre especies en cuyo caso deben manejarse muestras mayores.

Protección de cultivos

Hei L. , Nelson R. and Leach JE (1993)

Los esquemas modernos de producción agrícola prestan creciente atención a las estrategias de protección de cultivos que ofrezcan mayor estabilidad.

Componentes básicos de ésta estrategia son:

•Mejoramiento para formas estables de resistencia

•Despliegue de variedades resistentes de tal manera que se prolongue la vida útil de las variedades.

El conocimiento de la estructura poblacional de las PLAGAS(?) puede contribuir tanto a los esfuerzos de mejoramiento como al desarrollo de estrategias para la dispersión de la resistencia

ESTRUCTURA POBLACIONAL

Se refiere a :

• La cantidad de variación genética entre los individuos de la población

•La forma como esta variación genética puede particionarse en el tiempo y en el espacio.

Además:

Analizar la estructura poblacional de las plagas y la forma como sus poblaciones responden a restricciones experimentales o naturales permite deducir mecanismos de cambio o adaptación por parte de ellas.

Tradicionalmente, en caso de patógenos, las poblaciones han sido caracterizadas por análisis de virulencia sobre variedades con diferentes genes de resistencias llamados huéspedes diferenciales.

Este enfoque (virulencia), ofrece una visión limitada frente a la verdadera diversidad entre poblaciones filogenéticas y en tal sentido los marcadores moleculares han abierto un nuevo panorama.

EVOLUCION:

Conjunto de eventos simultáneos de diversificación y diferenciación concluyentes a particionar la variación genética.

Variación Genética

- Introducida por:
  - Mutación
  - Recombinación
  - Migración
- Existente causada por:
  - Selección
  - Deriva genética aleatoria

Al caracterizar una población en su estructura se puede inferir la importancia de las diferentes fuerzas interactuando para producir EVOLUCION.

Entre las fuerzas evolutivas, la selección de hospederos posiblemente es la más importante para perfilar una estructura poblacional.

En gran parte, es de naturaleza artificial, debido a los agroecosistemas en los cuales los hospederos resistentes a enfermedades o insectos, son periódicamente incorporados imponiendo fuertes procesos de selección a nuevas virulencias.

La comprensión del efecto de la selección de los hospederos es crucial para el manejo fitosanitario de los cultivos y es una fuerza que puede manejarse a través del mejoramiento genético y de la dispersión de variedades.

La respuesta al hospedero depende de los patrones reproductivos y de la recombinación genética de las plagas: sexual o clonal.

Otro factor importante es el **flujo genético** definido en términos amplios como el intercambio de genes entre poblaciones, con lo cual se reduce la diferenciación genética.

La migración puede llegar a una rápida colonización y a una adaptación de la población invadida, llegando a producir la evolución de una nueva subpoblación, pero también puede reducir el nivel de diferenciación geográfica sin garantizar el intercambio genético posterior.

La mezcla de dos poblaciones depende del nivel de intercambio genético entre ellas. Si la recombinación es activa, las dos poblaciones pueden unirse en una nueva, de reproducción aleatoria, pero tal recombinación generalmente es rara.

La erosión genética se define como el cambio de frecuencias génicas o genotípicas por efecto del azar y así vista, es antagonista del flujo genético para la promoción de la subdivisión poblacional.

NOTA: LAS FUERZAS EVOLUTIVAS RARA VEZ ACTUAN AISLADAS

M.C. Duque- CIAT

Análisis patotípicos - Análisis de Virulencias

Determinación del espectro de virulencia usando un conjunto de variedades (n) diferenciales portando diferentes factores de resistencia.

Si hay n diferenciales, pueden detectarse 2^n prototipos (patotipos)

Problemas:

Laborioso

Lento

Sensible al ambiente

Subjetivo

Los diferenciales pueden tener alelos de resistencia en más de un gen o factores de resistencia desconocidos

(solución : use NIL´s ... Pero.... Dispone de ellas?)

La estructura poblacional inferida de la virulencia puede "quedarse corta" al reflejar la diversidad genética de los aislamientos examinados.

Por ejemplo: 2 linajes pueden tener similares o idénticos patrones patotípicos porque han sido sometidos a igual presión de selección sobre un conjunto común de hospederos, pero no, porque genéticamente sean similares.

Para evitar el sesgo, la estructura poblacional debe inferirse de marcadores neutrales distribuidos aleatoriamente en el genoma.

Los marcadores moleculares son presumiblemente neutrales....

ESTRATEGIAS DE MUESTREO

Los métodos analíticos usados deben revelar polimorfismos entre individuos y permitir que las relaciones genéticas entre individuos puedan ser deducidas eficientemente.

La mejor estrategia de muestreo de poblaciones plagas depende del patrón de disposición espacial para lo cual se requiere un reconocimiento previo de campo.

TAMAÑO DE MUESTRA

Los tamaños de muestra deben determinarse a partir de la probabilidad de que los grupos mas escasos (???) sean detectados.

F: frecuencia de la variante

P: Probabilidad de encontrar la variante (al menos una vez) en la muestra de tamaño *n*

