ASPECTES DE LA PROBLEMÀTICA DE LA DISTÀNCIA TAXONÒMICA

per JORDI OCAÑA i REBULL

Unitat Docent de Bioestadística. Facultat de Biologia.
Universitat de Barcelona

INTRODUCCIÓ

A la figura 2 hi ha un esquema, necessàriament simplificat que tracta de reflectir les relacions existents entre els mètodes matemàtics utilitzables per un investigador que vulgui estudiar les semblances (o diferències) entre unitats taxonòmiques orgàniques (OTUs) i establir classificacions. La problemàtica de la distància taxonòmica es pot situar en els esglaons (a) i (b) + (b').

Generalment es parteix d'una taula de dades (figura 1) on cada unitat taxonòmica queda caracteritzada com un element d'un espai de n dimen-

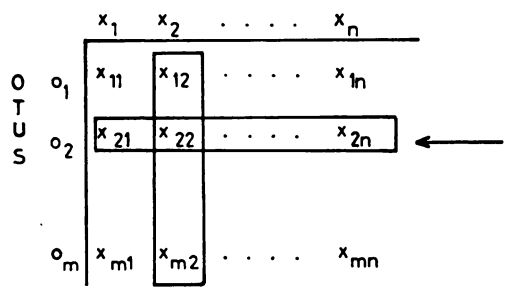


FIG. 1. — Exemple de taula de dades

sions, de coordenades (x_{i1}, ... ,x_{in}), i cada caràcter per un element d'un espai m-dimensional, amb coordenades (x_{1j}, ... ,x_{mj}). Però amb la matriu de dades no n'hi ha prou, cal considerar també quina és l'estructura d'aquests espais. Això depèn del tipus d'estudi realitzat, podríem dir de

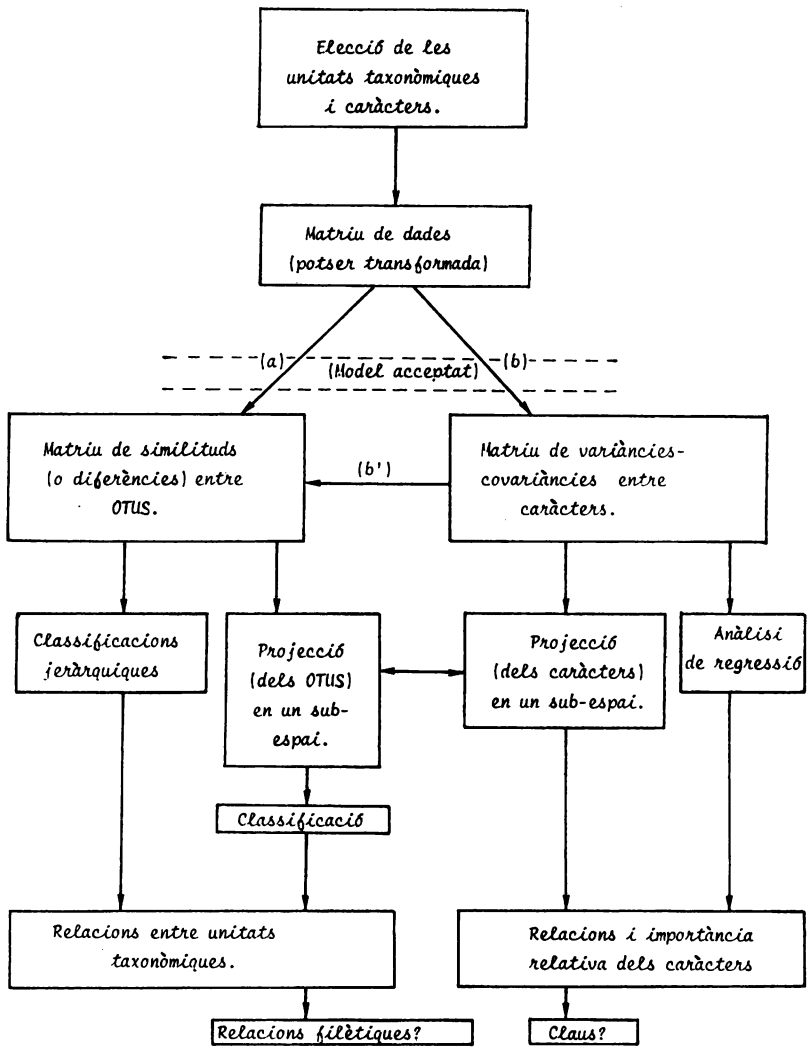


FIG. 2.

