

**UTILIZACIÓN DE MÉTRICAS RIEMANNIANAS
EN ANALISIS DE DATOS MULTIDIMENSIONALES
Y SU APLICACIÓN A LA BIOLOGÍA**

JOSE M^a OLLER SALA

BARCELONA, 25 de NOVIEMBRE de 1982.

Dendograma de las funciones de densidad condicionadas a cada una de las pautas que efectúa el pasivo (PAS).

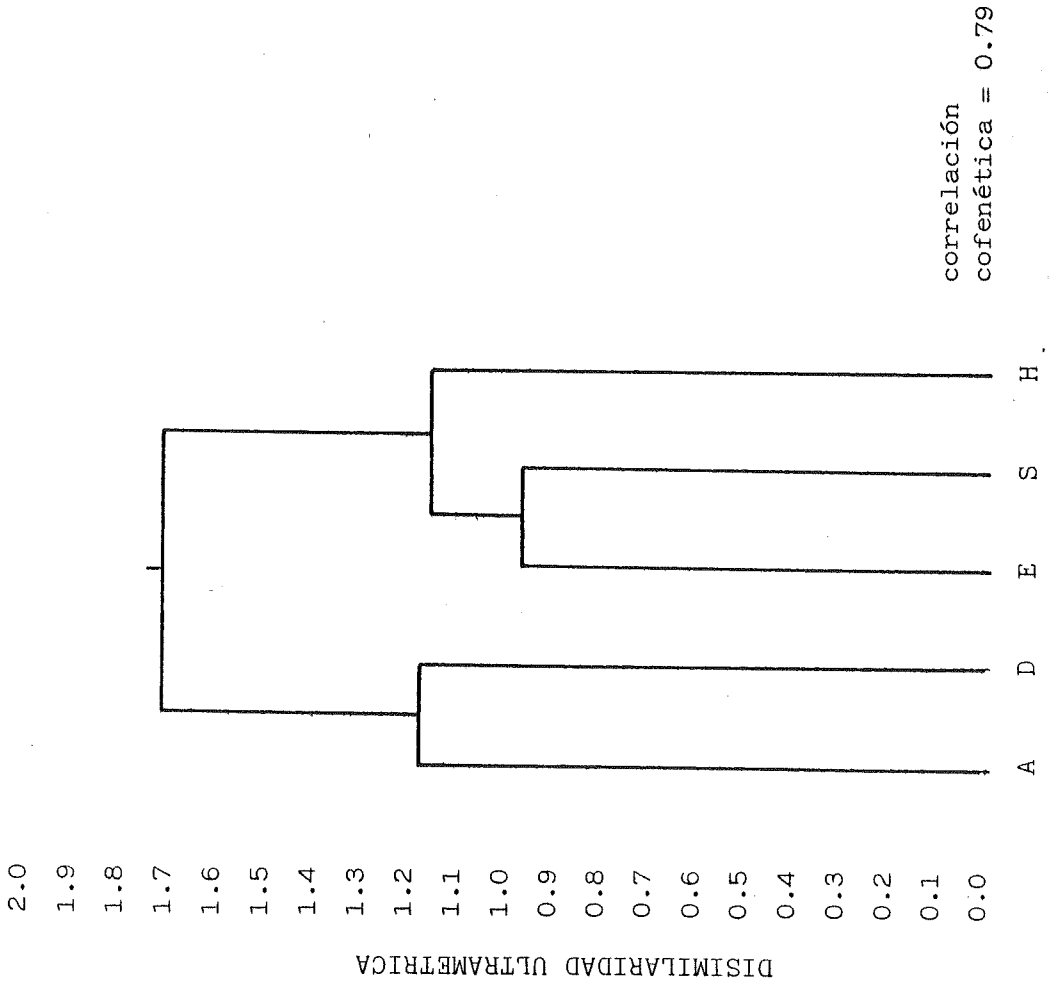
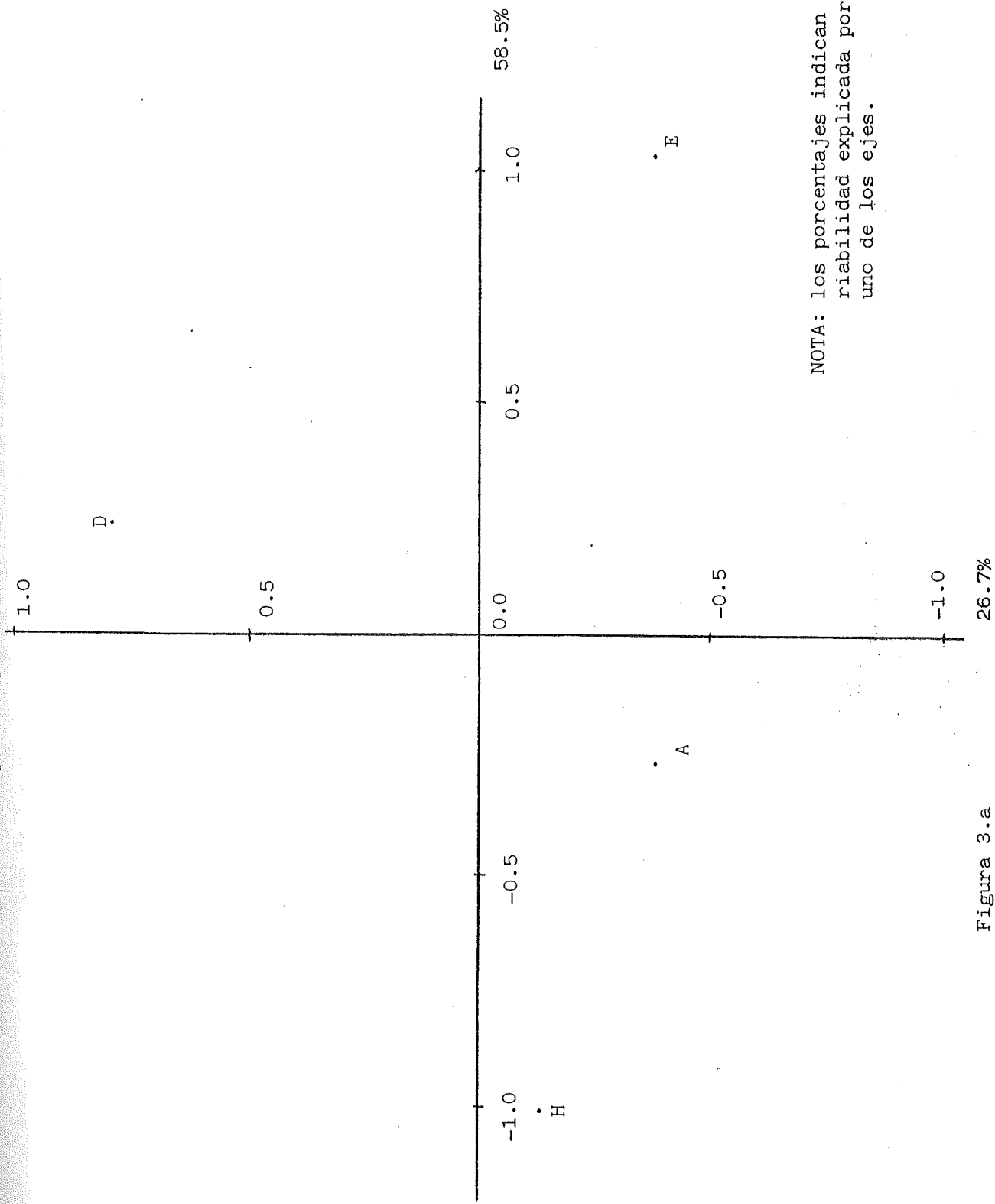


Figura 2.b



NOTA: los porcentajes indican la variabilidad explicada por cada uno de los ejes.