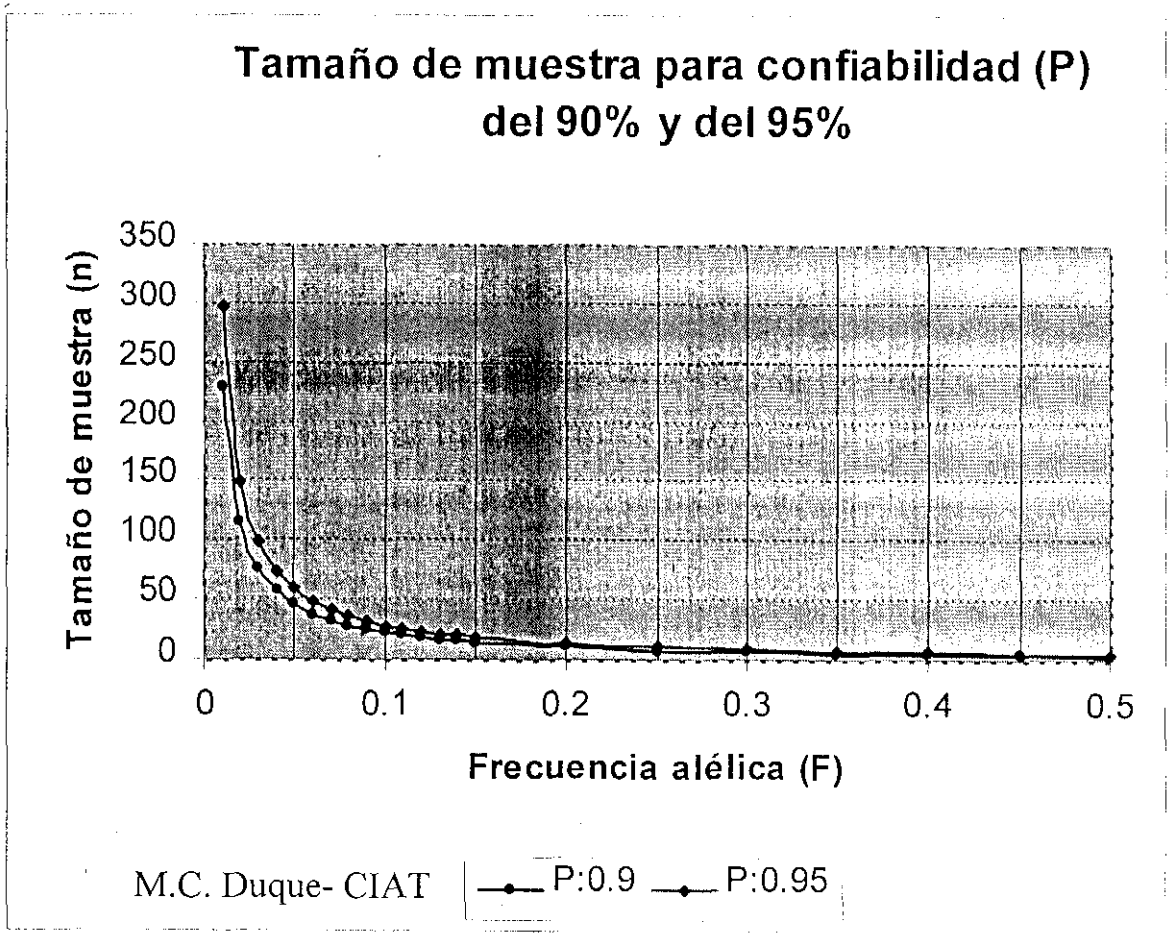
1-P = Probabilidad de no encontrar la variante

$$1 - P = (1 - F)^n,$$

por lo tanto,

$$P = 1\text{-}(1\text{-}F)^{\,n}$$

$$n = \frac{\ln(1-P)}{\ln(1-F)}$$



EL ANALISIS FILOGENETICO

La frecuencia génica y la haplotípica permiten construir árboles filogenéticos que en forma gráfica exhiben las relaciones genéticas entre individuos.

FORMAS DE CONTRUCCIÓN

- Métodos de distancia: Los datos originales llevan a matrices de similaridad o distancia

- Métodos de Máxima Parsimonia: los datos son usados directamente y las rutas que requieren el mínimo número de pasos para pasar de un estado a otro son las elegidas.

Bibliografía

Avise , J.C. (1994) "Molecular Markers, Natural History and Evolution" Chapman & Hall Inc New York

Bachmann, K .(1994)"Molecular Markers in Plant Ecology". New Phytol 126:403-418

Bavertock Peter and Moritz Craig 1996 "Project Design" Ch 2 In:"Molecular Systematics" Part I : Sampling. Edited by Hillis D.M, Moritz C. and Mable B.K. @nd Edition, Sinauer Associates Inc Publishers.

dos Santos J. B., J. Nienhuis , P. Scroch, J. Tivang, M. K. Slocum (1994) "Comparisson of RADP and RFLP genetic markers in determining genetic similarity among *Brassica oleraceae L.*" Theor Appl Genet 87:909-915

García, JA, Duque,MC, Tohme,JM, Xu s, Levy M (1995) " Un programa SAS para analisis de Clasificación". VI Simposio de Estadistica de la Universidad Nacionals de Colombia, V Seminario de Estadistica daplicada del Instituto Interamericano de Estadistica, IV Reunión de la Red de la SociedadInternacional de Biometria para Centro America, el Caribe, Colombia y Venezuela Santa Marta, Colombia

Leung H., Nelson R, and Leach J. 1993 "Population structure in plant pathogenic fungi and bacteria" , Advances in plant pathology, vol 10 157-206

Marshall D.R. and Brown A.H.D. 1975 "Optimum sampling strategies in genetic conservation" Ch 4 In: "Crop genetic resources for today and tomorrow" Edited by Frankel O.H and Hawkws K.J.G., Cambridge University Press

Nei M., Li W (1979) " Mathematical model for studying genetic variation on terms of restriction endonucleases". Proc Natl Acad Sci USA 76:5269-5273

Nei M(1973) "Analysis of Gene Diversity in Subdivided Populations".Proc Nat Acad Sci USA Vol 70, N 12 Part I pp:3321-3323.

