LE PROTISTOLOGUE ET LA TAXINOMIE NUMÉRIQUE (1)

PAR Pierre LEGENDRE (*)

SOMMAIRE

1. — INTRODUCTION

GOETHE, physicien, avocat, homme de lettres, n'aimait pas beaucoup les mathématiciens. Il croyait, en effet, que les « mathématiciens sont comme les Français. Parlez-leur et ils traduiront dans leur propre langue, où cela deviendra immédiatement quelque chose de tout à fait différent » (*in* MAGNUS, 1949). Je me passe de commentaires sur son opinion des

(1) Conférence présentée à la table ronde « Méthodes modernes en taxinomie », 18e Colloque du Groupement des Protistologues de Langue française, tenu à l'Université Pierre-et-Marie-Curie, Laboratoire Arago, Banyuls-sur-Mer, France, le 25 mai 1979 (P. DE PUYTORAC).

(*) Centre de Recherche en Sciences de l'Environnement, Université du Québec à Montréal, C. P. 8888, Succursale « A », Montréal, Québec H3C 3P8.

Français ! Pour lui, comme pour d'autres physiciens du XVIIIe siècle, la physique était affaire de perception essentielle de la réalité, alors que la formulation mathématique des phénomènes physiques représentait une gigantesque supercherie, prétendant formuler des semblants de démonstrations qui ne pouvaient rien apprendre de nouveau, puisque les conclusions devaient obligatoirement être contenues dans les prémisses. Les développements spectaculaires de la physique mathématique aux XIXe et XXe siècles devaient lui donner tort.

Moins violente a été la transition de la taxinomie classique à la taxinomie numérique, dont les fondements ont été posés, dans les années 1958-1960, par quatre équipes distinctes, formées, avant tout, aux méthodes classiques de la taxinomie : le « tandem » SOKAL et MICHENER, alors à l'Université du Kansas, SNEATH au National Institute for Medical Research de Londres, ROGERS et TANIMOTO au Jardin botanique de New York, enfin, CAIN et HARRISON à Oxford. SOKAL et SNEATH, en particulier, se sont réclamés, dès le départ, de principes énoncés deux siècles plus tôt par le botaniste français Michel ADANSON. Les bases néo-adansoniennes de la taxinomie numérique, que ROGERS appelait plutôt la taximétrie, sont les suivantes (SNEATH et SOKAL, 1973) :

1. Plus grand est le nombre de caractères considérés, meilleure est la classification produite.

2. *A priori*, tous les caractères ont un poids égal.

3. La mesure globale de ressemblance entre paires d'individus est une fonction de leur ressemblance quant à chacun des caractères servant à les comparer.

4. Les corrélations différentes entre caractères, d'un groupe d'organismes à un autre, permettent de reconnaître des taxa à l'aide de ces mêmes caractères.

5. On peut faire des déductions phylétiques à partir de la structure taxinomique d'un groupe, si l'on admet certaines hypothèses de base quant aux mécanismes d'évolution de certains caractères.

6. La taxinomie numérique est une science empirique plutôt que déductive.

7. La taxinomie numérique fonde ses classifications sur la ressemblance phénétique. Il faut d'abord remarquer, dans cette énumération de principes, que la taxinomie numérique est fondée sur l'hypothèse que la structure taxinomique recherchée émergera de la considération simultanée de l'ensemble des caractères, plutôt que de certains types de caractères pour chacune des catégories taxinomiques (principes 1 et 2). Cela résulte du fait que dans différents taxa, du niveau espèces, par exemple, les caractères sont en corrélation les uns avec les autres et que les taxa recherchés sont justement reconnaissables en tant qu'ils sont caractérisés par des combinaisons bien définies de l'ensemble des caractères (principe 4). Enfin, le principe 6, qui fait de la taxinomie

numérique une science empirique plutôt que déductive, montre bien qu'elle a été mise sur pied, non pas par des mathématiciens, mais bien par des biologistes désireux simplement d'appliquer les principes classiques de la systématique à des ensembles multi-caractères. C'est ainsi que le taxinomiste procède, avec les méthodes numériques, selon la même séquence d'étapes qu'en taxinomie classique : — 1. Le problème taxinomique étudié est matérialisé par un échantillonnage. — 2. On choisit des caractères qui permettent de décrire les spécimens de façon comparative. — 3. On évalue le degré de ressemblance entre les spécimens. — 4. On regroupe les spécimens en taxa à l'aide de ces évaluations de leurs ressemblances. — 5. Finalement, on peut étudier les implications de cette classification en termes de relations phylétiques ou de choix de caractères discriminants pouvant servir à leur identification.

La taxinomie numérique peut donc être définie comme *le groupement de spécimens en taxa, à l'aide de méthodes numériques, sur la base de leur description quant à leurs caractères observables*. Elle n'inclut donc pas les autres méthodes taxinomiques quantitatives, comme la sérologie ou la chromatographie, quoique les informations obtenues par ces autres méthodes puissent être utilisées comme caractères dans les études taxinomiques numériques. La modélisation numérique de chaînes évolutives est également exclue de cette définition de la taxinomie numérique.

En protistologie, les travaux pionniers de taxinomie numérique sont dus à GATES et BERGER, de l'Université de Toronto, qui ont tenté d'attaquer les problèmes de taxinomie des Ciliés à l'aide de méthodes numériques. Leurs premiers travaux (GATES et BERGER, 1974, 1976) ont exploré les variations biométriques de certains caractères quantitatifs et leur capacité de permettre, en les combinant, de séparer des espèces qui se ressemblent. Leurs travaux ont rapidement débouché sur une bonne connaissance des meilleurs caractères et combinaisons de caractères à employer dans les études quantitatives (BERGER, 1978) ainsi que sur les études phylogénétiques des Ciliés (GATES, 1978 *a*; LYNN, 1978). D'autres efforts ont enfin porté sur des études quantitatives de l'écologie des Ciliés (TAYLOR, 1978 *a*, 1978 *b*).

2. — LA MATRICE DE DONNÉES DU PROTISTOLOGUE

Les données employées en taxinomie numérique des Protistes forment un tableau dans lequel des spécimens sont décrits quant à un certain nombre de caractères qui peuvent être qualitatifs ou quantitatifs. Nous verrons plus loin comment combiner tous ces caractères en des mesures globales de la ressemblance entre spécimens.

Les caractères peuvent appartenir à différents types, dont le plus connu est le caractère métrique. Les études sur les Ciliés, par exemple,

ont permis d'établir une série de mesures standard entre des points clairement identifiables de la morphologie des spécimens (fig. 1); ces mesures sont souvent transformées, par la suite, en proportions, par exemple, par rapport à la taille totale du spécimen, de façon à dégager la variabilité taxinomique de la variabilité allométrique. Pour faciliter la prise de ces mesures, GATES et BERGER (1976) recommandent d'ailleurs l'emploi d'un coordinomètre électronique qui permet de relever aisément les coordonnées des points-repères sur une image du spécimen, projetée, par chambre claire, directement sur la table du coordinomètre : ces données, relevées sur ruban magnétique, sont ensuite fournies à un programme d'ordinateur qui calcule les distances entre les points-repères et qui transforme ces distances en pourcentages de la taille totale. Le travail du protistologue s'en trouve ainsi grandement simplifié.

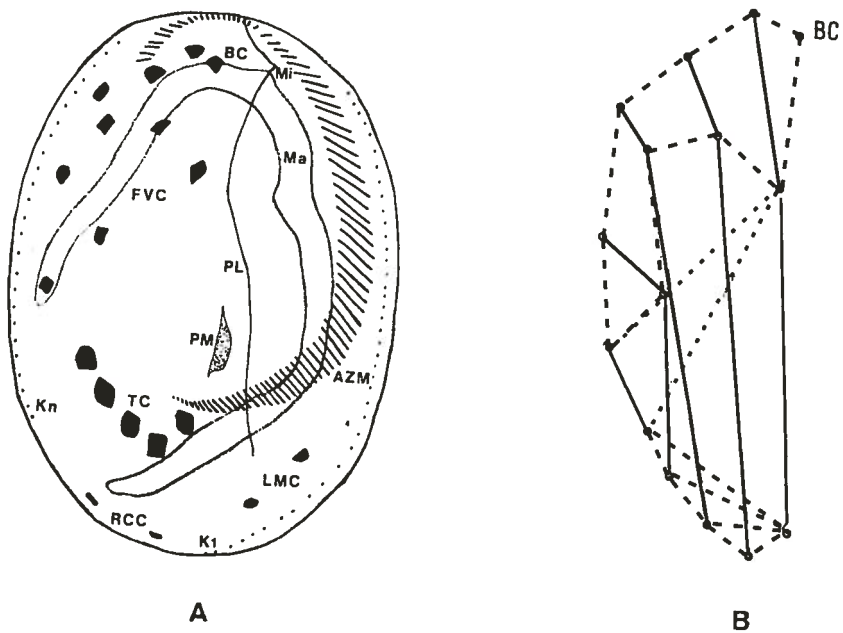

Fig. 1. — A, Diagramme représentant la surface ventrale d'*Euplotes harpa*; B, un schéma de 28 distances intercirrales. Le point BC (cirre buccal) est noté sur les deux diagrammes pour référence. Modifié de GATES (1978 *b*) que l'on pourra consulter pour les autres abréviations.

Un autre type de caractères est constitué de variables ordonnées mais non métriques. Cette classe comprend, par exemple, des caractères quantitatifs codés, sans mesure précise, en classe du type taille : « petite, moyenne, grande » ou encore coloration : « aucune, pâle, intermédiaire, foncée ». Ces caractères auraient avantage à être employés

aussi souvent que possible dans les études taxinomiques quantitatives puisqu'on sait maintenant les traiter mathématiquement et qu'ils peuvent souvent être relevés beaucoup plus rapidement, ou encore à moindre coût que les caractères métriques.