Figura 3.a 26.7%

Dendograma de las funciones de densidad condicionadas a cada una de las pautas que contesta el activo (ACT2).

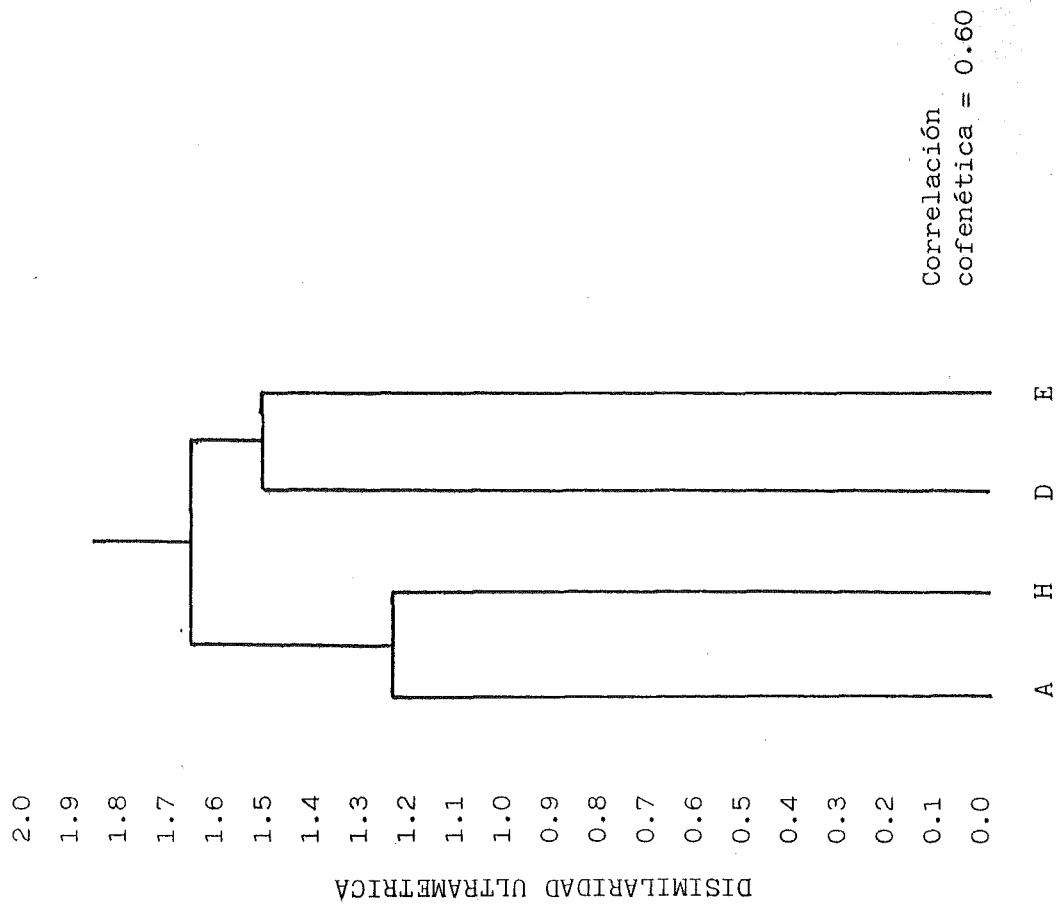
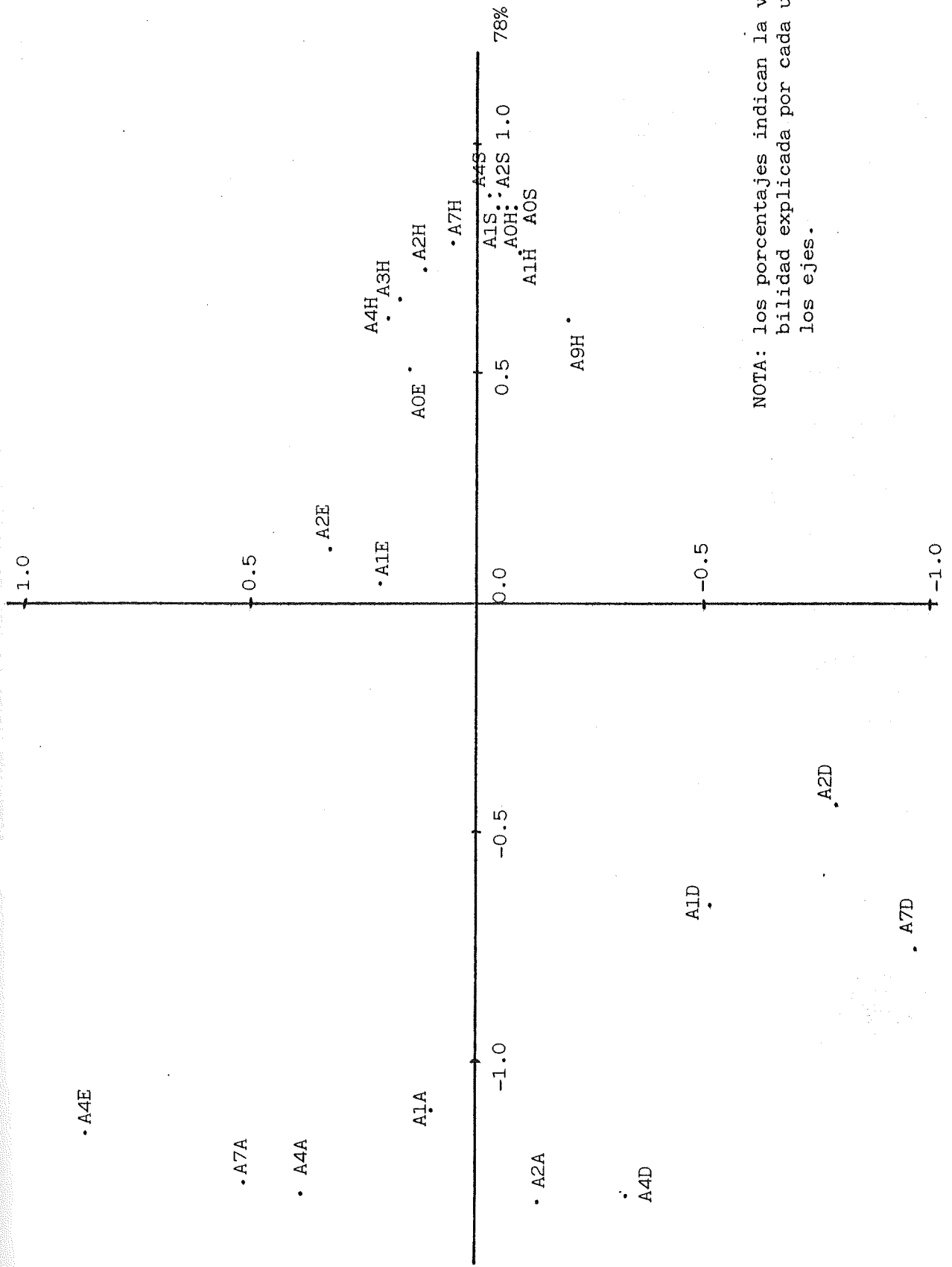


Figura 3.b



NOTA: los porcentajes indican la variabilidad explicada por cada uno de los ejes.

Figura 4.a 15.6%

Dendograma de las funciones de densidad condicionadas a las combinaciones de pautas ACT1 x PAS.