Khairallah, M. M., B.B. Sears and M.W.Adams (1992)" Mitochondrila restriction fragment length polymrphisms in wild *Phaseolus vulgaris L.*: insights on the domestication of the common bean". Theor Appl Genet. 84:915-922

Llaca V., A. Delgado-Salinas y P. Gepts (1994) "Chloroplasts DNA as an evolutionary marker in the *Phaseolus vulgaris* complex". Theor Appl Genet 88:646-652

Mazur , J. S. V. Tingey (1995) " Genetic mapping and introgression of genes of agronomic importance". Current Opinion in Biotrechnology 6:175-182

Powell W., M. Morgante, C. Andre, M. Hanafey, J Vogel, S Tingey, y A Rafalski

(1996) "The comparisson of RFLP, RAPD, AFLP, SSR (microsatellite) markers for germplasm analysis". Molecular Breeding 2:225-238

Rengifo J. (1997)" Caracterización de germoplasma de *Phaseolus vulgaris L.* cultivado andino por medio de AFLPS". Tesis de grado. Universidad de los Andes Facultad de Ciencias. Departamento de Ciencias Biológicas. Santafé de Bogotá Colombia

Roa AC, Maya MM, Duque,MC, Tohme J, Allem AC, Bonierbale,MW.(1997)" AFLP analysis of relationships among casavva and other *Manihot* species". Theor. Appl. Genet. 95:741-750

Rohlf,FJ (1994) "NTSYS-PC" Version 2.02 i Exeter Software, Setauket, New York

SAS Institute Inc (1990) SAS/STAT, SAS/GRAPH, Software: Reference, Version 6.12, Cary, NC:SAS Institute.

Tohme J. D.O. Gonzalez, S Beebe y M.C. Duque (1996) "AFLP analysis of gene pools of a wild bean core collection". Crop Sci 36:1375-1384 Yap,I,

Thormann C. E., M.E. Ferrira, L. E. A. Camargo , J. G. Tivang y T. C. Osborn(1994) "Comparison of RFLP and RAPD markrs to estimating genetic relationships within and among cruciferous species". Theop Appl Genet 88:973-980

Yap V. y R. J. Nelson (1996) "WinBoot: A program for performing bootstrap analysis of binary data to determine the confidence limits of UPGMA-based dendrograms" . IRRI Discussion Paper Series N14. International Rice Research Institute P.O. BOX 933 Manila, Philippines.

NTSYS -PC

NUMERICAL TAXONOMY AND MULTIVARIATE ANALYSIS SYSTEM

DESCRIPCION GENERAL

MYRIAM CRISTINA DUQUE E.

Consultora Estadística

CIAT

INTRODUCCION A NTSYS

1. Qué es NTSYS?

NTSYS: Numerical Taxonomy and Multivariate Analysis System. F. James Rohlf y Dennis E. Slice

NTSYS es un conjunto d programas que pretenden encontrar e ilustrar estructuras en conjuntos multivariados de datos.

2. Estructura de Trabajo en NTSYS

Módulos separados coordinados por un menú principal.

Cada módulo lee uno o varios archivos de entrada y producirá como salidas: Archivos de datos, Listados, Gráficas, que pueden consultarse por pantalla, pueden imprimirse o pueden conservarse con formato ASCII.

Definiciones:

En Biología Sistemática hay dos grandes orientaciones:

FENETICA: Busca descubrir y describir los patrones de diversidad genética existentes en un conjunto multivariado de datos.

FILOGENETICA: Pretende encontrar la historia evolutiva o árbol filogenético de los individuos considerados.

Requiere métodos que tomen en cuenta el modelo supuesto de evolución.

La **FILOGENETICA** parte de la base de que la mejor explicación biológica a una determinada situación de diversidad proviene de su historia evolutiva.

NTSYS Cumple con un enfoque **FENETICO**

PAUP, PHYLIP, MEGA ofrecen un enfoque **FILOGENETICO**

PASOS PRINCIPALES PARA EL ANALISIS DE DATOS CON NTSYS

1. Matriz con información de los individuos u OTU's en las diferentes variables evaluadas.

(OTU : Operational Taxonomic Unit)

2. Cálculo de medidas de similaridad / disimilaridad entre todos los pares de individuos.

3. Resumen de la información obtenida en términos de:

- Conjuntos anidados(jerárquicos) de individuos similares : **Clasificación**

-Disposición espacial en uno o más ejes de coordenadas:

Ordenamiento

PREPARACION DE DATOS

Deben ser archivos ASCII(frecuentemente usado : creado en Excel y guardado con extensión prn)

Cada archivo tiene 4 tipos de registro:

1. *Registro de comentarios*

El primer caracter debe ser uno de los tres siguientes:

‘ , “ , `

a continuación usted puede ingresar un texto que explique el contenido.

PREPARACION DE DATOS

2. ***Registro de parámetros***

a. tipo de matriz

1=matriz rectangular de datos

2=matriz simétrica de disimilaridad

3=matriz simétrica de similaridad

4=matriz diagonal

5=matriz de árbol de disimilaridad

6=matriz de árbol de similaridad

7=matriz de gráfica de disimilaridad

8='matriz de gráfica de similaridad

PREPARACION DE DATOS

b. Número de filas (*)

c. Número de columnas (**)

d. Identificador de datos faltantes

0= No hay

1= Si hay ➔ Después de un espacio en blanco debe codificar el número que lo represente.

3. ***Registros de labels o identificadores (opcional)***

En caso de usarse labels o identificadores (etiquetas), en el registro de parámetros debe escribirse "L", "B" o "E" inmediatamente después del número de filas (*) o de columnas (**), según sea a quien correspondan los labels.

PREPARACION DE DATOS

* B: Label en la misma línea de datos y antes de ellos.

* E: Label en la misma línea de datos y después de ellos.

*, ** L: Label en lista aparte y antes de los datos.

Nota 1 : Si se usan labels (etiquetas) para filas y columnas, y ambos con modalidad "L" va primero la lista correspondiente a filas.

Cada label tendrá 16(?) caracteres o menos (sin espacios intercalados) y se separan por comas o espacios en blanco.