Esquema simplificat (basat en BOYCE, 1969) de les relacions existents entre els mètodes matemàtics utilitzables per a establir classificacions d'unitats taxonòmiques orgàniques (OTUs)

quin és el tipus de model inicial considerat més adient (si es poden considerar els caràcters independents, si es tracta de variables aleatòries amb distribució normal, etc.) i de la mateixa natura dels caràcters emprats (normalment es poden considerar variables aleatòries, però el seu «nivell de mesura» pot ésser ben diferent: mesures contínues, comptatges, indicadors de presència-absència, etc...).

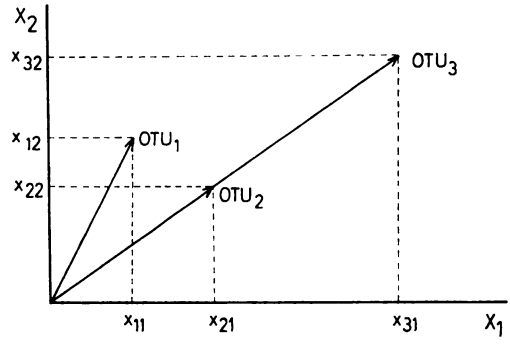


FIG. 3. — Esquema de la posició de tres OTUs en un espai de dues dimensions

Quan s'ha fixat la natura de l'espai de representació es pot intentar definir o escollir amb una certa precisió la distància taxonòmica més adequada.

La següent classificació dels índexs de semblança pot servir com un punt de referència en la discussió que a continuació detallem:

—Coeficients angulars
—Coeficients de distància
*Al marge de l'estructura de l'espai
°No mètriques
°Mètriques
*Índexs que tenen en compte (o ho proven) l'estructura de l'espai (normalment tenen propietats de mètrica).

COEFICIENTS ANGULARS

Un exemple característic n'és la distància de CAVALLI-SFORZA i EDWARDS (1967) i EDWARDS (1971). També ho són les expressions semblants als coeficients de correlació, quan s'utilitzen per relacionar OTUs. Els

darrers corresponen al cosinus de l'angle format pels dos OTUs, un cop representats en un cert espai. El primer correspon a la corda de l'arc associat a les dues unitats taxonòmiques representades en un espai transformat.

Aquests índexs reflexen les diferències degudes a l'orientació dels OTUs, no les degudes a la seva dimensió. Com es pot apreciar a la figura 3 la distància angular (angle, cosinus) entre OTU_1 i OTU_2 és la mateixa que entre OTU_1 i OTU_3. Hom podria considerar que la «punta» d'OTU_1 està més allunyada d'OTU_3 que d'OTU_2, ja que OTU_3 és més «llarg» i per tant esperar que la distància d'OTU_1 a OTU_3 fos més gran.

Coeficients de distància

Es podria considerar com a tals aquells que pretenen expressar les dues components (angular, dimensional), junt amb una llarga llista d'índexs d'interpretació geomètrica difícil o impossible, ja que normalment no es tracta de mètriques, malgrat que puguin ésser útils en algun cas determinat. Els darrers inclourien índexs basats en la probabilitat de trobar valors iguals d'un caràcter, en observacions independents, a cada OTU: Sneath (Sokal i Sneath, 1963), Hedrick (1971); índexs basats en la Teoria de la Informació: Kullback (1968), Orloci (1969); índexs d'associació: Jaccard (Sneath i Sokal, 1973), Gower (1971); distància de Nei (1970-71).

Dins de la classe dels índexs amb propietats de mètrica cal distingir clarament entre aquells que no tenen en compte l'estructura de l'espai (malgrat que la seva definició ja pressuposa una interpretació geomètrica dels OTUs, es suposa que aquests són punts d'un espai euclidi, expressats segons una base de vectors ortonormals, com si consideréssim els caràcters incorrelacionats i cadascú amb el mateix pes) i aquells que tracten de posar-la de manifest.

Les mètriques de Minkovski, de fórmula general

$$D_r = \left(\sum_{j=1}^{n} \left| x_{1j} - x_{2j} \right|^r \right)^{1/r}$$

donen lloc, quan r = 1, a la «distància segons les illes de cases a una ciutat» (per anar de l'OTU_1 a l'OTU_2, es va «trencant» de dimensió en dimensió, figures 4a i 4b)

$$D_1 = \sum_{j=1}^{n} \left| x_{1j} - x_{2j} \right|$$

i quan r = 2, a la distància euclídia usual (figura 4a)

$$D_2 = \left(\sum_{j=1}^{n} (x_{1j} - x_{2j})^2 \right)^{1/2}$$

La distància D_1, dividida per el nombre de caràcters, correspon a la distància M.C.D. (*mean character diference,* CAIN i HARRISON, 1957). Una expressió semblant va ésser proposada per PREVOSTI (1974), com a distància genètica entre poblacions, caracteritzades per les freqüències dels allels de n caràcters (s_j allels pel caràcter j, j = 1,...,n)

$$A = 1/2n \left(\sum_{j=1}^{n} \sum_{K=1}^{s_j} \left| p_{1jk} - p_{2jk} \right| \right)$$