Une dernière classe de caractères comprend les observations qualitatives, ni métriques ni même ordonnées, que l'on sait aussi maintenant traiter tout aussi bien que d'autres caractères. Des exemples en sont : couleur : « bleu, brun, jaune » ou forme : « comme le dessin A, comme le dessin B, comme le dessin C, etc. » ou encore texture : « lisse, granuleuse, ondulée ».

Nous allons maintenant examiner comment combiner ces caractères en des mesures de la ressemblance entre individus, puis, comment tirer de toutes ces mesures de ressemblance, une connaissance de la structure taxinomique, processus qui passe par des analyses de groupement et d'ordination en espace réduit.

3. — LES MESURES DE LA RESSEMBLANCE TAXINOMIQUE

Les mesures de ressemblance taxinomique disponibles dans la littérature sont nombreuses. Elles peuvent heureusement être classées en quelques grands types :

Mode Q ou *mode* R : les coefficients de type R sont ceux qui, comme le coefficient de corrélation linéaire appelé le *r* de PEARSON, visent à décrire le degré de dépendance entre les *caractères*. Le *r* de PEARSON peut être employé pour décrire les relations linéaires existant entre des caractères *métriques*. Pour des relations curvilinéaires, ou dans le cas de caractères ordonnés mais non métriques, il existe des coefficients de dépendance non paramétriques tel le τ de KENDALL ou le *r* de SPEARMAN, qui ne fondent pas leur perception de la dépendance sur les paramètres de la distribution normale que sont la moyenne et l'écart-type. Enfin, entre les caractères non-ordonnés, on peut calculer des coefficients de dépendance fondés sur la comparaison de ces caractères par tableaux de contingence : le mieux connu de ces coefficients est le χ^2, appelé aussi justement le carré de contingence. Les autres coefficients disponibles dans la littérature taxinomique servent plutôt à mesurer le degré de ressemblance entre les objets (ou spécimens) de l'étude et on les qualifie globalement de *coefficients de type Q*.

Ces coefficients Q ont été historiquement divisés en deux classes : les mesures de similarité et les mesures de distance. Les mesures de similarité prennent leur valeur maximum pour deux objets identiques et leur valeur minimum pour deux objets complètement différents. Les distances obéissent à la loi inverse, si bien qu'il est, en général, facile de passer de l'une à l'autre, soit par complémentarité ($D = 1 - S$)

ou encore à l'aide de formules de transformation du type $D = \sqrt{1 - S}$ ou encore $D = \sqrt{1 - S^2}$.

Enfin, *les coefficients de similarité*, dont il sera plus spécialement question ci-dessous, peuvent être divisés en deux classes, selon qu'ils traitent des caractères binaires (de types présence-absence) ou les caractères à descriptions multiples.

Pour comprendre comment on combine les caractères de façon à obtenir une mesure globale de la similarité entre spécimens, considérons, d'abord, un coefficient fondé sur des caractères binaires : la mesure globale sera fondée sur le nombre de caractères qui seront présents (1,1) ou absents (0,0) chez les deux spécimens, par rapport au nombre total de caractères considérés. Ces informations peuvent être rassemblées dans un tableau de fréquence 2 × 2 servant à comparer ces deux spécimens, qui est illustré à la fig. 2. Dans ce tableau, *a* est le nombre de caractères pour lesquels les deux spécimens sont codés 1, ou +, etc., *d* est le nombre de caractères qui codent les deux spécimens à 0, alors que *b* et *c* sont les nombres de caractères qui codent les deux spécimens différemment; *n* est le nombre total de caractères. Une façon naturelle d'établir la similarité des deux spécimens consiste à compter le nombre de caractères qui codent les deux spécimens de façon semblable et à diviser par le nombre total de caractères :

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

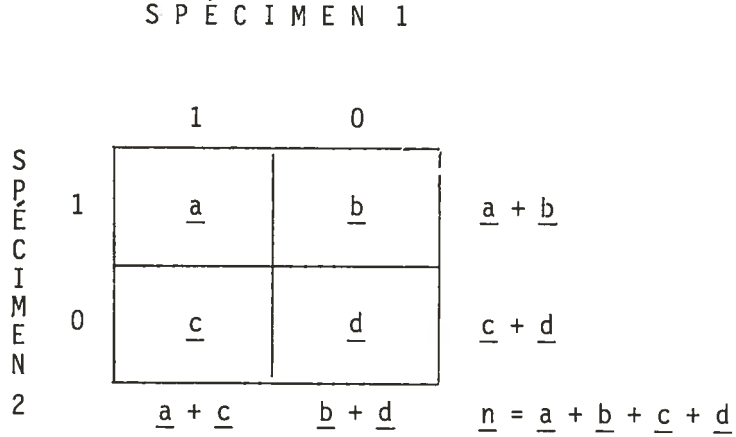

Fig. 2. — Comparaison de deux spécimens quant à *n* caractères binaires (présence-absence, ou vrai-faux).

Cette mesure est le coefficient de simple concordance (« *simple matching coefficient* ») de SOKAL et MICHENER (1958). Une variante de cette mesure est le coefficient de ROGERS et TANIMOTO (1960) qui donne

aux différences un poids plus important qu'aux ressemblances :

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

SOKAL et SNEATH (1963) ont ainsi présenté une série de variantes du coefficient de base. D'autres coefficients existent aussi, qui excluent les doubles zéros, donc la valeur *d*, de la comparaison, mais ceux-là sont plutôt employés en écologie. La mesure-type de cette famille est le coefficient de communauté de JACCARD (1908) :

$$S = a/(a + b + c)$$

Lorsque les données se présentent plutôt sous la forme de mesures, on peut les combiner en des mesures globales de la similarité en considérant l'écart entre les deux spécimens comparés, par rapport à l'écart dans la population de référence. GOWER (1971) a proposé un coefficient général de similarité, qui permet de combiner des caractères binaires et quantitatifs, après avoir traité chacun en accord avec son type mathématique. Ce coefficient a d'abord la forme suivante :

$$S(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} s_{i12}$$

où s_i est la valeur de similarité pour chacun des caractères. Si *i* est un caractère binaire, alors s_i prend la valeur 1 ou 0 selon qu'il y ait accord ou désaccord. Quand *i* est un caractère quantitatif, la différence est alors exprimée de façon fractionnaire par rapport à l'écart maximum R_i que l'on trouve dans la population (échantillon, ou autre population de référence) puis transformé en similarité en prenant son complément :

$$s_i = 1 - [|y_{i1} - y_{i2}|/R_i]$$

Finalement, ce coefficient peut être muni d'un artifice permettant d'éliminer de la comparaison les caractères pour lesquels l'information est non-disponible ou absente : le coefficient devient alors

$$S(x_1, x_2) = \frac{\sum_i w_{i12} s_{i12}}{\sum_i w_i}$$

pour les différents caractères *i*, où w_i est égal à 0 lorsque l'information, quant à ce caractère, est absente pour l'un ou l'autre des objets; $w_i = 1$ dans les autres cas. Par ailleurs, nous avons proposé (LEGENDRE et LEGENDRE, 1979) que w_i pourrait aussi prendre une valeur intermédiaire entre 0 et 1, de façon à donner des poids différentiels aux différents caractères, dans cette équation.

Un autre coefficient général de similarité a été proposé pour les études taxinomiques par ESTABROOK et ROGERS (1966). Ce coefficient, qui a la même forme générale que le coefficient de GOWER ci-dessus et qui comporte les mêmes facilités pour éliminer l'absence d'information ou pour donner des poids différents aux caractères, a été conçu de façon à permettre, en plus, la comparaison de caractères non-ordonnés, avec plus de finesse qu'un simple codage de ressemblance (1) ou de différence (0). Pour les caractères quantitatifs, ces auteurs proposent une autre formule que celle de GOWER, contenue dans l'équation suivante de similarité partielle pour chaque caractère :

$$s_{i12} = f(d_{i12}, k_i) = \frac{2(k + 1 - d)}{2k + 2 + dk} \quad \text{quand} \quad d \leqslant k$$
$$= 0 \quad \text{quand} \quad d > k$$

où d est la distance entre les deux objets quant à ce caractère, $|y_{i1} - y_{i2}|$, comme dans le coefficient de GOWER, et k est un paramètre, fixé par le taxinomiste pour chaque caractère, qui est égal au plus grand écart d pour lequel on désire avoir une similarité partielle plus grande que 0. On pourra consulter la littérature (par exemple : LEGENDRE et ROGERS, 1972; LEGENDRE et coll., 1972) pour étudier l'utilisation précise de cette mesure empirique de similarité partielle. Quant au cas des descripteurs non-ordonnés, lorsque le taxinomiste considère qu'il doit reconnaître une valeur non nulle de similarité partielle entre des descriptions différentes d'un caractère, ESTABROOK et ROGERS (1966) recommandent d'employer directement cette valeur comme s_i dans l'équation globale de similarité entre individus. Un exemple pourrait être le suivant : chez les Ciliés du super-ordre *Colpodidea*, l'arrangement de l'infraciliature buccale droite peut être décrite par un caractère formé des quatre descriptions qualitatives suivantes : 1. champ de rangées parallèles de cils; 2. cinétie à polarité inversée, morphogenèse somatique (du type *Platyophryia*); 3. cinétie à polarité inversée, morphogenèse semi-autonome (du type *Woodruffia*); 4. stichodyade.

En s'appuyant sur la morphogenèse, on pourrait considérer que le deuxième type ressemble partiellement aux types 3 et 4, alors que 3 ne ressemble pas à 4 et que 1 ne ressemble à aucun autre type. Cela peut se traduire par la matrice (symétrique) de similarité partielle suivante, dont le but est de modeler d'aussi près que possible les connaissances morphogénétiques acquises par le systématicien :

	1	2	3	4
1	1			
2	0	1		
3	0	0,4	1	
4	0	0,6	0	1

Ce tableau indique simplement qu'à deux spécimens, codés respectivement 2 et 3 pour ce caractère, on attribuera une valeur de similarité partielle de 0,4 qui s'additionnera aux valeurs de similarité partielle obtenues à l'aide des autres caractères; cette somme est finalement divisée par le nombre de caractères considérés (ou par la somme des n'_i si les caractères ont reçu des poids différents) pour donner la valeur totale de similarité de ces deux spécimens.