EJEMPLOS :

'DATOS DE PRUEBA (Reg de comentarios)

1 10L 4 0 (Reg de parámetros de una matriz rectangular de datos con 10 individuos identificados por labels en lista aparte, con cuatro variables por individuo y sin datos faltantes)

'DATOS DE PRUEBA (Reg de comentarios)

1 10B 4 0 (Reg de parámetros de una matriz

rectangular de datos con 10 individuos identificados por labels precediendo los datos, con cuatro variables por individuo y sin datos faltantes)

PREPARACION DE DATOS

1 10 4L 1 2

Matriz de datos con 10 individuos sin labels, con 4 variables, cada una de las cuales lleva su nombre. Hay datos faltantes y se identifican por 2

1 10B 4L 1 9

Matriz de datos con 10 individuos con labels antes de los datos, con 4 variables, cada una de las cuales lleva su nombre. Hay datos faltantes y se identifican por 9

1 10 4 1 5

Matriz de datos con 10 individuos y 4 variables, sin ningún tipo de labels. Hay datos faltantes y se identifican por 5.

PREPARACION DE DATOS

EJEMPLO:

Nota:" " representa al menos un espacio en blanco

1 10L 4 0

F1 F2 F3 F4

F5

F6 F7 F8 F9 F10

(Datos) .....

1 10L 4L 0

F1 F2 F3 F4 F5 F6 F7 F8 F9 F10

C1 C2 C3 C4

(Datos) .....

PREPARACION DE DATOS

4. Registros de datos

Los datos se entran por filas, con espacios en blanco o comas entre ellos.

Una línea no debe tener más de 255 caracteres (incluyendo los blancos y/o comas). Si se requiere continuar en otra debe hacerse en la siguiente línea.

El primer elemento de un individuo u OTU debe iniciar una línea.

EJEMPLO: estos son los datos :

```
1 0 0 1 1 1 0 0

0 1 0 1 0 1 1 1

1 1 0 0 1 1 0 0
```

A. Puede entrar como está ?

B. puede entrar de la siguiente manera ?

1_ 0_ 0_ 1

1_ 1_ 0_ 0

0_ 0 1_ 0_ 1_ 0_ 1

1 1

1_ 1_ 0_ 0_ 1 1_ 0_ 0

C. Es correcto lo siguiente?

1_ 1_ 0_ 1

1_ 1_ 0_ 0

0_ 1_ 0_ 1_ 0_ 1

1_ 1_ 1_ 1

0_ 0_ 1_ 1_ 0_ 0

DESCRIPCION DE PROGRAMAS

Hay muchos programas disponibles, pero se hará referencia solamente a los básicos para iniciar el trabajo.

1. OUTPUT: Permite hacer el control de inconsistencias. Convierte en tablas las matrices de entrada, utilizando los labels y permitiendo el paso a procesadores de texto.

2. SIMQUAL: Similaridad de datos cualitativos

Multiestado (5 opciones) :

Hamman, Rogers y Tanimoto, Simple y otros dos.

Binarias (11 opciones):

Dice, Jaccard, Phi, Kulcynski (2) , Rusell y Rao, Ochiai, Yule,y otros tres.

DESCRIPCION DE PROGRAMAS

3. SAHN (Sequential Aglglomerative Hierachical an Nested cluster methods)

Clasificación con 7 opciones :

Unión completa, simple, media (UPGMA), media ponderada (centroide) , unión de Spearman (Correlación);

4. CORRESP Análisis de Correspondencia

5. TREE Representación gráfica de un dendograma para impresión en papel o archivo para computador.

DOCUMENTO DE TRABAJO

UN PROGRAMA SAS PARA ANALISIS DE CLASIFICACION[1]

James A. García[2], Myriam Cristina Duque[3], Joe M. Tohme[4],
Shizong Xu[5] y Morris Levy[6]

---

[1] Modificación y adaptación de Xu, S. and Levy, M. (1991). IRRC, 1992 SAS *Handout.*

[2] CIAT, A.A. 67-13, Cali, Colombia.

[3] CIAT, A.A. 67-13, Cali, Colombia.

[4] Unidad de Biotecnología. CIAT, A.A. 67-13, Cali, Colombia.

[5] Purdue University, West Lafayette, IN, USA 47907.

[6] Purdue University, West Lafayette, IN, USA 47907.

UN PROGRAMA SAS PARA ANALISIS DE CLASIFICACION

1. INTRODUCCION

Los requisitos básicos de observación y medición de los procesos biológicos, hacen que en la investigación agrícola se consideren para análisis múltiples variables, las cuales pueden generar volúmenes de datos relativamente grandes. En algunos casos, el objetivo del estudio plantea la necesidad de agrupar individuos en clases sugeridas por los datos, de tal manera, que los individuos dentro de una de tales clases sean en alguna forma, similares entre sí, y los individuos en diferentes clases tiendan a ser diferentes.

El programa que se presenta, busca facilitar el manejo de los problemas de clasificación jerárquica sobre variables binarias, categóricas, cuantitativas o mezcla de ellas, y mediante el uso de re-muestreo generar el intervalo de confianza para cada uno de los nodos.

La presente, utiliza la versión 6.09 de SAS (Statistical Analysis System) y es una modificación y adaptación del trabajo realizado por Xu y Levy (1992).

2. MARCO TEORICO

2.1 Tipos de variables

a. Cualitativas

Son aquellas que describen una característica o atributo. Las variables cualitativas pueden ser:

a.1 Binaria, Dicótoma o Indicadora

Es aquella que puede tomar uno de dos posibles valores que denotan la presencia o ausencia de la característica (Ej.: si-no, vivo-muerto, éxito-falla). Frecuentemente se codifican con los valores 0 y 1.

a.2 Categórica

Es aquella que puede tomar un valor entre un limitado número de opciones.
Las variables categóricas pueden ser:

a.2.1 Ordinales

Cuando los niveles de la variable están ordenados de alguna manera. Ej.: malo-regular-bueno; bajo-medio-alto.

a.2.2 Nominales

Cuando no se puede establecer una relación de orden entre sus valores. Ej.: colores, regiones, especies, etc.

b. Cuantitativas

Son aquellas en las cuales todos sus valores tienen significado numérico. Las variables cuantitativas pueden ser:

b.1 Continuas

Pueden tomar cualquier valor dentro de un rango determinado. Generalmente son resultado de mediciones. Ej.: longitudes, áreas, volúmenes, pesos, tiempo, etc.

b.2 Discretas

Pueden tomar sólo algunos valores numéricos fijos y no pueden tomar valores intermedios entre ellos. Generalmente se presentan cuando se verifican conteos. Ej.: número de especies, número de colonias, etc.