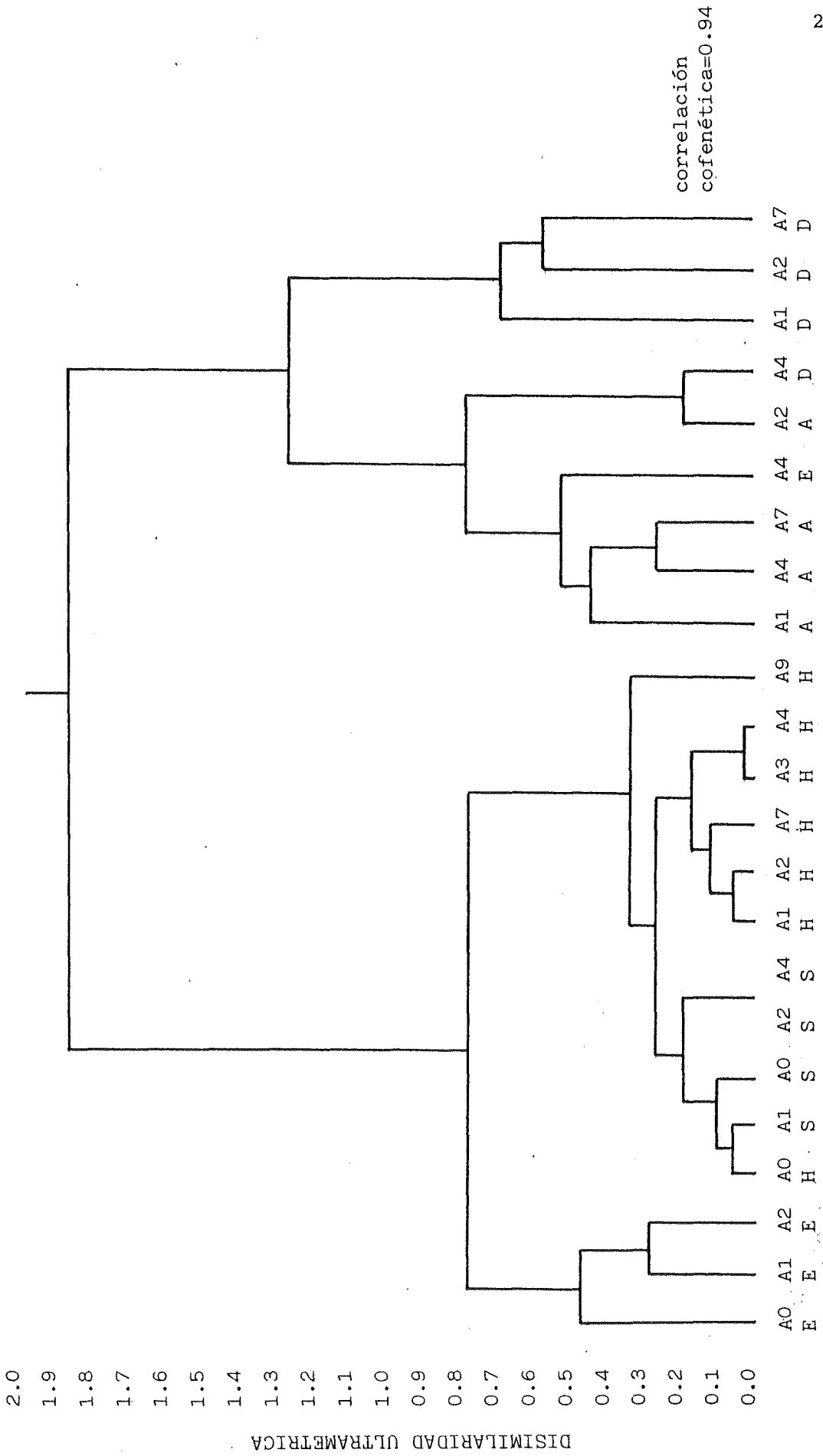
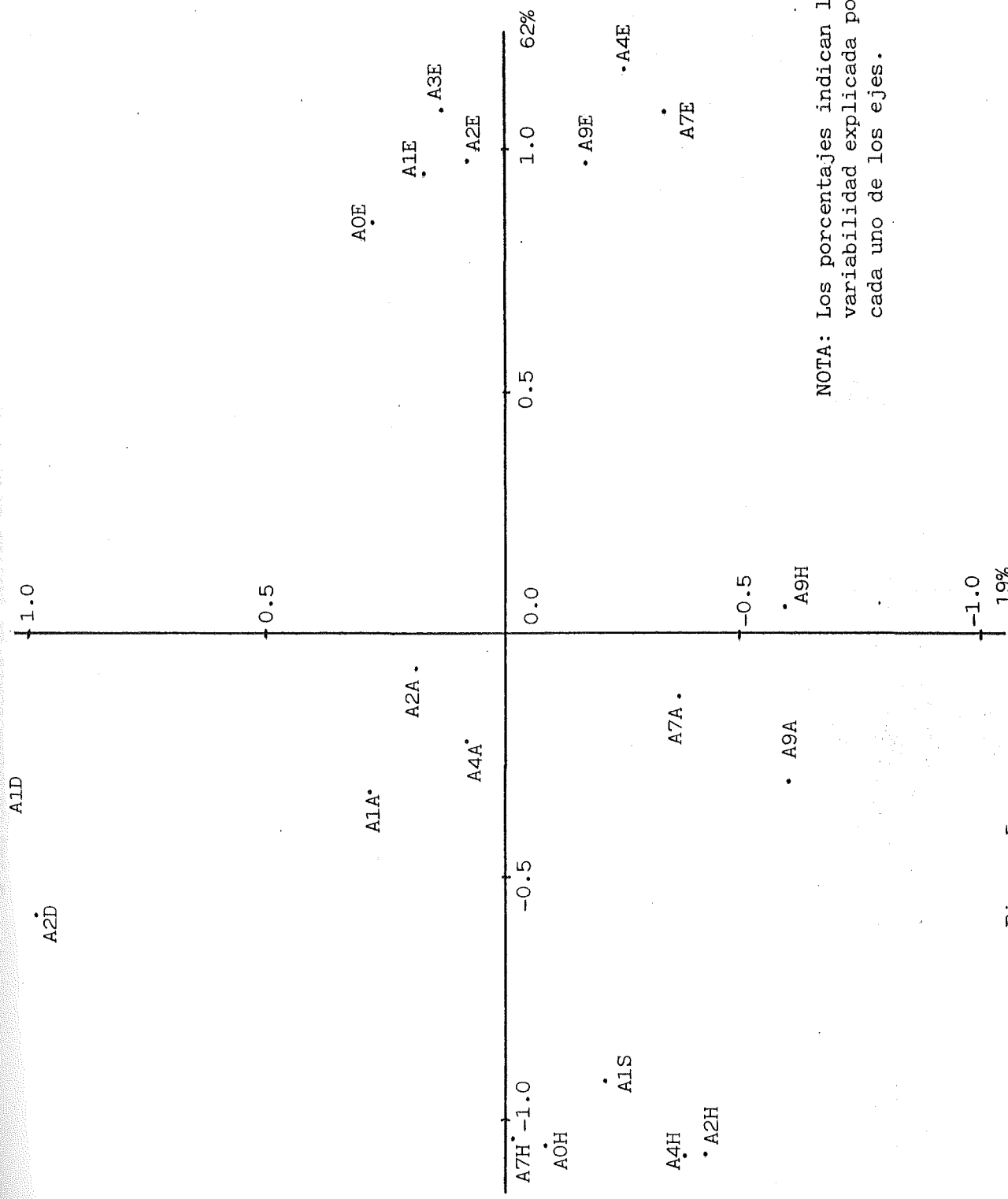


Figura 4.b



NOTA: Los porcentajes indican la variabilidad explicada por cada uno de los ejes.

