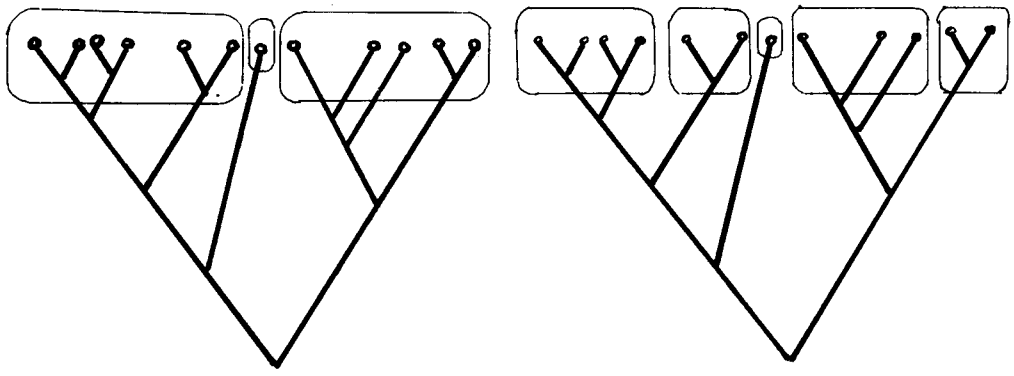
Ann. Sc. Math. Québec, 1977, vol. I, No. 2, pp. 325-336

D'UNE CLASSIFICATION HIERARCHIQUE A UNE CLASSIFICATION DISJOINTE

David Sankoff

I INTRODUCTION

Comment transformer une classification hiérarchique en une classification non-hiérarchique, en un nombre de groupes disjoints? Nous suggérerons deux conditions, l'une plus forte que l'autre, de compatibilité entre une hiérarchie et une partition ordinaire, et nous evaluerons quatre solutions à notre problème à l'aide de ces conditions. Nous porterons une attention particulière à une méthode conçue spécifiquement pour produire une partition dans laquelle les groupes sont de taille uniforme, propriété utile pour la taxonomie, surtout la taxonomie paléontologique.



GRAPHIQUE 1

Deux façons de décomposer le même arbre en groupes disjoints. La dimension verticale peut représenter le déroulement chronologique; les objets ou espèces classés correspondent aux sommets terminaux. Les groupes disjoints sont encerclés.

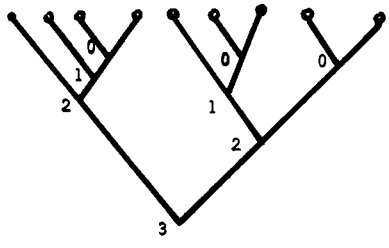
C'est dans le domaine de la systématique biologique plus généralement que l'on trouve le prototype de ce problème. Un arbre (une hiérarchie) donné (exemple: graphiques 1 à 3) représente les rapports entre des objets (espèces), eux-mêmes représentés par les sommets terminaux de l'arbre. Ces rapports sont présentés parfois en termes d'une hypothèse évolutioniste sur la parenté des espèces, qu'elle soit ou non associée à une échelle temporelle, parfois en termes des résultats d'un algorithme de regroupement (le *clustering*, *phonétique* ou *cladistique*, Sneath et Sokal 1973). Le problème de taxonomie qui se pose d'une façon quelque peu simplifiée au graphique 1, est de répartir les objets en un nombre de groupes disjoints (et ensuite de nommer ces groupes) en se servant de l'information inhérente à la structure de l'arbre.

II. LA CONDITION FORTE DE REGROUPEMENT

Une solution assez évidente et souvent empruntée serait de trancher la structure de l'arbre à un niveau critique, ce qui reviendrait soit à négliger les relations historiques trop éloignées, soit à terminer l'algorithme de regroupement hiérarchique à une valeur maximum de dissimilitude entre objets. L'arbre se décompose alors en un nombre de sous-arbres. Les objets ou sommets terminaux de chaque sous-arbre constituent donc un des groupes de la classification recherchée. Dans une variante de cette approche, le niveau critique est choisi à priori, en fonction de la dissimilitude maximum tolérable entre deux objets à l'intérieur d'un même groupe. Dans une autre variante le niveau est ajusté en fonction du nombre k de groupes voulus. Si ces échelles ne sont pas disponibles ou si on préfère ne pas s'en servir, par exemple lors d'une étude cladistique, il faut recourir à une autre méthode qui ne se base sur aucune échelle de temps ou de dissimilitude. On associe à chaque sommet non terminal S de l'arbre un chiffre qui indique le nombre maximum de sommets qui apparaissent entre S et un sommet terminal dominé (ou sous-tendu) par S (voir le graphique 2). Ces chiffres constituent en soi une échelle; nous pouvons alors trancher l'arbre à un niveau approprié.