2.2 Medidas de asociación: **Similaridad y distancia**

Muchas de las técnicas de análisis multivariado se basan en matrices simétricas que resumen la similaridad o diferencia (distancia) existente entre todos los posibles pares de individuos.

Las medidas de similaridad presentan un valor máximo para dos objetos idénticos y mínimo para dos objetos totalmente diferentes. Las distancias actúan a la inversa.

Hay muchas definiciones propuestas para medir la similaridad entre un par de individuos. La primera diferencia surge del tipo de variable. Para las **variables binarias** se consideran las posibilidades descritas en la Tabla 1, para lo cual se hará la siguiente definición:

		Individuo j: 1	Individuo j: 0	
Individuo i	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	n=a+b+c+d

a: Número de variables con valor 1 en los dos individuos.
b: Numero de variables con valor 1 sólo en individuo i.

c: Número de variables con valor 1 sólo en individuo j.
d: Número de variables con valor 0 en ambos individuos.
n: Número total de variables.

Tabla 1. Coeficientes de similaridad para variables binarias.

Definición	Nombre
$S_1 = \frac{a+d}{n}$	Coeficiente de concordancia simple Sokal & Michener (1958)
$S_2 = \frac{a+d}{a+2b+2c+d}$	Rogers & Tanimoto (1960)
$S_3 = \frac{2a+2d}{2a+b+c+2d}$	Sokal & Sneath (1) (1963)
$S_4 = \frac{a+d}{b+c}$	Sokal & Sneath (2) (1963)
$S_5 = \frac{1}{4}[\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}]$	Sokal & Sneath (3) (1963) *
$S_6 = \frac{a}{\sqrt{(a+b)(a+c)}} * \frac{d}{\sqrt{(b+d)(c+d)}}$	Sokal & Sneath (4) (1963)
$S_7 = \frac{a+d-b-c}{n}$	Hamann (*)
$S_8 = \frac{ad-bc}{ad+bc}$	Yule

Definición	Nombre
$S_9 = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$	Φ de Pearson
$S_{10} = \frac{a}{a+b+c}$	Jaccard (1900. 1901, 1908)
$S_{11} = \frac{2a}{2a+b+c}$	Dice (1945) Nei-Li (1979)
$S_{12} = \frac{a}{a+2b+2c}$	Sokal & Sneath (5) (1963)
$S_{13} = \frac{a}{n}$	Rusell & Rao (1940)
$S_{14} = \frac{a}{b+c}$	Kulczynski (1) (1928)
$S_{15} = \frac{1}{2}[\frac{a}{a+b} + \frac{a}{a+c}]$	Kulczynski (2)
$S_{16} = \frac{a}{\sqrt{(a+b)(a+c)}}$	Ochiai (1957)

A partir de los coeficientes de similaridad con rango entre 0 y 1 puede calcularse el valor de la distancia según las expresiones de la Tabla 2.

La similaridad en el caso de **variables categóricas** se calcula mediante la fórmula siguiente:

$$Sij = \frac{Nij}{N}$$

Donde "Nij" es el número de características en que los individuos son semejantes y "N" es el total de características evaluadas en cada individuo. Las distancias que aplican son las mismas definidas para las variables binarias.

Tabla 2. Expresiones para calcular distancia a partir de coeficientes de similaridad en variables binarias.

Definición	Nombre
$D_1 = 1 - S$	Complemento de similaridad
$D_2 = \sqrt{1 - S}$	----
$D_3 = \sqrt{1 - S^2}$	Transformación circular

* **Para las medidas S_5 y S_7, se recomienda usar solamente D_1.**

Para las **variables cuantitativas** no es necesario hacer un cálculo de similaridad para llegar al valor de distancia. Para calcular la distancia se recomienda estandarizar las variables cuantitativas. Si los datos ya vienen en esta forma el programa será informado con el parámetro STD=0. De no ser así, se hará una estandarización dividendo cada valor por el rango de la variable respectiva para que no se presenten diferencias de escala.

La distancia euclideana

$$D_{X_1X_2} = \sqrt{\sum_n (yi1 - yi2)^2}$$

es la utilizada en este programa.

Cuando hay **mezcla de variables cualitativas (binarias o categóricas) y cuantitativas** se procede así:

- -Para las variables cualitativas se calcula la similaridad y la distancia con las opciones

elegidas.

- Se hace un promedio ponderado por el número de variables, entre los valores de distancia de las binarias y la distancia euclidiana de las cualitativas.

Es de anotar que existen más formulas para el cálculo de coeficientes de similaridad, ver por ejemplo Gower (1971), Gower y Legendre (1986) y Legendre y Legendre (1986).

2.3 Tipos y métodos de clasificación

Hay diferentes tipos de clasificación:

a. Jerárquica
Tiene como objeto construir árboles o dendogramas en los que una clase mayor contiene subclases menores llamadas ramas. El árbol puede ser desarrollado ascendentemente, aglomerando desde los individuos o grupos pequeños a otros mayores o en forma descendente dividiendo sucesivamente desde el tronco a las ramas. Un grupo puede contener a otro y no hay otra forma de interacción posible.

b. No Jerárquica
Cada individuo pertenece a un único grupo. Se compara a cada individuo con las clases iniciales para asignarlo a la más adecuada.

c. Grupos translapados
Hay individuos que pertenecen a más de un grupo.