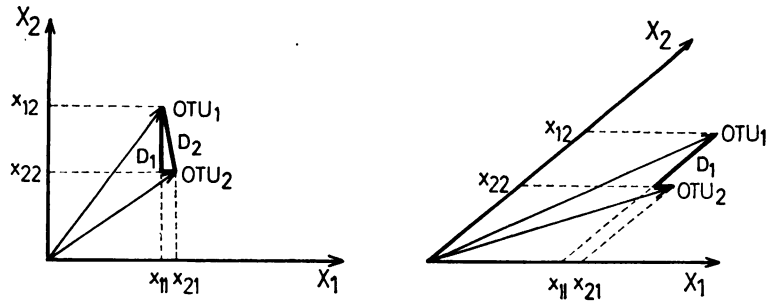
Com mostren les figures 4a i 4b, aquestes distàncies són independents de l'angle (correlacions) entre els vectors (variables aleatòries) que formen el sistema de referència, la qual cosa no passa amb D_2 o amb $1/nD_2$ (M.S.D. *mean square diference*) o amb la distància genètica de ROGERS (1972), que només tindria sentit en un sistema de referència com el de la figura 4a.

Certs índexs, com ara el coeficient de Pearson C.R.L. *(coefficient of racial likeness)* i la distància ji-quadrat de BENZECRI (1973), malgrat que no tinguin en compte les relacions entre caràcters, sí que ponderen d'alguna manera el «pes» de cada un d'ells. La distància ji-quadrat, utilitzada per comparar distribucions finites, s'expressa com

$$X^2 = \sum_{j=1}^{n} 1/p_{\cdot j} \, (p_{1j} - p_{2j})^2$$

essent $p_{\cdot j}$ la probabilitat de que es presenti l'estat j (per exemple, l'allel j) a la reunió de totes les poblacions estudiades. Aquesta distància té la propietat, realment interessant (BENZECRI, 1973), d'ésser pràcticament equivalent a la informació mútua (informació processada per un canal) entre les distribucions associades a les unitats taxonòmiques 1 i 2.

És interessant que les distàncies tinguin en compte les relacions entre caràcters, ja que si aquests són dependents, la informació continguda en un d'ells estarà en part ja expressada en els altres. Quan les unitats taxonòmiques han estat caracteritzades mitjançant variables aleatòries amb



FIGS. 4a-4b. — Esquema de la posició de dos OTUs en un espai de dos dimensions. *a:* eixos perpendiculars. *b:* eixos correlacionats

distribució conjunta normal multivariant, i quan la matriu de variàncies covariàncies és la mateixa a tots els OTUs, es pot utilitzar la distància de Mahalanobis.

$$\Delta^2 = (m_1 - m_2)' \; \Sigma^{-1} \; (m_1 - m_2)$$

essent

$$m_1 = (m_{11}, \ldots\ldots m_{1n})$$
$$m_2 = (m_{21}, \ldots\ldots m_{2n})$$

els vectors de mitjanes dels n caràcters, a les poblacions 1 i 2, i la matriu de variàncies covariàncies comunes.

Aquesta distància té certes propietats que la fan òptima. Això es pot deduir de raonaments basats en la raó de versemblança (MAHALANOBIS, 1936) o en raonaments basats en conceptes d'àlgebra lineal (per exemple, DEMPSTER, 1969). Es pot demostrar que té cura de les correlacions entre caràcters. Com en el cas de distribució multinormal, incorrelació implica independència estocàstica, el problema queda ben resolt.

Però quan les variables estan lligades per relacions no lineals (sovint les dades es poden considerar situades en hipersuperfícies, esfèriques per exemple) ja no té sentit considerar una distribució multinormal. Evidentment, tampoc en té quan es tracta de variables no contínues. En aquests casos s'ha fet servir la distància de Mahalanobis o distàncies formalment

anàlogues, adaptades a d'altres tipus de variables aleatòries (BALAKRISHNAN i SANGHVI, 1968; KURKZÒNSKI, 1970; OCAÑA, ALONSO i PREVOSTI, 1976). La validesa d'aquestes distàncies és dubtosa, car només tenen en compte les relacions lineals entre caràcters, això vol dir que la seva eficiència serà més gran quan el coeficient de correlació sigui més proper a la raó de correlació. En realitat també és dubtós que reflecteixin les relacions lineals ja que com la mitjana i la variància no es poden considerar independents, l'estimació de la matriu de variàncies-covariàncies comuns, és força problemàtica.