Chaque objet est ainsi comparé à chacun des autres objets de l'étude (spécimens ou taxa) et les valeurs ainsi obtenues sont assemblées en une matrice carrée (objets × objets) de similarités ou de distances, qui servira de base au groupement ou à l'ordination en espace réduit.

Les auteurs suivants présentent des revues de certains coefficients : COLE (1949, 1957), GOODMAN et KRUSKAL (1954, 1959, 1963), DAGNELIE (1960), SOKAL et SNEATH (1963), WILLIAMS et DALE (1965), CHEETHAM et HAZEL (1969), SNEATH et SOKAL (1973), ORLÓCI (1975), DAGET (1976), BLANC *et al.* (1976), LEGENDRE et LEGENDRE (1979).

4. — LE GROUPEMENT

Les groupements forment une famille de méthodes qui ont pour but d'effectuer une partition des objets en un certain nombre de classes, à partir des informations sur leurs ressemblances qui sont consignées dans la matrice de similarité (ou de distance) dont il a été question plus haut. Grouper des objets, c'est avant tout leur reconnaître un degré de similarité suffisant pour que ces objets soient réunis dans une même classe. Cette opération est toute naturelle pour le taxinomiste et elle constitue, le plus souvent, le premier but de son analyse.

Le *groupement* d'objets taxinomiques, fondé sur leur ressemblance globale quant à plusieurs caractères, est donc une opération de l'analyse multidimensionnelle qui consiste à « partitionner » la collection des objets de l'étude. Une *partition* est une division de la collection en sous-collections, telle que chaque objet appartienne à une et à une seule sous-collection, pour la partition en question (LEGENDRE et ROGERS, 1972). La classification des objets qui résulte du groupement peut être constituée d'une seule partition, ou, au contraire, de plusieurs partitions hiérarchisées des objets, selon le modèle de groupement employé (fig. 3).

Ces méthodes peuvent aussi être employées pour regrouper des caractères, mais nous nous limiterons ci-dessous au groupement des objets ou spécimens, pour la clarté de l'exposé.

Parmi toutes les méthodes de groupement, celle dont la logique apparaît la plus naturelle au spécialiste des sciences de la nature est le *groupement à liens simples* (LUKASZEWICZ, 1951; SNEATH, 1957; « *single linkage* » ou « *nearest neighbor clustering* »). Elle a comme point de départ une matrice de ressemblance quelconque entre les objets à grouper. On suppose que la mesure de ressemblance a été choisie avec soin,

Partition 1	Partition 2	Spécimens
Spécimens appartenant au genre A	Espèce 1	7, 12
	Espèce 2	3, 5, 11
	Espèce 3	1, 2, 6
Spécimens appartenant au genre B	Espèce 4	4, 9
	Espèce 5	8, 10, 13, 14

Fig. 3. — Exemple de deux partitions hiérarchisées d'un ensemble de spécimens : la première partition divise les spécimens selon leur genre et la deuxième, située hiérarchiquement sous la première, selon leur espèce.

de façon à tirer le meilleur parti de l'information taxinomique disponible. La méthode procède selon les étapes suivantes :

1) On récrit d'abord la matrice d'association (tableau I *a*) en ordre de similarité décroissante ou de distance croissante (tableau I *b*), mettant

TABLEAU I

Pour le groupement à liens simples, la matrice de similarité, a, calculée entre les cinq spécimens doit d'abord être récrite en ordre de similarité décroissante b.

a

Spéc.	Spécimens 212	214	233	431	432
212	1,000				
214	0,600	1,000			
233	0,000	0,071	1,000		
431	0,000	0,063	0,300	1,000	
432	0,000	0,214	0,200	0,500	1,000

b

S	Paires formées
0,600	212-214
0,500	431-432
0,300	233-431
0,214	214-432
0,200	233-432
0,071	214-233
0,063	214-431

en tête de liste les deux objets les plus similaires de la matrice d'association, puis la deuxième paire la plus similaire, et ainsi de suite jusqu'à ce que toutes les mesures comprises dans la matrice d'association aient été mises en ordre.

2) On forme ensuite les groupes de façon hiérarchique, en commençant par les deux objets les plus semblables, puis en laissant les objets s'agglomérer aux groupes, et les groupes s'agglutiner les uns aux autres, à mesure que l'on relâche le critère de similarité. L'exemple du tableau I donne naissance à deux types possibles de représentation ; les sous-graphes connexes à la fig. 4 *a* permettent de représenter chacun des liens de similarité unissant les objets au niveau considéré, alors que le dendrogramme de la fig. 4 *b* résume le groupement. Le dendrogramme est moins informatif que les sous-graphes puisque la position des liens entre groupes n'y est pas indiquée.

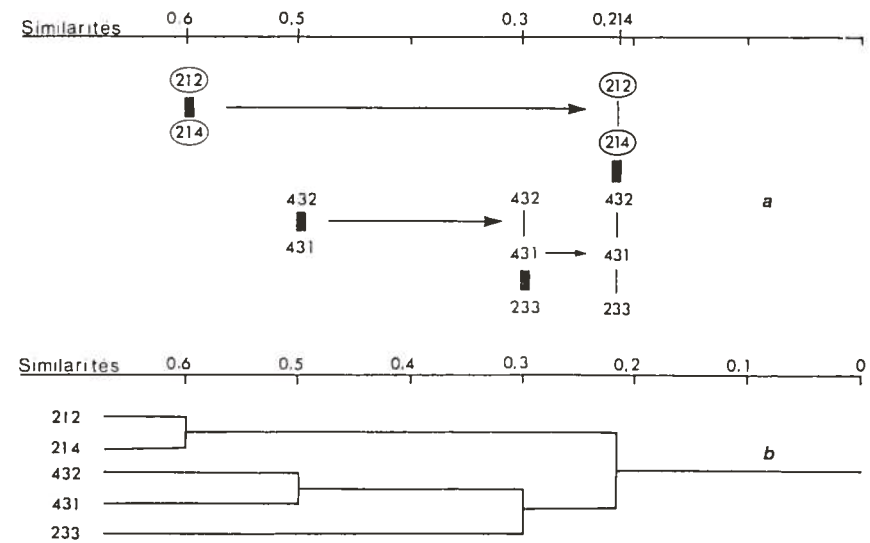

Fig. 4. — Les valeurs de similarité ordonnées du tableau I *b* permettent de grouper les objets à liens simples en relâchant le critère de similarité, ce qui peut être représenté par une série de sous-graphes connexes, *a* ou par un dendrogramme, *b*.

Dans ce modèle du groupement agglomératif à liens simples, deux groupes fusionnent si deux éléments, un de chaque groupe, atteignent la similarité de la partition considérée. Ce modèle rend compte de façon précise des relations entre paires d'objets voisins, mais il a aussi tendance à enchaîner les groupes les uns aux autres par des liens lâches. A l'opposé de cette méthode, on a donc conçu le groupement agglomératif à liens complets (SØRENSEN, 1948; « *complete linkage* » ou « *furthest neighbor sorting* ») dans lequel un objet ne peut se joindre à un groupe que lors-

qu'il est lié à tous les autres objets du groupe par des liens de similarité suffisants. Entre ces deux extrêmes ont été décrits différents types de groupement à liens intermédiaires.

D'autres modèles de groupement considèrent la moyenne arithmétique des valeurs de similarité, plutôt que le nombre de liens, ou encore procèdent selon un modèle géométrique dans lequel deux objets qui fusionnent sont remplacés par leur centroïde, ou centre de masse.

Il existe d'autres méthodes encore, faisant appel à différents critères pour décider de la fusion ou de la division des objets. Elles peuvent être « classées » en méthodes — séquentielles ou simultanées; — par agglomération ou par division; — monothétiques ou polythétiques; — hiérarchiques ou non; — probabilistes ou non et ainsi de suite, faisant ainsi état de la grande créativité qu'ont manifesté les taxinomistes pour résoudre le problème des classifications biologiques. SNEATH et SOKAL (1973) ainsi que LEGENDRE et LEGENDRE (1979) présentent des revues critiques de ces différentes méthodes.

5. — L'ORDINATION EN ESPACE RÉDUIT

Le taxinomiste peut aussi être intéressé d'obtenir une représentation graphique simplifiée de l'ensemble de la variabilité que recèle son échantillon. Nous verrons plus loin comment utiliser ces représentations en conjonction avec les résultats des groupements. L'ensemble des techniques permettant d'obtenir de telles représentations résumées porte le nom de méthodes d'ordination en espace réduit. ESCOUFIER (1975) les désigne aussi globalement sous le nom de méthodes de positionnement multidimensionnel. Ces méthodes impliquent la plupart du temps le calcul des valeurs propres et des vecteurs propres de la matrice décrivant la variabilité, et les premiers vecteurs propres, qui représentent les fractions les plus importantes de cette variabilité, forment donc ensemble un espace de référence qui résume en peu de dimensions une bonne partie de la variabilité des objets. Nous ne pouvons pas ici entrer dans le détail du calcul des vecteurs propres : il suffit cependant d'admettre pour la suite de cet exposé que les valeurs propres d'une matrice d'association sont des multiplicateurs de LAGRANGE (λ) par lesquels on calcule les valeurs optimales d'une expression sujette à certaines contraintes, et qui permettent de dégager les axes successifs de plus grande variabilité d'une matrice d'association et de « partitionner » cette variabilité en fractions orthogonales, indépendantes les unes des autres.