El programa que se presenta trabaja con clasificaciones jerárquicas únicamente.

Los métodos de clasificación se definen en los términos en que se calcula la distancia entre individuos o entre individuos y grupos de individuos. La Tabla 4 resume los métodos de clasificación jerárquica disponibles en el programa.

2.4 Error Estandar e Intervalo de Confianza

La distancia promedio entre dos individuos o grupos de individuos representa en el dendograma el nodo o punto donde se origina una ramificación. Dado que dicho valor es estimado a partir de una muestra y que no se conoce la población de la cual proviene, el muestreo repetido sobre la muestra original constituye una herramienta aceptable para llegar a estimar el error estándar. Este valor permitirá construir el intervalo de confianza del valor promedio de distancia en cada nodo.

Table 4. Métodos de clasificación jerárquica.

Método	Nombre
1. Ligamiento medio, UPGMA (unweighted pair-group method using arithmetic averages)	AVE
2. Centroide, UPGMC (unweighted pair-group method using centroids)	CEN
3. Ligamiento completo	COM
4. Método de la mediana de Gower WPGMC (weighted pair-group method using centroids)	MED
5. Ligamiento simple, vecino más próximo	SIN
6. Varianza mínima de Ward	WAR

Los métodos de re-muestreo como el Bootstrap, Jackknife y Special Jackknife se utilizarán en este programa para estimar el error estándar y construir el intervalo de confianza requerido.

a. Jackknife (Muestreo sin reemplazamiento)

Tukey en 1958, dio este nombre (Jackknife) a un enfoque que él propuso para prueba de hipótesis y cálculo de intervalos de confianza donde ningún otro método pudiera ser fácilmente usado.

En una muestra aleatoria de n valores $X_1, X_2, ..., X_n$ la media muestral se usa para estimar la media poblacional

$$\overline{X} = \frac{\sum Xi}{n}$$

Si la media muestral se calcula eliminando la observación "j". entonces

$$\overline{X}_{.j} = (\sum X_i - X_j)/(n-1)$$

Resolviendo para X_j a partir de estas ecuaciones se tiene que:

$$X_j = n\overline{X} - (n-1)\overline{X}_{.j}$$

Siendo posible determinar el valor muestral X_j a partir de la media global y de la media sin la observación "j".

En una situación general para un parámetro poblacional η tenemos que:

$$\hat{\eta}_j = n\hat{\eta} - (n-1)\hat{\eta}_{.j}$$

La expresión anterior permite demostrar que el promedio de estos pseudo-valores:

$$\hat{\eta} = \sum \eta_j / n$$

Es la estimación Jackknife de η.

Tratando los pseudo-valores como una muestra aleatoria de estimaciones independientes de η, la varianza del parámetro Jackknife es igual S^2/n donde S^2 es la varianza muestral de los pseudo-valores.

El intervalo de confianza para η es dado por:

$$\hat{\eta} \pm KS/\sqrt{n}$$

Donde K es el percentil de la distribución "t" con (n-1) grados de libertad al nivel de confianza apropiado.

Para el caso de clasificación de individuos evaluados sobre "N" caracteres, la muestra Jackknife contiene (N-1) caracteres pues el j-esimo caracter es eliminado cada vez, lo que

indica que el número de muestras está determinado por N y que una buena precisión no podrá alcanzarse para valores pequeños de N.

b. Bootstrap (muestreo con reemplazamiento)

Suponga que una muestra aleatoria de n valores (X_1, X_2,...X_n) tomada de una población, se usa para estimar algún parámetro η.

Los n valores observados son considerados como la mejor aproximación a la distribución poblacional de X, ya que la población verdadera es aproximadamente una población infinita en la que cada uno de los valores X es igualmente probable.

La variación muestral en el estimador η de η es determinada tomando muestras aleatorias de tamaño n sobre la muestra original.

Las muestras tomadas de esta manera son llamadas muestras Bootstrap y cada una de ellas provee una estimación η_i de η.

La manera más simple de determinar la distribución del parámetro en cuestión, es tomar gran número de muestras Bootstrap. Cien (100) muestras son suficientes para una buena estimación del error estándar según algunos autores.

La estimación Bootstrap de η está dada por:

$$\hat{\eta} = \sum \hat{\eta}_i / b$$

donde "b" es el número total de muestras Bootstrap y η_i es la estimación del parámetro en cada muestra.

La varianza del parámetro es igual S^2/n donde:

$$S^2 = \sum \frac{(\hat{\eta}_i - \hat{\eta})^2}{b - 1}$$

El intervalo de confianza $(1-\alpha)\%$ se encuentra entre los puntos $(\alpha/2)\%$ y $(1-\alpha/2)\%$, percentiles de la distribución de η_i, siendo α el nivel de confianza apropiado.

En el caso de clasificación de individuos evaluados sobre N caracteres, el muestreo aleatorio para conformar una muestra Bootstrap se efectúa sobre dichos caracteres y cada

uno de ellos tiene igual probabilidad de ser o no seleccionado. Igualmente cada muestra Bootstrap de tamaño N, por ser muestreo con reemplazamiento puede presentar un mismo caracter seleccionado por más de una ocasión.

c. **Special Jackknife (muestreo sin reemplazamiento)**

En este método, el 50% de los N valores de la muestra original son seleccionados sin reemplazamiento, es decir, un valor dado sólo puede aparecer una vez en la nueva muestra generada.

Debe seleccionarse un adecuado número de muestras para obtener una buena estimación de los parámetros. Las fórmulas de la Media y de la Varianza son las mismas empleadas con muestras Bootstrap.

Las muestras Special Jackknife al igual que las muestras Bootstrap producen un intervalo de confianza entre los percentiles $(\alpha/2)\%$ y $(1-\alpha/2)\%$ si un buen número de muestras es generado.