Figura 5.a

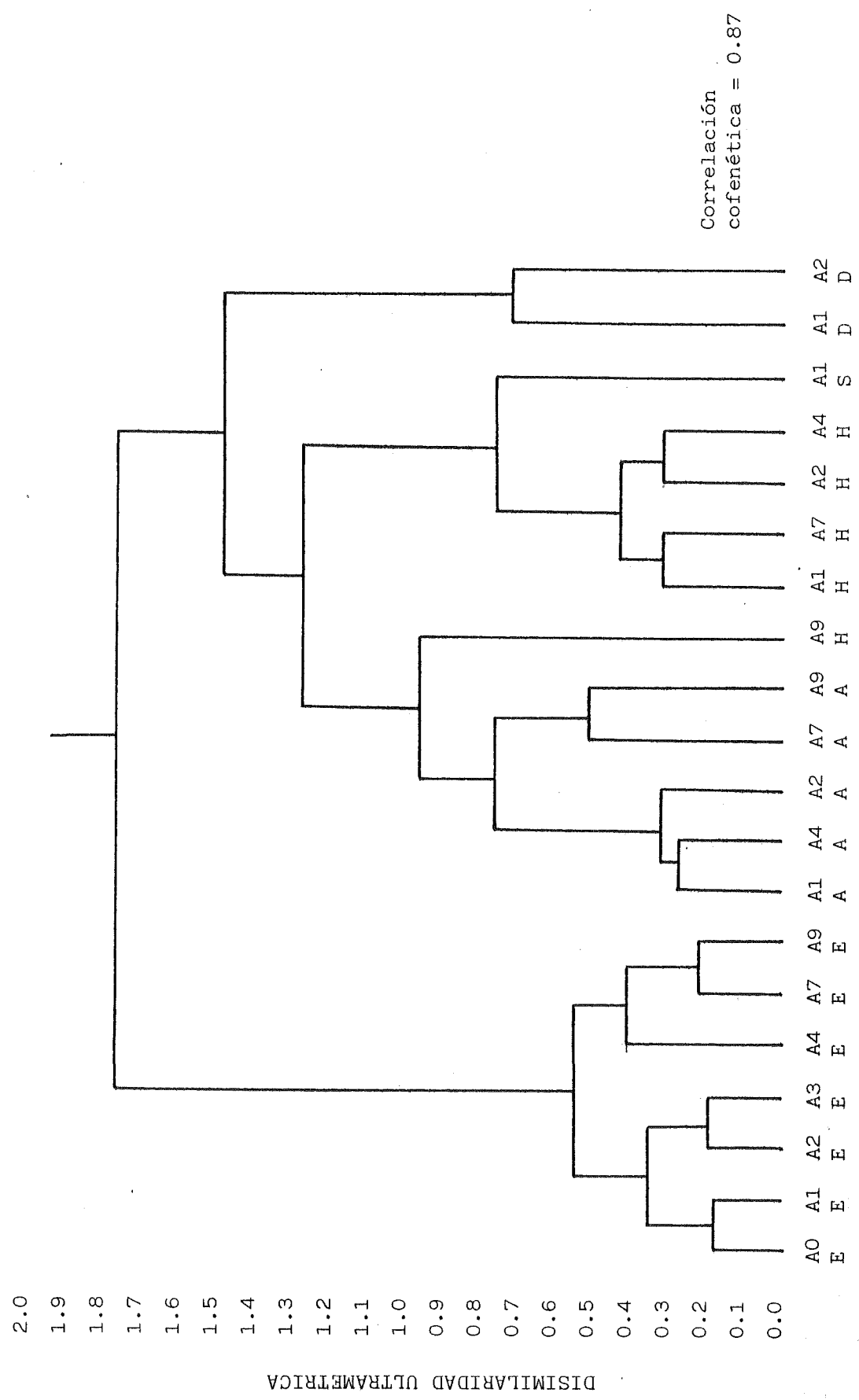
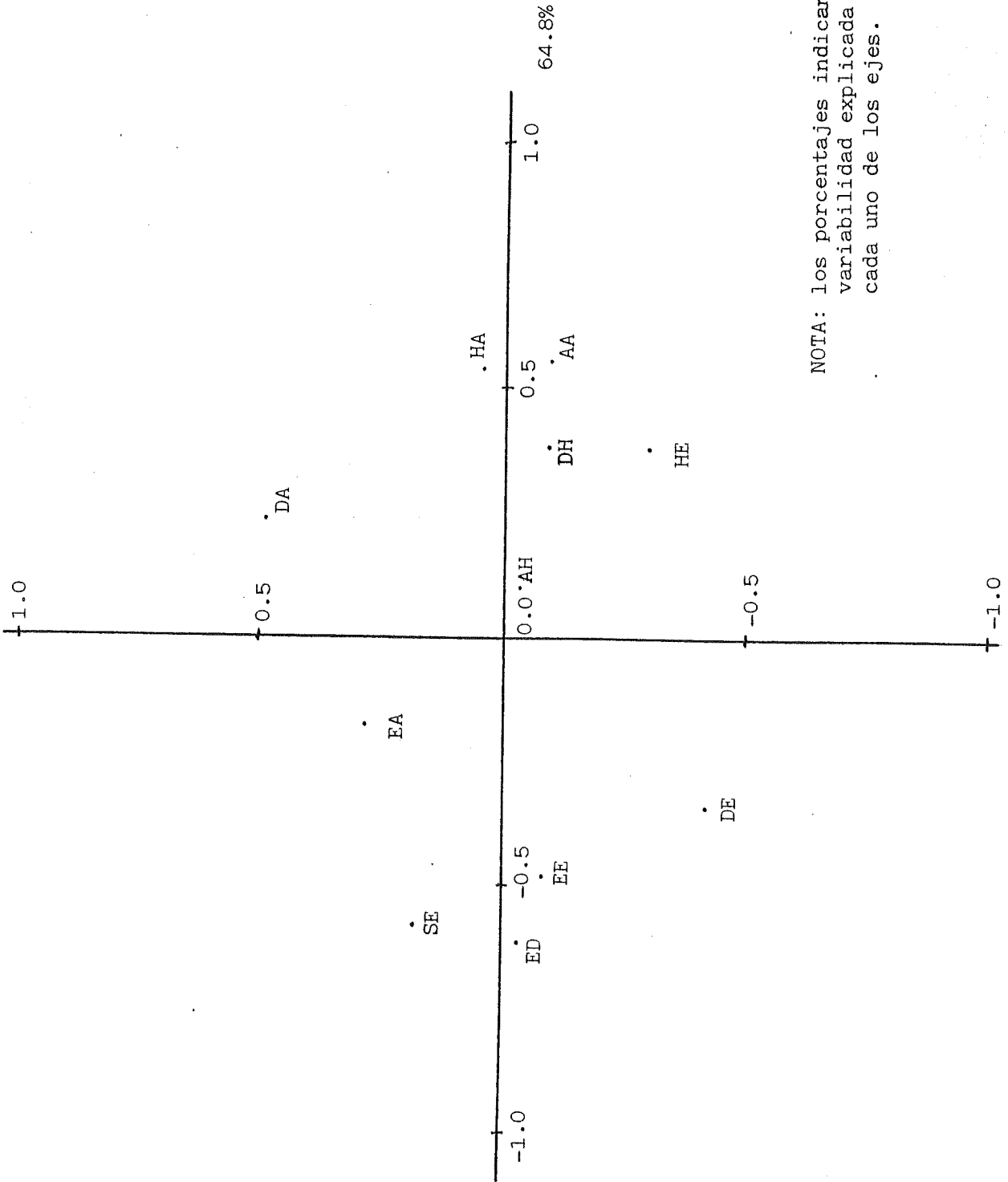


Figura 5.b

a las combinaciones de pautas PAS x ACT2.



NOTA: los porcentajes indican la variabilidad explicada por cada uno de los ejes.