En particulier, dans l'*analyse en composantes principales*, on exprime la variabilité de l'échantillon sous la forme d'une matrice de dispersion (ou : de variances-covariances) entre les caractères quantitatifs mesurés. Les vecteurs propres identifient les directions (indépendantes) de plus grande variabilité à travers la co-variation des caractères, alors que

les valeurs propres décrivent la quantité de variance qu'explique chacun des axes principaux ainsi trouvés. Géométriquement, si on imagine les objets positionnés dans un espace multidimensionnel comportant autant d'axes qu'il y a de caractères dans l'étude, les objets ou spécimens de l'étude forment comme un nuage de points dans cet espace. La solution des composantes principales consiste à trouver d'abord la direction de plus grande variabilité de ce nuage de points; une fois celle-là trouvée, on cherche une seconde direction de plus grande variabilité, à angle droit avec la première. Et ainsi de suite, jusqu'à ce qu'on ait épuisé toute la variabilité disponible. C'est ainsi que le plan, formé par les deux premiers axes principaux, est aussi la meilleure projection en deux dimensions de notre nuage de points multidimensionnel, donc le plan qui résume le mieux la variabilité totale de la matrice de données taxinomiques (fig. 5 et 6).

La méthode des composantes principales produit donc simplement une rotation des axes de référence (rotation orthogonale si les vecteurs propres sont normés à 1) dans les directions successives de plus grande variabilité. Cette rotation est telle qu'elle préserve les distances euclidiennes entre les paires d'objets. Une autre méthode, qui porte le nom d'*analyse factorielle des correspondances*, préserve plutôt la distance du χ^2 entre les paires d'objets, à travers un même mécanisme de rotation d'axes. Cela peut avoir beaucoup d'intérêt en taxinomie, puisque la distance du χ^2 compare, non plus des valeurs brutes, mais des profils de probabilités conditionnelles pondérées. En pratique, avec des données taxinomiques, cette distance permet de comparer deux objets quant à diverses mesures brutes, sans se soucier, par exemple, des différences de variance entre les caractères. Pour l'analyse en composantes principales de telles données, on doit souvent exprimer ces dernières en valeurs relatives d'une mesure générale de taille, afin d'éliminer les variations allométriques.

Les composantes principales ainsi que l'analyse des correspondances sont des méthodes qui permettent d'obtenir une représentation des objets en espace réduit à partir de caractères *quantitatifs*. Lorsque les données ne sont pas de ce type, ou si, pour toute autre raison, le taxinomiste préfère exprimer la ressemblance entre ses spécimens par une mesure de similarité ou de distance de son choix, qui ne correspond pas nécessairement à une distance euclidienne ou à une distance du χ^2, il est alors possible de procéder à une ordination de ses objets en espace réduit à l'aide de méthodes de positionnement multidimensionnel qui préservent le mieux possible les relations mêmes de distance entre les objets qu'exprime la matrice de similarité que l'on vient de calculer. Deux modèles sont surtout employés pour ce faire : l'analyse des *coordonnées principales* (GOWER, 1976) est une technique de vecteurs propres qui produit une ordination en espace réduit à partir de toute matrice de similarité ou de distance entre objets, en autant que ces distances soient métriques. Lorsque les distances appartiennent plutôt

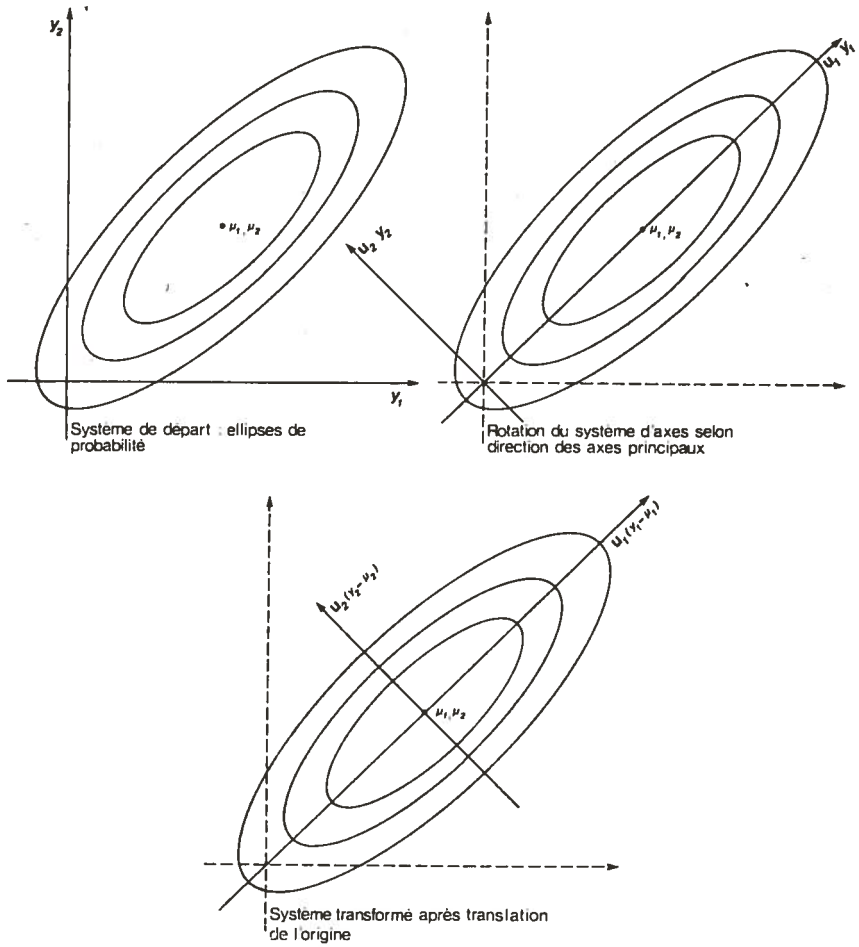

Fig. 5. — Attendu un nuage de points représenté ici par des ellipses d'équiprobabilité, l'analyse en composantes principales produit une rotation orthogonale des axes d'origine dans les directions successives de plus grande variance (D'après LEGENDRE et LEGENDRE, 1979).

à la classe des semimétriques (distances qui violent l'inégalité du triangle $D(a, b) + D(b, c) \geqslant D(a, c)$), on peut encore réaliser une ordination de ces objets en espace réduit par une méthode itérative d'approximation qui porte le nom de *cadrage multidimensionnel non métrique* (SHEPARD, 1962, 1966; KRUSKAL, 1964 *a*, 1964 *b*).

Enfin, lorsque le taxinomiste s'intéresse aux relations entre des taxa déjà constitués (β-taxinomie), il lui est loisible d'obtenir une représentation de ses spécimens dans un espace réduit de *fonctions discriminantes*. Ce modèle est apparenté à l'analyse en composantes principales, mais

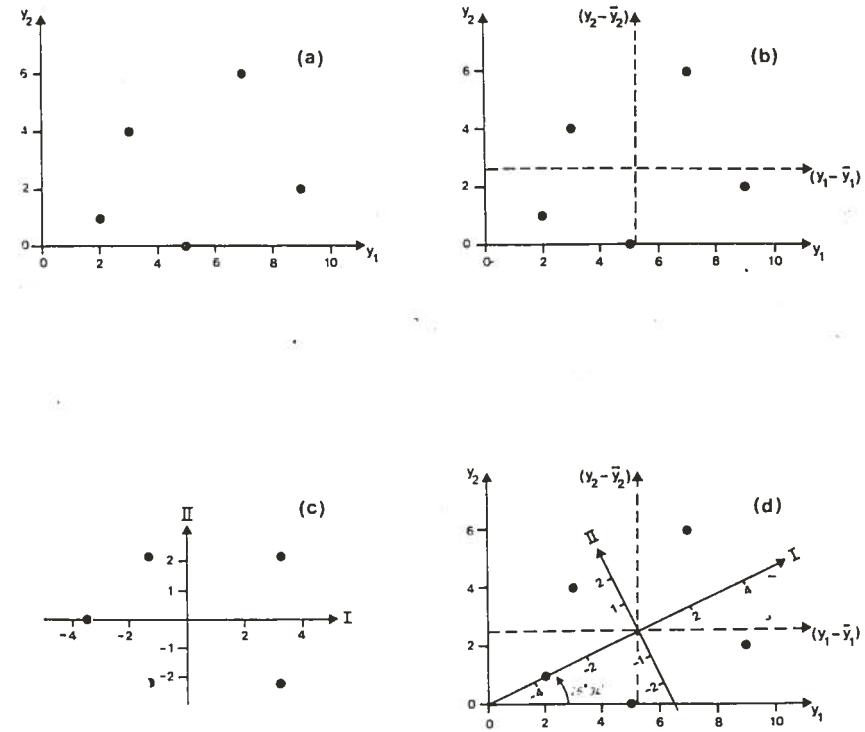

Fig. 6. — Exemple fictif d'analyse en composantes principales. (*a*), Les cinq objets disposés dans un graphique des deux descripteurs Y_1 et Y_2; (*b*), Le centrage des données produit les descripteurs centrés $(Y_1-\bar{Y}_1)$ et $(Y_2-\bar{Y}_2)$, représentés en tirets; (*c*), Les objets représentés dans le système des axes principaux I et II, centrés par rapport au nuage de points; (*d*), Les objets dans les différents systèmes d'axes (*b* et *c*) se superposent après une rotation de 26°34′ (D'après LEGENDRE et LEGENDRE, 1979).

il permet de maximiser les axes de la variabilité *entre les taxa* tout en minimisant la variabilité à l'intérieur des différents taxa. Ce modèle a son équivalent dans les mesures de ressemblance, puisqu'il est possible de calculer la distance entre deux groupes quant à l'ensemble des caractères mesurés, soit par la *distance générale* de MAHALANOBIS (1936), soit à l'aide du *coefficient de ressemblance raciale* de PEARSON (1926).

L'intérêt de telles représentations en espace réduit peut être illustré par un exemple, tiré de GATES et BERGER (1974), portant sur trois souches du Cilié *Tetrahymena pyriformis*. Une analyse en composantes principales de la matrice de dispersion de 27 caractères, exprimés en pourcentages de la taille, a permis de distinguer clairement les trois souches, comme l'illustre la figure 7. Les caractères provoquant surtout cette séparation des souches sont : — largeur maximum/largeur au niveau

du cinétosome antérieur; — largeur maximum/longueur totale; — largeur au niveau du cinétosome antérieur/longueur totale; les 24 autres mesures sont moins reliées à cette séparation des trois souches.