3. MACRO CLASIFI

La macro **CLASIFI** es un programa SAS (versión 6.09) que básicamente utiliza Macro Lenguaje SAS, SAS/IML (Interactive Matrix Language) y los procedimientos CLUSTER, TREE, MEANS y UNIVARIATE.

El procedimiento **CLUSTER** realiza el agrupamiento jerárquico de los individuos basado en la matriz de distancia que es calculada por el programa. El procedimiento dispone de 11 métodos de agrupamiento, 6 de los cuales aplican a nuestro propósito, los cuales se originan en la forma de manejo de la distancia entre individuos o grupos de individuos.

El procedimiento **TREE**, basado en los resultados del **CLUSTER** construye el dendograma que muestra la aglomeración jerárquica entre individuos. **MEANS** y **UNIVARIATE** calculan promedios, varianza e intervalos de confianza.

3.1 Parámetros

Dos grupos de parámetros suministran información a la macro para su ejecución:

a. **Parámetros posicionales:** Son parámetros que en la macro se han definido con valor nulo y en un orden específico, por lo tanto le corresponde al usuario definirlos respetando dicho orden de acuerdo a su necesidad particular.
   a.1 Obligatorios: Indispensable definirlos

ARCH: Nombre del archivo de datos que contiene información básica de cada uno de los individuos a clasificar.

IDENTIF: Nombre de la variable de identificación de los individuos. Debe ser alfabética.

TIPOVAR: Parámetro que define el tipo de cada una de las variables numéricas que caracterizan a los individuos. El tipo puede ser:

1: Si la variable es binaria.
2: Si la variable es categórica (ordinal, nominal).
3: Si la variable es cuantitativa (continua, discreta).
4: Si se presenta mezcla de variables binarias y cuantitativas.
5: Si se presenta mezcla de variables categóricas y cuantitativas.

a.2 **Opcionales:** Sólo es necesario definirlos cuando se desea calcular el intervalo de confianza de cada uno de los nodos o puntos de ramificación del árbol de clasificación.

NODOS: Número de nodos a estudiar

NODINI: Número del nodo a partir del cual se desea realizar el análisis. El programa considera como nodo número uno (1) la raíz del árbol. (MAX = # individuos -1).

NMUESTR: Número de muestras aleatorias que se desean generar para poder estimar el intervalo de confianza de cada nodo.

b. **Parámetros de palabras claves:** Son parámetros a los cuales internamente se les ha definido un valor (Default), pero que el usuario puede modificar indicando una opción válida.

METODOM: Indica el método de muestreo a usar

1: Bootstrap (Default) - Muestreo con reemplazamiento.
2: Jackknife - Muestreo sin reemplazamiento.
3: Special Jackknife - Muestreo sin reemplazamiento.

METODOC: Indica el método de clasificación que el procedimiento Cluster de SAS usará: AVE, CEN, COM. MED, SIN y WAR. Default: AVE.

Cualquiera de estos métodos puede ser usado para la producción del dendograma o lista de individuos en cada grupo, pero el análisis de cada nodo está definido con el método AVERAGE (AVE).

LCLUST: Cuando se desea listar los individuos que conforman cada uno de los grupos que se quieren formar, se define este parámetro indicando el número total de grupos deseados. Default = 0 (no imprime).

LVCUALI: Lista de variables cualitativas que participarán en el análisis.

LVCUANT: Lista de variables cuantitativas que participarán en el análisis.

IMPR: Parámetro para definir si se imprimen las matrices de distancia (todos los tipos de variables) y de similaridad (sólo en binarias y categóricas).

1: Si imprime
0: No imprime (default)

ARCHISAL: Genera archivos de salida tipo SAS (*.ssd01)

0: No genera (Default)
1: Archivo con matriz de similaridad (similari.ssd01)
2: Archivo con matriz de distancias (distance.ssd01)
3: Genera ambos archivos

STD: Parámetro para definir si es necesario estandarizar o no las variables cuantitativas.

1: Si estandariza (default)
0: No estandariza

SIMILB: Coeficiente de similaridad para variables binarias

S1: Simple (Sokal y Michener)
S2: Rogers y Tanimoto
S3: Sokal y Sneath (1)
S4: Sokal y Sneath (2)
S5: Sokal y Sneath (3)

S6: Sokal y Sneath (4)
S7: Hamann
S8: Yule
S9: Phi de Pearson
S10: Jaccard
S11: Sorensen = Dice = Nei (Default)
S12: Sokal y Sneath (5)
S13: Russell y Rao
S14: Kulczynski (1)
S15: Kulcznski (2)
S16: Ochiai

TDISTBI: Tipo de distancia a calcular con variables binarias o categóricas

1: (1- S) (S=Similaridad)
2: SQRT (1-S)
3: SQRT (1-S**2)

TDISTCU:
1: Bray - Curtis
2: Métrica de Canberra
3: Distancia taxonómica media (Default)
4: Distancia media al cuadrado
5: Distancia Euclideana
6: Distancia Euclideana al cuadrado
7: Distancia Media de Manhattan (City Block)

SALIDA: Nombre de archivo de salida del cluster que identifica el grupo a que pertenece cada individuo.

3.2. Modo de uso

% CLASIFI (ARCH, IDENTIF, TIPOVAR, NODOS, NODINI, NMUESTR, METODOM = *, METODOC = *, LCLUST = *, LVCUALI = *, LVCUANT = *, IMPR = *, ARCHISAL = *, STD = *, SIMILB = *, TDISTBI = *, TDISTCU = * , SALIDA = *);

*: Indica un valor adecuado entre las opciones de cada parámetro.