GRAPHIQUE 2 :

Une échelle définie par le nombre maximum de sommets entre un sommet non-terminal et un sommet terminal.



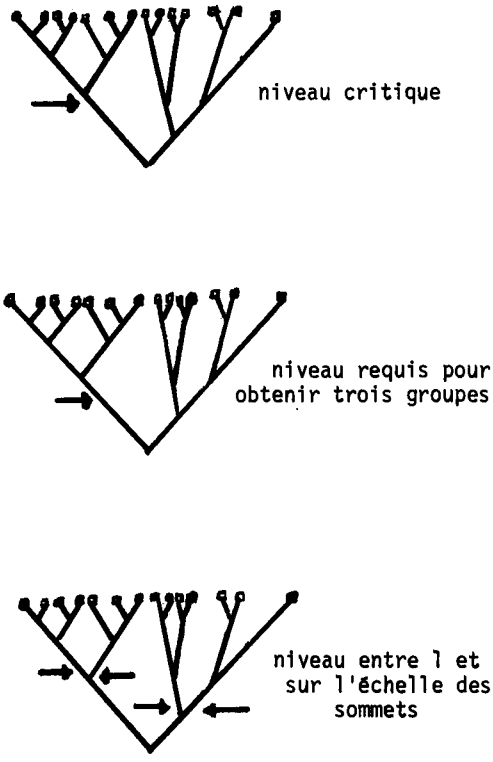
En somme, il existe au moins trois façons différentes de découper un arbre, basées sur

(i) un niveau critique sur une échelle donnée,

(ii) un critère de k groupes, ou

(iii) un niveau critique sur une échelle de sommets.

Les résultats de ces trois approches peuvent différer même lorsqu'ils sont appliqués à un même arbre, comme le démontre le graphique 3. Cependant elles ont toutes une propriété en commun, à savoir que les partitions des sommets terminaux en groupes disjoints qu'elles produisent satisfont une condition fondamentale que nous appelons la condition forte de regroupement. Une partition des sommets terminaux satisfait cette condition par rapport à l'arbre donné si *chaque sommet non-terminal ne domine que des groupes entiers, ou bien s'il ne domine qu'une partie d'un seul groupe*. Cette condition exclut qu'un sommet domine quelques membres de plusieurs groupes à moins qu'il ne domine entièrement chacun de ces groupes.



GRAPHIQUE 3

Résultats des trois méthodes appliquées au même arbre

Il y a un désavantage inhérent à toutes ces méthodes. Les groupes qui constituent une partition sont très souvent de tailles très différentes. Il y a généralement des groupes d'un ou deux membres seulement aussi bien que des groupes qui contiennent un quart, un tiers ou même une demie de tous les objets à classifier. Cependant, il est souhaitable qu'une classification consiste en groupes de tailles plus ou moins égales. Comme nous allons le voir, que cette propriété de tailles comparables ne soit pas respectée découle inévitablement de la condition forte.

III UN SEUIL POUR LA TAILLE DES GROUPES

Malgré le fait que la condition forte semble correspondre à une restriction naturelle sur les regroupements, comme en témoignent les trois méthodes de partition que nous avons examinées, la biologie systématique abonde en exemples qui ne respectent pas ce principe. Dans des études paléontologiques en particulier, on rencontre des schémas évolutionnistes d'un type assez différent de ceux dont nous venons de traiter. Le graphique 4 présente un exemple hypothétique d'un arbre d'évolution comme on en trouve souvent dans les oeuvres paléontologiques. Ce qui nous y intéresse n'est pas spécifiquement le fait que les objets se situent à différentes époques dans le temps, quoique ce fait soit relié au phénomène que nous allons étudier. L'important c'est que certains groupes contiennent des ancêtres d'autres groupes disjoints. On pourrait penser que ce phénomène n'apparaît pas dans les regroupements des graphiques 1 à 3 simplement parce que les objets qui y sont classés sont tous contemporains alors que dans le graphique 4, les objets sont dispersés dans le temps. Cependant, aux graphiques 1 à 3 il y a aussi des objets non contemporains, c'est-à-dire les objets hypothétiques ou ancestraux, correspondant aux sommets non terminaux. Ceux-ci peuvent être classés, pour la plupart, selon la classification construite pour les sommets terminaux, et ce d'une seule façon naturelle. Un sommet non terminal appartient à un groupe s'il domine au moins un sommet terminal du groupe et ne domine aucun membre des autres groupes. De cette façon la condition forte n'est jamais violée. Ainsi il nous reste encore à expliquer en quoi diffèrent les principes d'organisation de la classification au graphique 4 de ceux des autres graphiques.