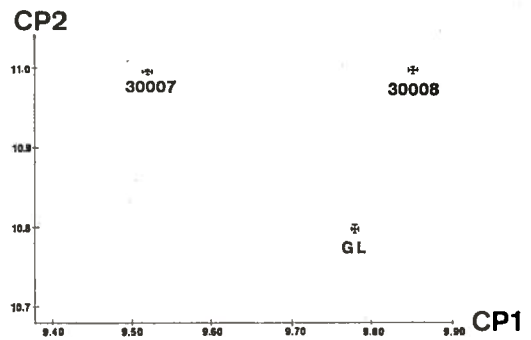

Fig. 7. — Position des souches 30007, 30008 et GL de *Tetrahymena pyriformis* dans l'espace des composantes principales 1 et 2 (CP1 et CP2) de la matrice de dispersion de 27 caractères. La moyenne des spécimens de chaque souche est entourée de l'intervalle de confiance de 99 pour 100 sur chacun des axes principaux. Modifié de GATES et BERGER (1974).

6. — GROUPEMENT ET ORDINATION

Alors que les ordinations en espace réduit fournissent une information quant aux axes principaux de la variabilité de l'échantillon, on a vu plus haut que les méthodes de groupement fournissent une information plus fine, souvent au niveau des paires d'objets (dans les groupements à liens), là où, à la limite, le groupement lui-même peut être faussé par un enchaînement excessif. C'est pourquoi plusieurs auteurs (GOWER et ROSS, 1969; ROHLF, 1970; SCHNELL, 1970; JACKSON et CROVELLO, 1971; LEGENDRE, 1976) ont proposé indépendamment de profiter des avantages des deux méthodes pour dégager la structure taxinomique, en associant groupement et ordination sur un même graphique. Un exemple tri-dimensionnel en est présenté à la figure 8, quoique, en général, deux dimensions suffisent.

7. — INTERPRÉTATION DE LA STRUCTURE TAXINOMIQUE

La structure taxinomique une fois découverte, il peut être intéressant de la mettre en relation, soit avec les caractères taxinomiques qui lui ont donné naissance (appelée l'information interne de la structure), soit encore avec d'autres caractères morphologiques, chémotaxinomiques, chromosomiques ou autres qui n'étaient pas considérés dans

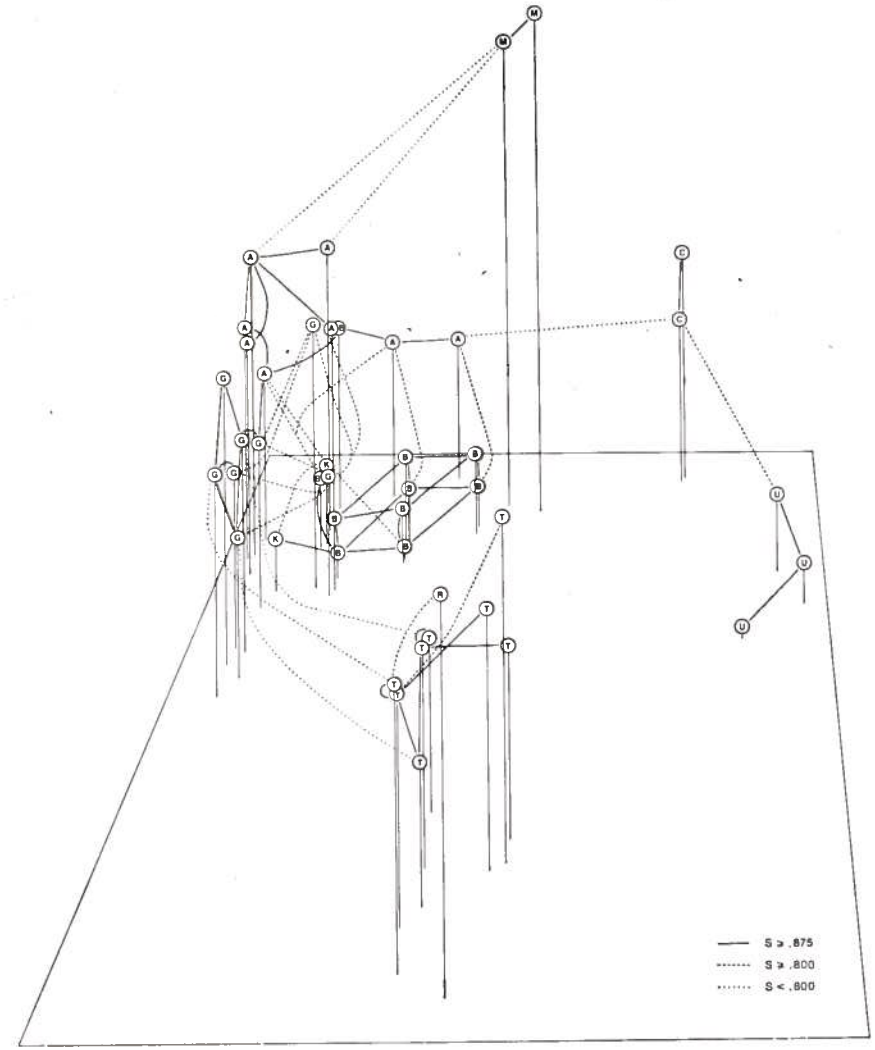

Fig. 8. — Chaîne de liens primaires d'un groupement à liens simples projetée sur une représentation dans un espace tri-dimensionnel de coordonnées principales. Le premier axe principal (profondeur) rend compte de 35 pour 100 des sommes de carrés, le second (largeur) de 20 pour 100 et le troisième (hauteur) de 16 pour 100. Les lettres représentent des spécimens tirés de neuf populations de Poissons Salmonidés de l'ouest de l'Amérique du Nord (D'après LEGENDRE, 1976).

l'analyse et qui constituent donc de l'information externe à la structure. Il peut enfin s'agir de descripteurs écologiques ou de comportement. Je ne dirai qu'un mot de cette phase de l'analyse qui demanderait un traitement beaucoup plus détaillé puisqu'elle débouche sur l'identifi-

cation automatique, d'une part, et sur l'écologie numérique, d'autre part.

Mentionnons tout de même que du point de vue de la mathématique impliquée, les fonctions discriminantes, dont il a été question plus haut, peuvent être employées pour trouver quels sont les caractères discriminants entre les taxa que l'on a délimités. Un exemple en est donné par GATES et BERGER (1974) qui, dans la même étude précédemment mentionnée, des trois souches de *Tetrahymena pyriformis*, ont trouvé par cette méthode cinq caractères discriminants, trois servant à distinguer la souche 30008 des deux autres, les deux derniers caractères établissant la différence entre les souches GL et 30007 (fig. 9).

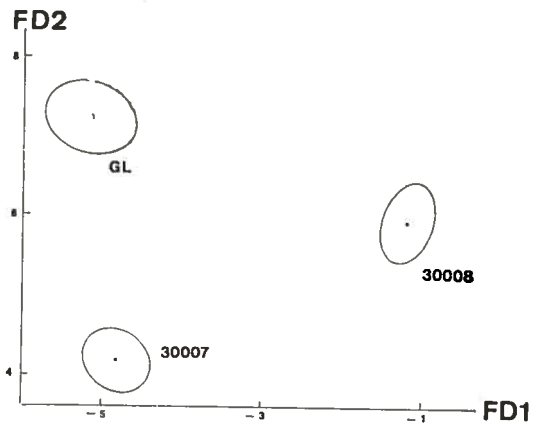

Fig. 9. — Position des souches 30007, 30008 et GL de *Tetrahymena pyriformis* dans l'espace des fonctions discriminantes 1 et 2 (FD1 et FD2) calculées à partir de 20 caractères quantitatifs. Chaque souche est représentée par son point moyen ainsi que par l'ellipse de confiance de 99 pour 100. Modifié de GATES et BERGER (1974).

En second lieu, mentionnons que chaque axe d'une ordination en espace réduit peut être considéré comme un caractère, sur le plan mathématique, et qu'il peut donc être mis en corrélation avec tout autre caractère métrique (par corrélation paramétrique) ou simplement ordonné (par corrélation non paramétrique, à employer aussi lorsque les relations caractères-ordination ne sont pas linéaires).

En troisième lieu, toute classification, qui est une partition des objets en classes mutuellement exclusives, peut être considérée elle-même comme un caractère non-ordonné, ce qui permet de la mettre en relation avec tout autre descripteur non-ordonné par un tableau de contingence. Ce sujet à lui seul demanderait un développement beaucoup plus important, développement que l'on trouvera au chapitre 4 du manuel de LEGENDRE et LEGENDRE (1979). Il convient toutefois de souligner l'importance de cette technique, puisqu'elle permet de *comparer*

entre elles différentes classifications des mêmes objets, obtenues à l'aide de méthodes ou de caractères différents.

Il convient enfin de souligner que les meilleures méthodes mathématiques ne sauraient mettre en évidence des différences entre espèces qui ne se trouvent pas d'abord dans les caractères employés comme base d'étude. C'est ainsi que GATES et BERGER (1976) ont conclu en l'impossibilité de séparer sur des caractères morphologiques les Ciliés *Paramecium primaurelia* et *P. pentaurelia* à la suite d'une analyse en composantes principales, alors qu'il avait été possible de distinguer par cette même méthode d'autres espèces appartenant à ce complexe d'espèces biologiques (GATES et coll., 1975; POWELSON et coll., 1975).