Ejemplos

a. **% CLASIFI (DATA1, IDENTIF, 3);**

Realiza CLUSTER ANALYSIS mediante el método AVERAGE a los individuos identificados por la variable IDENTIF cuya información sólo sobre variables cuantitativas está almacenada en el archivo DATA1.

b. **% CLASIFI (DATA1, IDENTIF, 3, LCLUST = 4);**

Además de lo anterior, lista los individuos de cada uno de los cuatro grupos que se desea formar.

c. **% CLASIFI (DATA1, IDENTIF, 3, 5, 2, 100);**

Realiza un análisis sobre 5 nodos, empezando con el nodo número 2, para lo cual la macro genera aleatoriamente 100 muestras tipo BOOTSTRAP (Default).

d. **% CLASIFI (DATA1, IDENTIF, 3, 5, 2, METODOM = 2);**

Igual al anterior, pero generando muestras JACKKNIFE, las cuales dependen del número de variables en que son evaluados los individuos. Con este método de muestreo, el parámetro que indica el número de muestras aleatorias no es necesario.

e. **% CLASIFI (DATA2, IDENTIF, 4, 5, 1, 100, METODOM = 3, LVCUALI = X_1-X_{10}, LVCUANT = X_{11}-X_{16}, IMPR = 0);**

En este ejemplo se efectúa un análisis de clasificación a los individuos en el archivo DATA2, identificados con la variable IDENTIF sobre una mezcla (4) de variables cualitativas (X_1-X_{10}) y de variables cuantitativas (X_{11}-X_{16}).

Se desea analizar 5 nodos empezando con el nodo número 1 y tomando 100 muestras SPECIAL JACKKNIFE sin imprimir las matrices de distancia y de similaridad.

Cuando el archivo de datos contiene gran número de individuos y se desea hacer una solución fragmentada del problema, se recomienda efectuar el siguiente procedimiento:

1. Estandarice cada una de las variables cuantitativas dividiendo cada valor por el rango correspondiente y construya un nuevo archivo de datos con dichos valores.

2. Invoque la macro CLASIFI con el objeto de generar el dendograma con todos los individuos.

3. Con base en el dendograma, seleccione del archivo de datos estandarizados los individuos que conformen una rama.

4. Invoque la macro sobre el subconjunto de datos anteriores, defina el valor de cero (0) al parámetro STD (los datos ya han sido previamente estandarizados) y analice los nodos empezando desde el 1.

De esta manera, un problema con gran número de individuos puede ser fraccionado y analizado por etapas.

Otra manera de ayudar a reducir el número de individuos originales es el aglomerar en un solo elemento, los individuos que no posean una distancia mayor a 0.10 (por ejemplo) o al nivel que el usuario crea conveniente.

4. REFERENCIAS

GENSTAT 5 Reference Manual (1987). Claredon, Oxford.

Gower, J.C. (1971). A General Coefficient of Similarity and Some of its Properties. Biometrics 27. 857-874.

Gower, J.C. and Legendre, P. (1986). Metric and Euclidean Properties of Dissmilarity Coefficients. Journal of Classification 3, 5-48.

Hall, P. and Wilson, S.R. (1991). Two Guidelines for Bootstrap Hypothesis Testing. Biometrics 47. 757-762.

Legendre L. and P. Legendre (1984). Ecologie Numérique 2. La Structure des données écologique. Collection d'écologie 13, 2a. ed. Masson, Presses de l'université du Québec, Québec, Canada.

Liu, R.Y. and Singh, K. (1992). Efficiency and Robustness in Resampling. The Annals of Statistics. 20 (1): 370-384.

Manly, B.F.J. (1992). Workshop Notes on Computer Intensive Statistics. III Simposio de Estadística. Universidad Nacional de Colombia. pp 1-20.

Pigeot, I. (1991). A Jackknife Estimator of a Combined Odds Ratio. Biometrics 47, 373-381.

SAS Guide to Macro Processing, Version 6, Second Edition, Cary, NC: SAS Institute Inc.

SAS/IML Software: Usage and Reference, Version 6, First Edition, Cary, NC: SAS Institute Inc. 1989. 501 pp.

SAS Procedures Guide, Version 6, Third Edition, Cary, NC: SAS Institute Inc., 1990. 705pp.

SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 1 and 2, Cary, NC: SAS Institute Inc., 1989. 943pp y 846pp.

Xu, S. and Levy, M. (1992). SAS Programs for Inferring Phylogenies. Working Document. Purdue University. IRRC SAS Handout.

Palmira, Julio 5 de 2001

Señores
Fondo para el Desarrollo del Recurso Humano
CIAT

El pasado 15 de Junio terminó el entrenamiento adelantado por mí en North Carolina State University, atendiendo al evento "Summer Institute in Statistical Genétics".

Anexas van las fotocopias del certificado de asistencia al curso y de los recibos de matrícula y hotel, rubros a los cuales se aplicará el auxilio ofrecido por ustedes.

Inscripción al curso	$US 1500
Alojamiento Mayo 28- Junio 15	$US 734
Perdiem Mayo 28-Junio 16	$US 880
Total	$US 3114

Deseo expresarles mi total satisfacción por la calidad de los cursos recibidos, no sólo por la instrucción ofrecida sino por la motivación enorme para continuar profundizando en éstse tipo de temas.

Adicionalmente deseo, en forma muy enfática, agradecer el apoyo brindado por ustedes para poder aprovechar ésta magnífica oportunidad.

Atentamente,

Myriam Cristina Duque E.

Myriam Cristina Duque E.
Consultora Estadística

c.c. Dr Joe Tohme

WinBoot:

A program for performing bootstrap analysis of binary data to determine the confidence limits of UPGMA-based dendrograms

Developed by Immanuel V. Yap and Rebecca J. Nelson

The authors may be contacted via the Internet at these e-mail addresses:
i.yap@cgnet.com (for technical support)
r.nelson@cgnet.com (for distribution/general inquiries)

1996

IRRI
INTERNATIONAL RICE RESEARCH INSTITUTE
P.O. Box 933, Manila 1099, Philippines

Contents