Figura 6.a

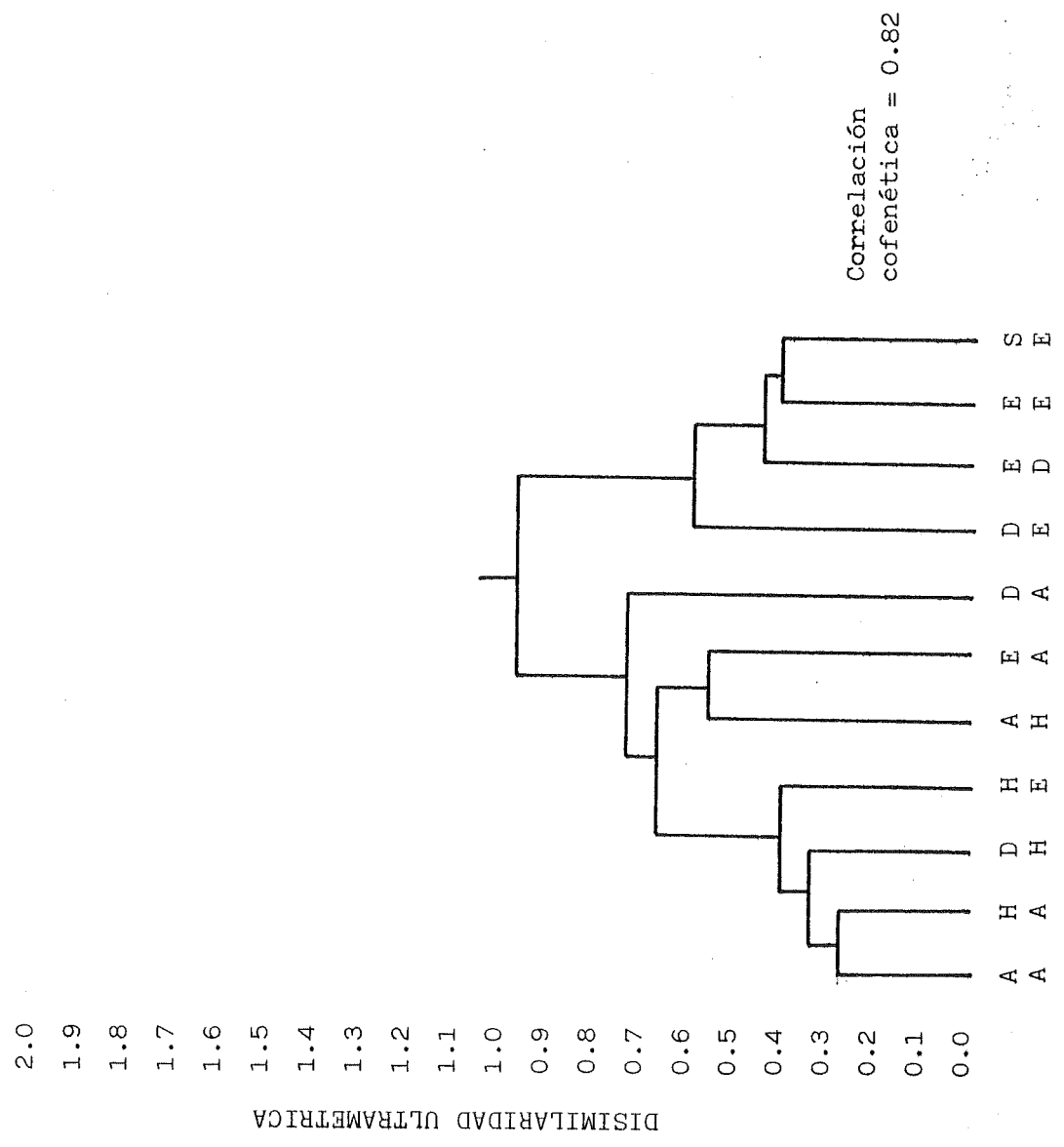


Figura 6.b



7.4. DISCUSION

A continuación vamos a intentar interpretar los resultados con las técnicas de representación de datos empleadas y con la ayuda de la tabla I, sin pretender realizar un estudio exhaustivo, que se deja para especialistas en la materia.

En las figuras 1a y 1b se representan las pautas que inicialmente efectúa el activo (ACT1) en función de la respuesta que provoca al pasivo (PAS) y la réplica de sí mismo (ACT2). En la figura 1a, las pautas han quedado ordenadas tanto más a la izquierda cuantas más huidas del pasivo ha provocado la pauta del activo (ACT1), mientras que se han dispuesto hacia arriba las pautas iniciales del activo que provocan más "estar" en la respuesta del activo al pasivo (ACT2) y en cierto modo son más eficientes.

Podríamos considerar que en el comportamiento agonístico hay, entre otras, dos factores o componentes, una es la componente de "acción" o movimiento hacia el oponente, otra es la componente de aumento de tamaño aparente. El eje horizontal reflejaría la "acción" con valores crecientes hacia la izquierda, mientras que el segundo podría ser una medida del "tamaño aparente", creciendo hacia abajo (recordar que la pauta A4 se caracteriza por un total despliegue de alas). La pauta A3 tal vez no quedará demasiado bien situada, de acuerdo con esta interpretación, pero hay que destacar que es

la pauta más difícil de precisar. Las pautas A7 y A9 serían movimientos hacia el oponente sin aumento del tamaño aparente, mientras que la pauta A4 combinaría el movimiento con el aumento del tamaño aparente. El primer eje reflejaría más la agresividad, en cuanto a acción basada en la esperanza de éxito, y el segundo la amenaza, en cuanto a acción disuasoria, por ésto las pautas más agresivas (situadas hacia la izquierda) provocan más huídas y las pautas con mayor contenido de amenaza provocan mayor variabilidad en la respuesta del pasivo y por tanto mayor variabilidad en la réplica del activo, disminuyendo consiguientemente el "estar" en ACT2.

En las figuras 2a y 2b se representan las pautas del pasivo en función de las causas (pautas ACT1) y los efectos (pautas ACT2). En la figura 2a, se sitúan hacia la izquierda las pautas del pasivo que provocan más "estar" en ACT2, mientras que el eje vertical situa hacia arriba las pautas del pasivo asociadas a respuestas más agresivas del ACT2. El eje horizontal podría interpretarse como un eje de agresividad-no agresividad del pasivo, con valores crecientes de agresividad a la derecha, mientras que el eje vertical sería más bien un eje de temor-confianza, del pasivo, con valores crecientes de confianza hacia arriba.

La naturaleza del fenómeno estudiado hace que, ante una pauta agresiva del activo, no se den respuestas del pasivo, agresivas y a la vez con temor elevado, o respuestas no agresivas confiadas,

por ello las pautas se disponen principalmente en el primero y tercer cuadrantes. Nótese que el dendograma refleja bien las similitudes de las pautas del pasivo de acuerdo con las ideas que tenemos preformadas de ellas.

En las figuras 3a y 3b se representan las pautas con que replica el activo al pasivo (ACT2), en función de su propia actitud original y la respuesta del pasivo. En la figura 3a, el eje horizontal puede interpretarse como un eje asociado a la tolerancia-intolerancia del pasivo, con valores crecientes de intolerancia hacia la izquierda, mientras que el eje vertical puede ser interpretado como un eje de pasividad-actividad del pasivo, con valores altos de actividad hacia abajo (ataques, despliegues, huídas) y bajos arriba (estar).

El ataque del activo, en ACT2, suele ser una respuesta al ataque del pasivo, PAS, que es una actitud intolerante activa, mientras que la huída está asociada a una actitud intolerante más pasiva (despliegue), constituyendo ésta un medio de evitar una pelea.

El ataque o la huída se plantean pues como una alternativa al ACT2, favoreciéndose uno u otra según el grado de actividad de la respuestas intolerante del pasivo.

El dendograma muestra la proximidad del ataque y la huída del ACT2, (réplicas activas al pasivo), frente al despliegue y "estar" (réplicas menos activas al pasivo).