8. — ÉTUDE PHYLÉTIQUE

Nous ne pouvons terminer ce survol de la taxinomie numérique sans évoquer la construction des arbres décrivant les relations phylétiques entre les taxa, puisque les classifications de niveau supra-spécifique tendent à décrire ou du moins prétendent souvent décrire, tant bien que mal, les chemins parcourus par l'évolution des espèces. GATES (1978 *a*) a défini la phylogenèse comme le processus, opérant sur la variabilité naturelle des populations, qui a pour résultat de diviser celles-ci en sous-populations.

Une certaine confusion existe dans la terminologie relative à ces études, comme l'ont noté SNEATH et SOKAL (1973). En effet, les relations dégagées par les études taxinomiques classiques, avec le matériel biologique contemporain, sont de niveau *phénétique* lorsque la ressemblance entre individus y est établie à l'aide de caractéristiques phénotypiques, alors que d'autres études visent à dégager des relations *cladistiques* qui rendent compte des relations de descendance d'un ancêtre commun. Certains auteurs, à la suite de HENNIG (1966), aimeraient que la phylogenèse ne soit exprimée qu'en termes cladistiques alors que d'autres, comme MAYR (1965), veulent y exprimer aussi la divergence phénétique qui a pu se produire depuis le point de séparation des lignées. Ces différences d'écoles montrent essentiellement l'incapacité du système de nomenclature actuel d'exprimer toutes les facettes de l'évolution biologique mais ne sauraient en aucun cas empêcher la systématique de progresser vers une meilleure compréhension de ces relations phylétiques, même en l'absence d'évidences fossiles, comme c'est la règle en protistologie. Il peut donc arriver qu'une classification phylétique de type cladistique ne soit pas en accord avec la classification phénétique, lorsque se posent certains problèmes pour lesquels ces deux méthodes connaissent des différences conceptuelles. Par conséquent, il importe de séparer les deux démarches et d'identifier clairement celle que l'on suit, même si les deux emploient, en bonne partie, les mêmes méthodes

de calcul, la différence se situant au niveau des hypothèses de base ainsi qu'à celui des conclusions, comme nous le verrons.

Les travaux en taxinomie cladistique semblent représenter au moins quatre écoles de pensée.

1. La première méthode proposée, celle d'EDWARDS et CAVALLI-SFORZA (1964), mesure la distance entre taxa quant à la fréquence allélique de différents gènes. Ces fréquences sont donc des caractères quantitatifs. A partir de cette matrice de distances, ils calculent un arbre de longueur minimum représentant les relations phylétiques en un réseau qui ne postule pas des ancêtres communs.

2. Avec CAMIN et SOKAL (1965) sont apparus les cladogrammes, c'est-à-dire des arbres comportant des nœuds ou embranchements, qui sont les ancêtres communs supposés des lignées divergentes. Ces nœuds sont appelés des hypothèses d'évolution par ESTABROOK (1968) ou des HTU (« *hypothetical taxonomic unit* ») par FARRIS (1970). CAMIN et SOKAL emploient des caractères discontinus comme données de base et cherchent le cladogramme le plus parcimonieux capable de produire le vecteur de description de chaque objet. Une solution topologique au problème de l'arbre de CAMIN et SOKAL a été trouvée par ESTABROOK, en 1968.

3. Des solutions analytiques à la théorie cladistique de HENNIG ont été développées par FARRIS et son groupe depuis 1966. Ils utilisent des caractères continus aussi bien que codés pour mesurer la ressemblance entre objets à l'aide de la distance de Manhattan, puis établissent des graphes et des arbres de longueur minimum qui représentent des estimations des séries évolutives.

4. Les deux dernières décennies ont connu des progrès importants dans la connaissance des séquences d'acides aminés dans les protéines, ce qui a conduit des chercheurs et, en particulier FITCH et MARGOLIASH (1967), à calculer des cladogrammes à partir des différences entre les séquences de protéines, qui permettent d'estimer la distance minimale de mutation entre objets.

Tous les graphes ou arbres construits par les cladistes ont ceci de commun qu'ils cherchent à être le plus parcimonieux possible : cela ne signifie pas que les cladistes prétendent que la parcimonie est une caractéristique des mécanismes de l'évolution, mais simplement que la parcimonie (l'arbre le plus court possible) donne un critère permettant de choisir entre plusieurs solutions mathématiquement équivalentes. Les autres hypothèses nécessaires à la formation de cladogrammes sont : — 1. La nature des ancêtres (l'hypothèse d'évolution mentionnée plus haut) qui sont postulés, ou qui sont plus souvent calculés à partir des données sur les objets de l'étude, à l'aide d'un mode de calcul déterminé; — 2. Les types de branchements qui sont permis : branchements binaires seulement ou branchements multiples qui permettent d'obtenir des arbres plus courts; — 3. Des inversions sont-elles permises

le long de l'arbre dans l'ordre supposé d'évolution des caractères ? — 4. Il faut enfin postuler quels sont les états primitifs et les états avancés de chaque caractère. Ceci peut être établi, par le biologiste, à l'aide de quatre types d'évidence qu'a décrits HENNIG (1966) : *a*) lorsque des données paléontologiques sont disponibles, ce qui ne saurait être le cas en protistologie, les états primitifs peuvent être associés aux strates écologiques les plus anciennes; *b*) les séries chorologiques, décrivant la progression géographique des espèces, peuvent aussi indiquer quels sont les états primitifs; *c*) le développement embryonnaire des pluricellulaires, de même que la morphogenèse des unicellulaires, peut indiquer la nature des caractères primitifs, suivant la loi de la récapitulation de HAECKEL; *d*) enfin, l'évolution d'un caractère, lorsqu'elle est connue, peut fournir des indications quant à l'évolution d'autres caractères.

Les travaux de phylogenèse cladistique ont été peu nombreux jusqu'ici chez les Protistes. Cependant, GATES (1978 *a*) en a établi le cadre théorique et il en a montré l'intérêt par l'exemple d'un cadre cladistique établi pour le genre *Euplotes* (Ciliés) en suivant les principes de HENNIG résumés plus haut. Les séquences d'évolution de différents caractères ont été établies soit à l'aide d'évidences biogéographiques, soit par comparaison de caractères avec d'autres dont l'évolution était connue. L'établissement de l'arbre évolutif le plus parcimonieux lui a permis de résumer la variabilité, décrite jusqu'ici en quelque 51 espèces morphologiques (ou typologiques) qui ne représenteraient, selon cet auteur, qu'une dizaine d'espèces vraies. LYNN (1978) a établi par la même méthode le cladogramme reliant cinq espèces de Ciliés des genres *Colpoda* et *Tillina*.

9. — CONCLUSION

En résumé, la taxinomie numérique n'est qu'un instrument à la disposition des systématiciens, comme le sont la cytogénétique, la génétique expérimentale, ou encore les études taxinomiques par voie biochimique ou d'ultrastructure. Cet instrument a cependant la caractéristique de permettre au systématicien de résumer la variabilité observée par l'une ou l'autre, ou par plusieurs des méthodes analytiques à sa disposition. Cette méthode peut être employée soit à des fins de classification stricte, soit pour des études de phylogenèse, la différence entre ces deux branches de la taxinomie numérique ne résidant d'ailleurs pas tant dans les méthodes employées pour résumer la variabilité que dans les hypothèses de base, qui déterminent les conclusions que l'on peut tirer de telles études. Ceci supporte la règle à l'effet qu'il faut avant tout être un systématicien pour faire de la systématique, le support du mathématicien, du biochimiste ou du microscopiste électronicien étant surtout dirigé plutôt vers le développement des méthodes propres à son art que vers l'interprétation taxinomique ou évolutionnaire pro-

prement dite. En effet, les conclusions systématiques peuvent différer des conclusions strictement numériques puisque la biosystématique est avant tout une démarche intégrant des évidences provenant de sources diverses : je ne citerai comme exemples que les espèces morphologiquement semblables du genre *Paramecium* que l'on peut cependant distinguer par leur isolement reproducteur, ou encore la variabilité morphologique exceptionnelle observée chez *Tetrahymena paravorax*.

Les travaux réalisés depuis quelques années en taxinomie numérique des Protistes montrent l'intérêt de cette approche pour le protistologue qui, déjà en possession de certaines données de base sur la morphologie et l'ultrastructure de son matériel, ainsi que des connaissances nécessaires sur la valeur relative des caractères, du moins chez les Ciliés (GRAIN, 1977), devrait y trouver un moyen d'intérêt pour la synthèse de l'information systématique.

BIBLIOGRAPHIE

BERGER (J.), 1978. — Multivariate morphometric characterization of Ciliophoran species. *Biosystems*, **10**, 65.

BLANC (F.), CHARDY (P.), LAUREC (A.), REYS (J.-P.), 1976. — Choix des métriques qualitatives en analyse d'inertie. Implications en écologie marine benthique. *Mar. Biol. (Berl.)*, **35**, 49-67.

CAMIN (J. H.), SOKAL (R. R.), 1965. — A method for deducing branching sequences in phylogeny. *Evolution*, **19**, 311-326.

CHEETHAM (A. H.), HAZEL (J. E.), 1969. — Binary (presence-absence) similarity coefficients. *J. Paleontol.*, **43**, 1130-1136.

COLE (L. C.), 1949. — The measurement of interspecific association. *Ecology*, **30**, 411-424.

COLE (L. C.), 1957. — The measurement of partial interspecific association. *Ecology*, **38**, 226-233.

DAGET (J.), 1976. — *Les modèles mathématiques en écologie*. Masson, Paris. Collection d'Écologie, n° 8, 172 p.

DAGNELIE (P.), 1960. — Contribution à l'étude des communautés végétales par l'analyse factorielle. *Bull. Serv. Carte Phytogéogr.*, **B5**, 7-71, 93-195.

EDWARDS (A. W. F.), CAVALLI-SFORZA (L. L.), 1964. — Reconstruction of evolutionary trees. P. 67-76 *in* : HEYWOOD (V. H.), MCNEILL (J.), eds., *Phenetic and Phylogenetic Classification*. Syst. Ass. Pub. 6. 164 p.