Probablement caldria definir una distància basada en una teoria matemàtica més general que l'àlgebra lineal, potser basada en la raó de correlació. Es tractaria de trobar una transformació que permetés passar d'un conjunt de variables dependents a un conjunt de variables independents. Potser la Teoria de la Informació tingui quelcom a dir al respecte.

BIBLIOGRAFIA

1. BALAKRISHNAN, V.; SANGHVI, L. D.: (1968). Distances between populations on the basis of attribute data. *Biometrics,* 24: 859-866.
2. BENZECRI, J. P.: (1973). L'analyse des données. I. La Taxinomie. Ed. Dunod.
3. BOYCE, A. J.: (1969). Mapping Diversity: a comparative study of some numerical methods. *Proc. Coll. Num. Tax. Univ. St. Andrews,* September 1968), (1-31). A. J. Cole Ed. Academic Press.
4. CAIN, A. J.; HARRISON, G. A.: (1958). An analysis of the taxonomist's judgemnt of affinity. *Proc. Zool. Soc. Lond.,* 131: 85-98.
5. CAVALLI-SFORZA, L. L.; EDWARDS, A. W. F.: (1967). Phylogenetic analysis: models and estimation procedures. *Evolution,* 21: 550-570.
6. DEMPSTER, A. P.: (1969). Elements of continous Multivariate Analysis. Addison-Wesley.
7. EDWARDS, A. W. R.: (1971). Distances between populations on the basis of genc frequencies. *Biometrics,* 27: 873-881.
8. GOWER, J. C.: (1971). A general coeficient of similarity and some of its properties. *Biometrics,* 27: 857-871.
9. HEDRICK, P. W.: (1971). A new approach to measuring genetic similarity. *Evolution,* 25: 276-280.
10. JACCARD, A.: (1973). Distances généalogiques et distances génétiques. *Cahiers d'Antropologie et d'Ecologie Humaine,* 1: 11-124.
11. KULLBACK, S.: (1968). Information Theory and Statistics. J. Wiley.
12. KURKZYNSKY, T. W.: (1970). Generalized distance and discrete variables *Biometrics,* 26: 525-534.
13. MAHALANOBIS, P. C.: (1936). On the generalized distance in Statistics. *Proc. Nat. Inst. Sci. India,* 2: 49-55.
14. NEI, M.: (1971). Interespecific gene differences and evolutionary time estimated from electrophoretic data on protein identity. *Amer. Nat.,* 105: 385-398
15. NEI, M.: (1972). Genetic distance between populations, Amer. Nat., 106: 283-292.
16. OAÑA, J.; ALONSO, G. i PREVOSTI, A.: (1977). La estructura matemática de las poblaciones y la definición de distancias entre ellas. 2.º S.I.N.A.P.E. Universidade Estadual de Campinas. Brasil.

17. ORLOCI, L.: (7969). Information Theory models for hierarchic and non hierarchic classifications. *Proc. Coll. Num. Tax. Univ. St. Andrews.* September 1968 (148-164). Academic press.
18. PREVOSTI, A.: (1974). La distancia genética entre poblaciones. Miscelanea Alcobé: 109-118.
19. SNEATH, P. H. A.; SOKAL, R. R.: (1973). Numerical Taxonomy. W. H. Freeman.
20. SOKAL, R. R.; SNEATH, P. H. A.: (1963). Principles of Numerical Taxonomy. W. H. Freeman.

DISCUSSIÓ

CUADRAS

M. G. KENDALL introduí una geometria estocàstica que considerava distàncies geodèsiques. Utilitza mètodes molt complicats de geometria diferencial i exigeix recursos molt difícils d'anàlisi matemàtica. El problema de certes distàncies no euclídies és que es relacionen amb matrius de «covariàncies» que tenen valors propis negatius. Les coordenades euclídies per a representar poblacions, tenen, aleshores, unes coordenades reals i unes altres imaginàries, és a dir, la «distància» és negativa per alguns determinats eixos. Existeix, altrament un mètode de formació molt diferent, però amb més aplicació que la geometria estocàstica, que permet trobar una representació euclídia raonable de distàncies no euclídies. És l'análisi de proximitats de Torgerson, Shepard, Kruskal, Caroll i altres.

ALONSO

En aquest camp, caldria més comunicació entre biòlegs i matemàtics.