En las figuras 4a y 4b se representan las combinaciones de pautas ACT1 x PAS, en función de la respuesta final del activo, ACT2. En la figura 4a, en el eje horizontal, a la derecha, se situarían las combinaciones de pautas que implicarían frecuencias de "estar", en el ACT2, elevadas, mientras que el eje vertical, diferenciaría los ataques de las huídas del ACT2, situándose hacia arriba las combinaciones de pautas más asociadas a los ataques y abajo a las huídas, del ACT2. Nótese que la combinación A4E provocó sólo ataques en el ACT2, por ésto ha quedado arriba a la izquierda, aunque hay que tener presente que esta combinación de pautas fue muy poco frecuente y la estimación de la respuesta del activo, por tanto impreciso. El primer eje podría interpretarse como un eje de agresividad-no agresividad del pasivo, con agresividad creciente hacia la izquierda, mientras que el segundo evaluaría las combinaciones de pautas según la actitud más o menos agresiva que provocan en el activo (ACT2), con respuestas agresivas de éste hacia arriba. El comportamiento intolerante, en el ACT2, estaría influido por dos factores, uno sería la esperanza de éxito del activo en un enfrentamiento con el pasivo, ante una actitud intolerante o indiferente de éste. Otro factor sería la intolerancia del pasivo, en forma de ataque, que tiende a generar una respuesta agresiva del activo.

Hay que notar también que las huídas y sumisiones del pasivo producen una respuesta muy similar en el activo, independientemente de la primera pauta realizada por éste (ACT1). Esto es lógico ya que la respuesta predominante del ACT2 será "estar". Sin embargo,

las respuestas más intolerantes del pasivo (ataques, despliegues) producen una mayor variabilidad, es decir, la respuesta del activo depende, en esta circunstancia, de la primera acción realizada por éste (ACT1). Queda también claro, al igual que en las figuras 3a y 3b, que los despliegues del pasivo ocasionan más huidas del ACT2 que los ataques, debido a que dan la opción del activo a iniciar una pelea o no, mientras que los ataques fuerzan la lucha.

El dendograma, figura 4b, muestra que, globalmente considerado el proceso estudiado, es más importante la pauta realizada en último lugar, que la primera, cara a la respuesta del activo. Hay que notar también un coeficiente de correlación cofenética elevado, indicador de una buena clasificación jerárquica.

En las figuras 5a y 5b se representan las combinaciones de pautas del activo, ACT1 x ACT2 (actitud global de éste) respecto la distribución de pautas del pasivo. En la figura 5a, respecto el eje horizontal se disponen a la derecha las combinaciones de pautas asociadas a huidas o sumisiones del pasivo. En el centro, aquellas combinaciones más relacionadas con ataques del pasivo, y a la izquierda aquellas más relacionadas con los despliegues del mismo. Respecto el eje vertical se situarían hacia arriba las combinaciones de pautas del activo más relacionadas con una actitud indiferente, "estar", del pasivo.

El eje horizontal pone en claro una vez más la relación, ya

comentada del ataque o despliegue del pasivo y las respuestas ataque/huída del activo.

El dendograma, figura 5b, muestra también lo que hemos comentado anteriormente: la respuesta del pasivo determina grandemente el comportamiento del ACT2, con cierta independencia de la pauta inicial del activo (ACT1).

En las figuras 6a y 6b se representan las combinaciones de pautas PAS x ACT2 (efectos) en función de la distribución de pautas iniciales del activo (causas).

En la figura 6a el eje horizontal podría interpretarse como un eje de tolerancia-intolerancia inicial del activo, con valores crecientes de tolerancia hacia la izquierda.

De todos los comentarios anteriores podemos resumir algunas conclusiones, algunas de las cuales corroboran los hallazgos de Senar (1982), pero la aplicación sistemática de la metodología expuesta en el apartado 7.1. ha permitido hallar algunas nuevas. En general, parece claro que el comportamiento agonístico del Lúgano viene influido por varios factores: un factor de tolerancia-intolerancia, un factor de actividad-pasividad, la esperanza de éxito de un enfrentamiento, lo que podríamos llamar el "determinismo agresivo", ya que una conducta intolerante activa (como respuesta a una pauta intolerante inicial) suele implicar una réplica intolerante activa, etc.

También hay que destacar que, principalmente en pautas tolerantes, la conducta del animal viene básicamente determinada por la última pauta que le han realizado, más que por otras anteriores, quedando claro el valor comunicativo de los despliegues agresivos. Ello puede justificar, al menos en este caso, el uso de cadenas de Markov para el estudio del comportamiento animal. Finalmente cabe destacar, en los despliegues agresivos, dos componentes que podemos denominar componentes de movimiento y tamaño respectivamente.

8. UNA APLICACION AL DIAGNOSTICO AUTOMATIZADO.

Resumen:

En el presente capítulo se sugiere un algoritmo, basado en una función distancia y en la aplicación de una decisión Bayesiana, con ciertas propiedades de aprendizaje, utilizable para la diagnosis automatizada, dado algún tipo de análisis clínico. Se ilustra con el diagnóstico de algunas enfermedades a través de la interpretación de mielogramas.

Sumario:

8.1. PLANTEO GENERAL.

8.2. DIAGNOSTICO DE ALGUNAS ENFERMEDADES A PARTIR DE LA INTERPRETACION DE MIELOGRAMAS.

8.2.1. Descripción de las líneas celulares.

8.2.2. Consideración de algunas de las alteraciones de la médula ósea roja.

8.2.3. Una distancia en el conjunto de resultados.

8.3. ANALISIS CLINICOS Y RESULTADOS.

8.4. DISCUSION.

8.1. PLANTEO GENERAL

Vamos a partir de una población humana, usualmente formada por personas de una determinada área geográfica que presentan una sintomatología específica previa. Se consideraran k clases patológicas, E_1, \dots, E_k . Cada una de ellas puede ser una determinada enfermedad, un grupo de ellas o bien una variante de una sólo. Los conjuntos E_i deberán ser por construcción disjuntos y entre todos ellos deberán cubrir todas las posibles enfermedades de la población estudiada. En la práctica una clase patológica deberá estar formada por todas las enfermedades que no se tenga interés en estudiar.

Las probabilidades de los sucesos E_i , $p(E_i)$, son fácilmente estimables considerando un determinado número de elementos de la población y, determinando a que clase patológica pertenecen: frecuencia relativa del suceso E_i , a partir de una muestra de la población.

Dado un individuo de la población, si no disponemos de más información sobre el mismo, hay a priori una determinada probabilidad $p(E_i)$ de que pertenezca a la clase patológica E_i . Si efectuamos sobre este individuo algún tipo de análisis clínico, la información que nos da éste debe determinar un cambio en la distribución de probabilidad de las distintas clases patológicas. Generalmente, aunque no siempre, obtendremos una distribución de probabilidad menos entrópica.

Si pretendemos efectuar un diagnóstico, una regla de decisión razonable consiste en asignar a este individuo a la clase patológica más probable.

A continuación examinemos el problema de la estimación de las probabilidades de los sucesos E_i condicionadas a un determinado resultado de un análisis clínico. Si hay un número finito o numerable de resultados posibles, R_j , aplicando el teorema de Bayes podremos escribir:

$$p(E_i/R_j) = \frac{p(E_i) \cdot p(R_j/E_i)}{\sum_{h=1}^k p(E_h) \cdot p(R_j/E_h)} \quad (1)$$

donde la estimación de las $p(R_j/E_h)$ no ofrece demasiados problemas prácticos: basta escoger un número suficientemente grande de personas que pertenezcan a la clase patológica E_h y determinar en que proporciones se presentan los sucesos R_j . En cambio si hay un número no numerable (continuo) de resultados posibles, r , deberemos escribir:

$$p(E_i/r) = \frac{p(E_i) \cdot f(r/E_i)}{\sum_{h=1}^k p(E_h) \cdot f(r/E_h)} \quad (2)$$

donde $f(r/E_h)$ es la función de densidad, de los resultados, condicionada a la clase patológica E_h . La estimación de una función de densidad de probabilidad, en el caso continuo, presenta numerosos problemas

y requiere, para ser minimamente efectiva, tamaños muestrales excesivamente elevados, Duda & Hart (1973). Una alternativa razonable consiste en definir una partición finita $P = R_1, \dots, R_n$ del conjunto de todos los resultados posibles, distinguiendo sólo como distintos a dos análisis que pertenezcan a distintos elementos de la partición. Así podremos aplicar la expresión (1) para hallar las probabilidades de las distintas clases patológicas, una vez conocido el resultado de la prueba clínica aplicada.