ESCOUFIER (Y.), 1975. — Le positionnement multidimensionnel. *Revue de Statistique appliquée*, **23**, 5-14.

ESTABROOK (G. F.), 1968. — A mathematical model in graph theory for biological classification. *J. theor. Biol.*, **12**, 297-310.

ESTABROOK (G. F.), ROGERS (D. J.), 1966. — A general method of taxonomic description for a computed semilarity measure. *Bioscience*, **16**, 789-793.

FARRIS (J. S.), 1966. — Estimation of conservatism of characters by constancy within biological populations. *Evolution*, **20**, 587-591.

FARRIS (J. S.), 1970. — Methods for computing Wagner trees. *Syst. Zool.*, **19**, 83-92.

FITCH (W. M.), MARGOLIASH (E.), 1967. — Construction of phylogenetic trees. *Science*, **155**, 279-284.

GATES (M. A.), 1978 *a*. — An essay on the principles of ciliate systematics. *Trans. Amer. Micros. Soc.*, **97**, 221-235.

GATES (M. A.), 1978 *b*. — Morphometric variation in the hypotrich genus *Euplotes*. *J. Protozool.*, **25**, 338-350.

GATES (M. A.), BERGER (J.), 1974. — A biometric study of three strains of *Tetrahymena pyriformis* (Ciliatea : Hymenostomatida). *Can. J. Zool.*, **57**, 1167-1183.

GATES (M. A.), BERGER (J.), 1976. — Morphological inseparability of *Paramecium primaurelia* and *Paramecium pentaurelia*. *Trans. Amer. Micros. Soc.*, **95**, 507-514.

GATES (M. A.), POWELSON (E. E.), BERGER (J.), 1975. — Syngenic ascertainment in *Paramecium aurelia*. *Syst. Zool.*, **23**, 482-489.

GOODMAN (L. A.), KRUSKAL (W. H.), 1954. — Measures of association for cross classification. *J. amer. statist. Ass.*, **49**, 732-764.

GOODMAN (L. A.), KRUSKAL (W. H.), 1959. — Measures of association for cross classification. II. Further discussion and references. *J. amer. statist. Ass.*, **54**, 123-163.

GOODMAN (L. A.), KRUSKAL (W. H.), 1963. — Measures of association for cross classification. III. Approximate sampling theory. *J. amer. statist. Ass.*, **58**, 310-364.

GOWER (J. C.), 1966. — Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325-338.

GOWER (J. C.), 1971. — A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857-871.

GOWER (J. C.), ROSS (G. J. S.), 1969. — Minimum spanning trees and single linkage cluster analysis. *Appl. Statist.*, **18**, 54-64.

GRAIN (J.), 1977. — Les critères actuellement utilisables dans l'établissement de la classification des protozoaires ciliés. P. 24 *in* : *Abstracts, Fifth intern. Congress Protozoology*, New York.

HENNIG (W.), 1966. — *Phylogenetic systematics*. Univ. of Illinois Press, Urbana. 263 p.

JACCARD (P.), 1908. — Nouvelles recherches sur la distribution florale. *Bull. Soc. vaudoise Sci. nat.*, **44**, 223-270.

JACKSON (R. C.), CROVELLO (T. J.), 1971. — A comparison of numerical and biosystematic studies in *Haplopappus*. *Brittonia*, **23**, 54-70.

KRUSKAL (J. B.), 1964 *a*. — Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1-27.

KRUSKAL (J. B.), 1964 *b*. — Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115-129.

LEGENDRE (L.), LEGENDRE (P.), 1979. — *Écologie numérique*. Tome 1 : *Le traitement multiple des données écologiques*. Tome 2 : *La structure des données écologiques*. Masson, Paris et les Presses de l'Université du Québec. Collection d'Écologie, nos 12 et 13. 196 et 252 p.

LEGENDRE (P.), 1976. — An appropriate space for clustering selected groups of western North American *Salmo*. *Syst. Zool.*, **25**, 193-195.

LEGENDRE (P.), ROGERS (D. J.), 1972. — Characters and clustering in taxonomy: a synthesis of two taximetric procedures. *Taxon*, **21**, 567-606.

LEGENDRE (P.), SCHRECK (C. B.), BEHNKE (R. J.), 1972. — Taximetric analysis of selected groups of western North American *Salmo* with respect to phylogenetic divergences. *Syst. Zool.*, **21**, 292-307.

LUKASZEWICZ (J.), 1951. — Sur la liaison et la division des points d'un ensemble fini. *Colloquium math.*, **2**, 282-285.

LYNN (D. H.), 1978. — Size increase and form allometry during evolution of ciliate species in the genera *Colpoda* and *Tillina* (Ciliophora : Colpodida). *Biosystems*, **10**, 201-211.

MAGNUS (R.), 1949. — *Goethe as a scientist*. Translated by Norden (H.). Henry Schuman, New York. 259 p.

MAHALANOBIS (P. C.), 1936. — On the generalized distance in statistics. *Proc. natl. Inst. Sci. India*, **2**, 49-55.

MAYR (E.), 1965. — Numerical phenetics and taxonomic theory. *Syst. Zool.*, **14**, 73-97.
ORLÓCI (L.), 1975. — *Multivariate analysis in vegetation research*. Dr. W. JUNK B. V., The Hague. 276 p.
PEARSON (K.), 1926. — On the coefficient of racial likeness. *Biometrika*, **18**, 105-117.
POWELSON (E. E.), GATES (M. A.), BERGER (J.), 1975. — A biometrical analysis of 22 stocks of four syngens of *Paramecium aurelia*. *Can. J. Zool.*, **53**, 19-32.
ROGERS (D. J.), TANIMOTO (T. T.), 1960. — A computer program for classifying plants. *Science* (*Wash. D. C.*), **132**, 1115-1118.
ROHLF (F. J.), 1970. — Adaptive hierarchical clustering schemes. *Syst. Zool.*, **19**, 58-82.
SCHNELL (G. D.), 1970. — A phenetic study of the suborder Lari (Aves). I. Methods and results of principal components analyses. *Syst. Zool.*, **19**, 35-57.
SHEPARD (R. N.), 1962. — The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, **27**, 125-139, 219-246.
SHEPARD (R. N.), 1966. — Metric structures in ordinal data. *J. math. Psychol.*, **3**, 287-315.
SNEATH (P. H. A.), 1957. — The application of computers to taxonomy. *J. gen. Microbiol.*, **17**, 201-226.
SNEATH (P. H. A.), SOKAL (R. R.), 1973. — *Numerical taxonomy—the principles and practice of numerical classification*. W. H. Freeman, San Francisco. 573 p.
SOKAL (R. R.), MICHENER (C. D.), 1958. — A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.*, **38**, 1409-1438.
SOKAL (R. R.), SNEATH (P. H. A.), 1963. — *Principles of numerical taxonomy*. W. H. Freeman, San Francisco. 359 p.
SØRENSEN (T.), 1948. — A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons. *Biol. Skr.*, **5**, 1-34.
TAYLOR (W. D.), 1978 *a*. — Growth responses of ciliate protozoa to the abundance of their bacterial prey. *Microb. Ecol.*, **4**, 207-214.
TAYLOR (W. D.), 1978 *b*. — Maximum growth rate, size and commonness in a community of bactivorous ciliates. *Oecologia* (*Berl.*), **36**, 263-272.
WILLIAMS (W. T.), DALE (M. B.), 1965. — Fundamental problems in numerical taxonomy. *Adv. bot. Res.*, **2**, 35-68.

DISCUSSION

P. DE PUYTORAC. — Il n'est peut-être pas inutile de rappeler la position de la taxinomie numérique par rapport aux autres systèmes taxinomiques, comme l'a fait GUINOT (1977).

Dans la *systématique traditionnelle*, chaque espèce étant immuable, on établissait des archétypes sur des critères morphologiques. Dans la *systématique phénétique*, comme il vient d'être expliqué, les taxons sont construits en calculant statistiquement les concordances d'après un très grand nombre de caractères considérés comme homologues. Dans la *systématique évolutionniste*, on veut classer les organismes et non pas les caractères, en appréciant la totalité des caractères comme un ensemble intégré. Le but est la découverte de groupements naturels dont les éléments sont réunis par une origine commune. Alors que la classification par grade tient compte des ressemblances dues à l'identité des niveaux atteints par diverses formes lors d'une évolution parallèle, le classement par clade tient compte des ressemblances dues à l'origine commune de deux ou plusieurs éléments d'un groupe. Dans la

systématique cladistique, l'exigence de base est celle d'une origine monophylétique de toutes les unités taxinomiques : le taxon doit être seulement composé des espèces ayant un ancêtre commun; un taxon monophylétique doit comprendre toutes les espèces qui descendent de l'espèce souche. La cladogenèse se manifeste par deux phénomènes évolutifs : la *stasigenèse* ou préservation des caractères primitifs (plésiomorphes) et l'*anagenèse* ou apparition et développement successif de caractères nouveaux (apomorphes). Dans l'analyse cladistique, il faut donc faire la distinction entre caractères plésiomorphes et caractères apomorphes et considérer que tous les caractères apomorphes n'ont pas la même valeur pour le systématicien.

Les critères de distinction entre caractères plésiomorphes et apomorphes peuvent être l'antériotité paléontologique, l'antériorité dans l'ontogenèse, la direction évolutive d'une série de transformations de caractères homologues, la biogéographie. Parmi les difficultés qui sont posées alors au taxinomiste, il faut retenir la convergence due à l'apparition de caractères analogues mais non homologues, dans deux taxons n'appartenant pas au même groupe naturel, et le parallélisme, apparition indépendante de caractères similaires chez deux ou plusieurs membres d'un groupe naturel. Le recours aux ordinateurs, dans la systématique phénétique, pour être conciliable avec la systématique évolutionniste devrait prendre en compte, dans la programmation, une hiérarchie des caractères, ce qui est contraire au principe selon lequel, *a priori*, tous les caractères ont une valeur égale.