Si llamamos A al conjunto de todos los resultados posibles, usualmente algún subconjunto de R^m , y en éste hay definida una distancia, d , hay una forma razonable de definir una partición P de A , de k elementos.

Definamos $\bar{d}(a, E_j)$ como la "distancia promedio" del resultado $a \in A$ a la clase patológica E_j , definida por:

$$\bar{d}(a, E_j) = E(d(a, r) / E_j) = \int_A d(a, r) f(r / E_j) dr \quad (3)$$

entonces podemos definir la partición P de A a través de:

$$a \in R_j \Leftrightarrow \begin{cases} \bar{d}(a, E_j) < \bar{d}(a, E_i) & i=1, \dots, j-1 \\ \bar{d}(a, E_j) \leq \bar{d}(a, E_i) & i=j+1, \dots, k \end{cases} \quad (4)$$

Dado $a \in A$, estas distancias promedio pueden ser fácilmente estimadas a partir de una muestra de enfermos de cada una de las clases patológicas.

Si, para cada individuo, los resultados de los análisis clínicos los podemos considerar como una muestra aleatoria simple de una cierta función de densidad paramétrica $f(x_1, \dots, x_n, \theta^1, \dots, \theta^m)$ podemos resumir a aquella a través de los estimadores (máximo-verosímiles) de los parámetros. Así, el conjunto de resultados A será idéntico a la variedad paramétrica asociada a f (ver capítulo 3) y podemos extender a aquel la métrica riemanniana definida en ésta última.

Si los análisis clínicos proporcionan mucha información, cara al diagnóstico de las distintas clases patológicas, es de esperar que:

$$p(R_j/E_j) = \int_{R_j} f(r/E_j) dr \approx 1 \quad (5)$$

y para $i \neq j$:

$$p(R_j/E_i) = \int_{R_j} f(r/E_i) dr \approx 0 \quad (6)$$

abreviadamente:

$$p(R_j/E_i) \approx \delta_{ij} \quad (7)$$

siendo δ_{ij} las deltas de Kronecker. Entonces al aplicar (1) resultará:

$$p(E_i/R_j) \approx \frac{p(E_i) \cdot p(R_j/E_i)}{\sum_{h=1}^k p(E_h) \cdot p(R_j/E_h)} = \frac{p(E_i) \cdot \delta_{ij}}{p(E_j)} = \delta_{ij} \quad (8)$$

por tanto, conocido R_j la regla de decisión nos conducirá a clasificar al enfermo dentro de la clase patológica E_j , con probabilidad de error casi nula. Por otra parte, la distribución de las probabilidades a posteriori posee la menor entropía posible.

Describamos a continuación el algoritmo de clasificación. Se parte de una muestra controlada de enfermos, M , es decir enfermos cuya clase patológica es conocida. A partir de esta muestra controlada se evalúan en primer lugar las probabilidades "a priori", $p(E_i)$. A continuación se determina para cada enfermo a que elemento de la partición $P = \{R_1, \dots, R_k\}$ pertenece, calculando su distancia media a cada uno de los individuos, de cada una de las enfermedades por separado. Si en M hay n_i individuos de la clase patológica E_i , la distancia media de un individuo cuyo resultado de análisis es ω , a la clase patológica E_i vendrá estimada por:

$$\bar{d}(\omega, E_i) = \frac{1}{n_i} \sum_{e \in M \cap E_i} d(\omega, e) \quad (9)$$

Una vez conocido, para cada individuo, a que elemento de la partición P pertenece, se procede a la estimación de las probabilidades condicionadas, $p(R_j/E_i)$, a través de las proporciones de los

sucesos R_j dentro de los enfermos de la muestra perteneciente a la clase patológica E_i .

A continuación, dado un individuo que desconocemos la clase patológica a la que pertenece, con un resultado de análisis ω , determinamos a través de (9) al elemento de la partición P al que pertenece y posteriormente aplicamos el teorema de Bayes (1).

La regla de decisión será asignar a este individuo a la clase patológica más probable a posteriori (en el improbable caso de empate habrá que recurrir a una decisión aleatorizada) con una probabilidad de error estimable:

$$P(\text{error}) = 1 - \max_i p(E_i/R_j) \quad (10)$$

Podemos también medir el cambio de información (variación de entropía) asociado al procesamiento del resultado del análisis. Si llamamos $p_i = p(E_i)$ y $q_i = p(E_i/R_j)$, entonces el cambio de información medido en bits viene dado por:

$$\Delta I = I_F - I_I = \sum_{i=1}^k q_i \log_2 q_i - \sum_{i=1}^k p_i \log_2 p_i \quad (11)$$

Si $\Delta I > 0$ significa que se ha ganado información (perdido entropía) con todo el proceso. $\Delta I < 0$ significa lo contrario, pudiendo darse el caso cuando el enfermo pertenece a alguna clase patológica cuya probabilidad a priori sea muy baja.

Una vez se haya podido determinar la enfermedad real del individuo que pretendíamos diagnosticar, puede ser utilizado este caso como un elemento más de la muestra controlada M , mejorando así la definición de la partición P y la estimación de las probabilidades, a priori y condicionadas.

Algunas modificaciones en la definición de la partición pueden ser sugeridas si consideramos no sólo la clase patológica más próxima a un resultado clínico sino las dos o tres más próximas. La partición contendría $k(k-1)$ ó $k(k-1)(k-2)$ elementos, podría mejorarse el diagnóstico pero habría necesidad de un mayor tamaño muestral.

8.2. DIAGNOSTICO DE ALGUNAS ENFERMEDADES A PARTIR DE LA INTERPRETACION DE MIELOGRAMAS

Vamos a aplicar lo expuesto en el apartado anterior al diagnóstico de enfermedades partiendo de los resultados obtenidos al efectuar un mielograma.

Para efectuar un mielograma se hace una punción casi siempre en el esternón o en la cresta iliaca (algunas veces también en la tibia, vértebras, etc.) y posteriormente se aspira, obteniendo médula ósea. A partir de la misma, se efectúan extensiones en portaobjetos que son posteriormente fijadas (generalmente con metanol) y teñidas, siguiendo alguno de los tres métodos más corrientes: Giemsa, Pappenheim y Wright, Miale (1982), Sandoz (1972), Rozman (1978).

Posteriormente se efectúa, al microscopio óptico, un recuento de los distintos tipos celulares presentes en la médula ósea, algunos de ellos puestos de relieve no tanto por su morfología sino a través de pruebas citoquímicas, como la reacción de la esterasa, el PAS, la fosfatasa alcalina y la fosfatasa ácida, la peroxidasa, etc. Una vez efectuado un recuento de un número elevado de células, se calculan los porcentajes. Variaciones de éstos están fuertemente asociadas a algunas enfermedades. Veamos a continuación los principales tipos de células presentes en la médula ósea.

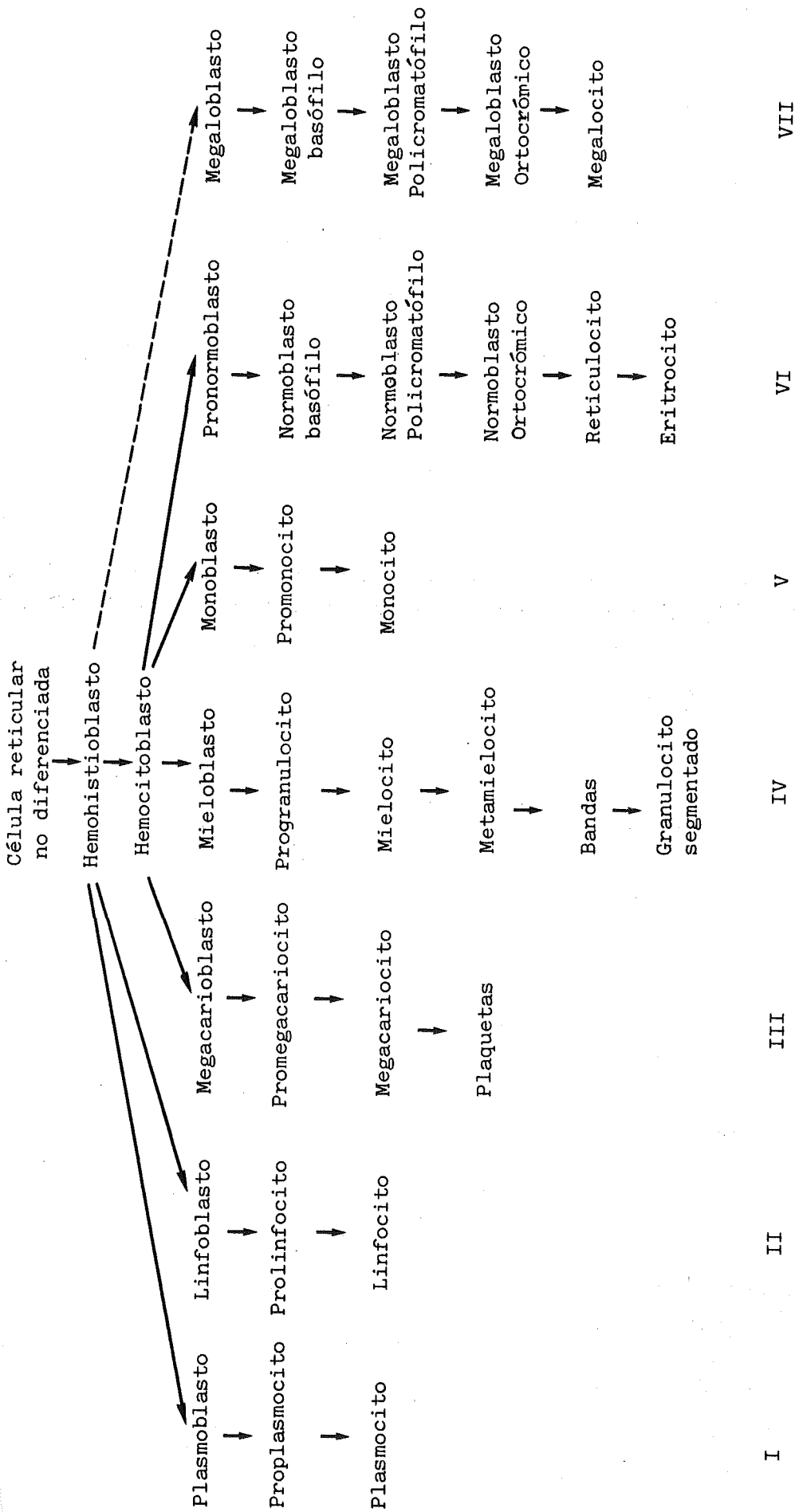
8.2.1. Descripción de las líneas celulares.

Las células que se presentan normalmente en la médula ósea son derivadas de un tipo de célula madre indiferenciada del sistema retículo-endotelial (stem cell). La posterior diferenciación de ésta dará lugar a varias líneas celulares, distinguiéndose las células en cada una de éstas por el distinto grado de madurez. Hay varias características que son indicativas de la madurez celular, y son comunes a las distintas líneas: pérdida de la basofilia citoplasmática, debida a una disminución del ARN_m, producción de granulos citoplasmáticos y síntesis de sustancias específicas (peroxidasa en la serie granulocítica, hemoglobina en la serie eritrocítica, etc.), cambios morfológicos en el núcleo, de redondo u oval pasa en algunos casos a ser segmentado, etc., condensación de la cromatina, pérdida de nucleolos, disminución de la relación núcleo/citoplasma, disminución del tamaño nuclear (salvo megacariocitos que son poliploides) etc., Miale (1982), Bernard & Bessis (1965).

Las líneas celulares a distinguir son: I) serie plasmocítica cuyo elemento más maduro es el plasmocito o célula plasmática, II) serie linfocítica que comprende desde los linfoblastos a los linfocitos, que juegan un importante papel en la síntesis de anticuerpos. III) Serie trombocítica cuyas células (megacariocitos) forman las plaquetas. IV) Serie granulocítica o mieloide, cuya célula más diferenciada es el granulocito segmentado, en sus variedades neutrófila, eosinófila y basófila, de importantes propiedades fagocitarias. V) Serie monocítica, que comprende desde el monoblasto al monocito, células macrofágicas muy activas. VI) Serie eritrocítica, cuyo exponente más diferenciado es el eritrocito, portador de hemoglobina. VIII) Serie megalocítica, puede considerarse como una anomalía de la anterior, ligada a una ausencia del proceso de maduración nuclear, asociado a algunos tipos de anemias. El elemento más diferenciado es el megalocito o eritrocito macrocítico.

Para una descripción detallada de los distintos tipos celulares que forman las anteriores series, puede consultarse Bernard & Bessis (1965), Sandoz (1972), Miale (1982).

El siguiente esquema completa lo anteriormente expuesto.



LINEAS CELULARES PRESENTES
EN MEDULA OSEA, Miale (1982)

8.2.2. Descripción de algunas alteraciones de la médula ósea

Las proporciones promedio de los distintos tipos celulares en una médula ósea de un adulto normal pueden verse en la siguiente tabla:

Mieloblastos	1.33%	Linfoblastos	-
		Prolinfocitos	-
Progranulocitos	3.32%	Linfocitos	10.84%
Mielocitos	9.62%	Monoblastos	-
Neutrófilos	7.74%	Promonocitos	-
Eosinófilos	1.55%	Monocitos	1.55%
Basófilos	0.33%		
		Plasmoblastos	-
Metamielocitos	12.17%	Proplasmocitos	-
Neutrófilos	9.95%	Plasmocitos	-
Eosinófilos	1.66%		
Basófilos	0.33%	Megacarioblastos	-
		Promegacariocitos	-
Bandas	19.70%	Megacariocitos	0.22%
Neutrófilos	17.70%		
Eosinófilos	1.77%	Promegaloblastos	-
Basófilos	0.22%	Megaloblastos basófilos	-
		Megaloblastos policromatófilos	-
Segmentados	17.25%	Megaloblastos ortocrómicos	-
Neutrófilos	14.82%		
Eosinófilos	2.21%	Pronormoblastos	0.55%
Basófilos	2.22%	Normoblastos basófilos	2.65%
		Normoblastos policromatófilos	12.9%
		Normoblastos ortocrómicos	4.98%
		Células reticulares	0.22%
		No identificadas	1.88%

Algunas enfermedades ocasionan alteraciones en las proporciones celulares, dependiendo de la serie o series celulares afectadas.

La proliferación, no maligna, de cada una de las líneas celulares ocasiona las llamadas hiperplasias. Las hiperplasias granulocíticas pueden ser de origen infeccioso, pueden ocurrir por un descenso relativo de las demás series celulares, o pueden ser de origen desco-