J. GRAIN. — S'il n'est pas fait de distinctions entre les homologies et les convergences, peut-on vraiment accorder une valeur phylogénétique à une classification fondée sur la taxinomie numérique ?

P. LEGENDRE. — Oui et non. Tout dépend de la structure qu'on a donnée à la classification. C'est au taxinomiste de savoir quel modèle correspond le mieux au phénomène qu'il veut mettre en évidence et quelles hypothèses il veut intégrer. Si on admet, au départ, une classification phénétique, les conclusions ne pourront pas donner d'interprétation évolutive. Si on choisit, au contraire, d'ordonner les caractères, on pourra espérer s'approcher du tracé d'un ordre phylogénétique.

J. GENERMONT. — Il peut y avoir des convergences, mais en travaillant sur un grand nombre de caractères, ce qui est un des principes de la taxinomie numérique, statistiquement, on peut espérer s'approcher des relations phylogénétiques.

P. CHARDY. — A l'issue d'une classification phénétique, on peut dégager les caractères déterminants de cette classification et s'interroger alors, sans *a priori*, sur la signification phylétique de ces caractères, s'ils en ont une.

M. M. COUTEAUX, J. F. PONGE et A. MUNSCH (Laboratoire d'Écologie générale, Muséum National d'Histoire Naturelle, 91800 Brunoy). — Grâce à la microscopie à balayage, il a été possible de voir, chez des individus du genre *Euglypha*, des structures peu ou pas discernables en microscopie optique. La place et la valeur de ces structures doivent être précisées dans la classification. On peut aisément les traduire en données biométriques. Il nous a paru donc souhaitable de faire une tentative de taxinomie numérique.

Quatre-vingt-onze individus du genre *Euglypha*, provenant de 24 échantillons de sol d'origines diverses, ont servi d'unités de base à des analyses des correspondances ayant pour but de mettre en évidence leurs ressemblances et leurs divergences et à les représenter graphiquement. Au départ, un maximum de données ont été réunies, soit 37 caractères concernant la taille et la forme de la thèque, des écailles pariétales et des écailles buccales. L'analyse a montré que les données relatives à la forme de la thèque et des écailles pariétales n'avaient aucun pouvoir discriminant; aussi ont-elles été éliminées pour ne conserver que 19 caractères ayant trait à la taille de la thèque, des écailles pariétales et des écailles buccales et à la forme des écailles buccales.

Deux analyses des correspondances ont été réalisées, l'une sur l'ensemble des 19 caractères, l'autre uniquement sur les écailles buccales. Les deux analyses mettent en évidence des groupes d'individus où l'on peut reconnaître *Euglypha hyalina*, *E. rotunda* var. *minor*, *E. cristata*, *E. cristata* f. *decora*, *E. capsiosa*, *E. simplex*, *E. filifera* et *E. strigosa*. L'axe 1 dispose les individus selon un gradient de taille. L'axe 2 concerne la présence d'une digitation à l'extrémité postérieure de l'écaille buccale et la présence d'un épaississement à son extrémité antérieure. L'analyse des correspondances sur les écailles buccales seulement établit mieux la distinction entre les espèces. On voit donc l'importance de la morphologie des écailles buccales comme facteur discriminant dans la classification des *Euglypha*.

D. Izard, F. Gavini et P. A. Trinel (INSERM, Unité nº 146, CERTIA, 59650 Villeneuve-d'Ascq). — Nous avons appliqué les méthodes modernes de taxinomie à la classification bactérienne. La démarche empruntée dans la classification des *Enterobacteriaceae* se décompose en 3 étapes : — 1) Étude par analyse numérique d'environ 200 souches appartenant à un même genre et provenant d'origines diverses (8, 6, 7, 5). Le nombre des caractères phénotypiques étudiés est voisin de 250. Cette étape conduit à la définition de phénons; certains correspondent à des espèces bien connues, d'autres semblent représenter des entités (sur le plan phénotypique) nouvelles; — 2) Étude génétique : dans un premier temps, le GC pour 100 moyen des groupes nouveaux est mesuré (2, 1, 10). Cette détermination donne une première idée sur l'homogénéité génétique des phénons nouveaux. Dans un deuxième temps, l'homogénéité génétique est étudiée par hybridation ADN/ADN. L'ADN de la souche type définie lors de l'étude phénotypique est marqué, puis hybridé avec un échantillonnage d'ADN représentant au moins 50 pour 100 des souches du phénon (3, 5, 12, 13, 11, 14). Dans un troisième temps, les relations génétiques du phénon nouveau sont étudiées en hybridant l'ADN marqué de la souche type avec des ADN extraits des différentes espèces de la famille des *Enterobacteriaceae* (3, 5, 11, 12, 13, 14). Dans un quatrième temps, la mesure relative de la longueur du chromosome de la souche type est effectuée par hybridation réciproque avec d'autres souches types représentant différentes espèces (11); — 3) Étude de para-génétique : elle consiste à étudier les variations immunologiques (traduisant par conséquent des variations au niveau de la structure primaire) d'une enzyme clé du métabolisme glucidique, la glucose 6-phosphate déshydrogénase (G 6-PD). Par réaction de microfixation du complément, on peut ainsi définir tout comme la distance phénotypique et la distance génétique (= pour 100 d'hybridation), la distance immunologique. A l'heure actuelle, ces travaux ont conduit à la mise en évidence d'un genre nouveau, *Rahnella* (12), de 3 espèces nouvelles, *Enterobacter amnigena* (14), *Enterobacter intermedium* (11) et *Serratia fonticola* (5). Deux groupes atypiques sur le plan des caractères phénotypiques ont été reliés à des espèces connues : *Klebsiella pneumoniae* mobile et acétoïne négative (3), *Enterobacter cloacae* acétoïne négative (14). Trois autres phénons nouveaux font l'objet d'études d'hybridation ADN/ADN : F, K et L.

Une clé d'identification de ces espèces nouvelles, fondée sur l'utilisation d'un ordinateur a été proposée; la matrice de cette clé est formée de 20 caractères; l'identification est pondérée par une probabilité.

Ces travaux débouchent sur deux aspects : l'étude des germes de l'environnement (la plupart des groupes nouveaux sont issus de l'environnement) et l'établissement d'une classification naturelle.

Références mentionnées dans la dernière intervention.

1. Ferragut (C.), Gavini (F.), Izard (D.) et Leclerc (H.), 1978. — Étude du pour 100 GC dans un groupe d'entérobactéries H_2S^- apparentées au genre *Citrobacter*. *Can. J. Microbiol.*, **24**, 473-479.
2. Ferragut (C.) et Leclerc (H.), 1976. — Étude comparative des méthodes de détermination du Tm de l'ADN bactérien. *Ann. Microbiol.* (Inst. Pasteur), **127 A**, 223-235.
3. Ferragut (C.) et Leclerc (H.), 1978. — Study of motile and negative acetoin *Klebsiella pneumoniae* strains. *Antonie van Leeuwenhoek*, **44**, 407-424.
4. Gavini (F.) (résultats non publiés).
5. Gavini (F.), Ferragut (C.), Izard (D.), Trinel (P. A.), Leclerc (H.), Lefebvre (B.) et Mossel (D. A. A.), 1978. — Studies of *Serratia fonticola* sp. nov. isolated from the aquatic environment. *Int. J. Syst. Bacteriol.* (envoyé pour publication).
6. Gavini (F.), Ferragut (C.) et Leclerc (H.), 1976. — Étude taxonomique d'entérobactéries appartenant ou apparentées au genre *Enterobacter*. *Ann. Microbiol.* (Inst. Pasteur), **127 B**, 317-335.
7. Gavini (F.), Leclerc (H.), Lefebvre (B.), Ferragut (C.) et Izard (D.), 1977. — Étude taxonomique d'entérobactéries appartenant ou apparentées au genre *Klebsiella*. *Ann. Microbiol.* (Inst. Pasteur), **128 B**, 45-59.
8. Gavini (F.), Lefebvre (B.) et Leclerc (H.), 1976. — Positions taxonomiques d'entérobactéries H_2S^- par rapport au genre *Citrobacter*. *Ann. Microbiol.* (Inst. Pasteur), **127 A**, 275-295.
9. Izard (D.), 1976. — Les techniques d'hybridation en taxonomie bactérienne. Mémoire pour l'obtention du Certificat d'Études Supérieures de Bactériologie, Faculté de Pharmacie de Lille, 1976.
10. Izard (D.), Ferragut (C.), Gavini (F.) et Leclerc (H.), 1978. — Variations of the moles percent guanine plus cytosine within a group of *Enterobacteriaceae* belonging or related to the genus *Enterobacter*. *Int. J. Syst. Bacteriol.*, **28**, 449-452.
11. Izard (D.), Gavini (F.) et Leclerc (H.), 1979. — Polynucleotide sequence relatedness among *Enterobacter intermedium* sp. nov. (type strain : CIP 79-27) and the species *Enterobacter cloaceae* and *Klebsiella pneumoniae*; influence of the genome sige. *Zent. Blatt* (envoyé pour publication).
12. Izard (D.), Gavini (F.), Trinel (P. A.) et Leclerc (H.), 1978. — *Rhanella aquatilis* un nouveau membre dans la famille des *Enterobacteriaceae*. *Ann. Microbiol.* (Inst. Pasteur) (accepté pour publication).
13. Izard (D.), Gavini (F.), Trinel (P. A.) et Leclerc (H.), 1978. — Étude d'un groupe nouveau d'*Enterobacteriaceae* (groupe H_1) apparenté à l'espèce *E. cloacae*. *Can J. Microbiol.* (envoyé pour publication).
14. Izard (D.), Gavini (F.), Trinel (P. A.) et Leclerc (H.), 1979. — Deoxyribonucleic acid relatedness between *Enterobacter cloaceae* and *Enterobacter amnigena* sp. nov. with a description of *E. amnigena* (type strain : ATCC 33072). *Int. J. Syst. Bacteriol.* (envoyé pour publication).