GRAPHIQUE 4

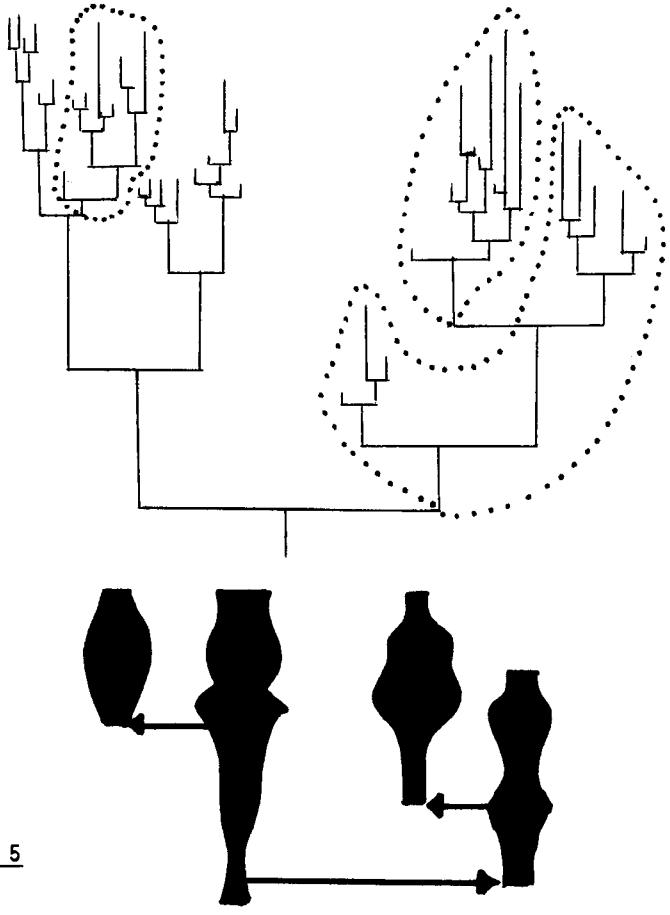
Arbre d'évolution tel qu'on en trouve dans les ouvrages de paléontologie. Les parties foncées représentent des ensembles nommés contenant plusieurs lignées.

Existe-t-il une autre condition de regroupement qui permettrait des classifications du type de celle du graphique 4? Comment définir un algorithme qui trouverait de telles classifications? Les recherches par Raup et al. (1973) nous aident à répondre à ces questions.

Ces auteurs ont simulé par ordinateur le processus de l'évolution. Le modèle qui engendrait les arbres évolutionistes ressemblait à un processus de ramification ou d'embranchement (Karlin et Taylor 1975, ch.8) du type Galton-Watson. Après chaque intervalle de temps, le destin de chaque espèce (ou lignée) était déterminé par hasard; ou bien l'espèce demeurait inchangée pour un autre intervalle de temps, ou bien elle se divisait en deux espèces qui, à partir de ce moment, évoluaient indépendemment, ou bien encore elle s'éteignait complètement. Les probabilités de division et d'extinction étaient beaucoup plus petites que la probabilité de continuation, et elles étaient revisées de temps en temps afin que le nombre de lignées ne s'éloigne pas trop d'une valeur fixée au préalable. Le programme d'ordinateur produisait des schémas évolutionistes tels qu'au graphique 5. On considérait comme lignée toute ligne non interrompue entre deux époques dans le temps, correspondant à l'évolution d'une espèce à partir de l'un de ses ancêtres.

Une fois le schéma engendré, ses lignées étaient regroupées en fonction du principe suivant: *pour qu'un ensemble de lignées soit regroupé, il faut que ces lignées aient toutes pour origine directe un ancêtre commun. De plus, la longueur totale des lignées dans le groupe doit excéder un seuil donné, sans qu'aucun des*

sous-groupes qu'il contient n'atteigne ce même seuil. (Remarquons que plusieurs lignées peuvent être identiques pendant une période de temps et ensuite prendre chacune son propre chemin; en définissant la longueur totale des lignées dans un groupe, nous ne comptons qu'une fois de telles portions partagées par des lignées.) Le graphique 5 montre un regroupement compatible avec ce principe de seuil. On voit immédiatement que certains groupes contiennent des ancêtres des autres, ce qui constitue une violation de la condition forte de regroupement.



GRAPHIQUE 5

Arbre d'évolution d'après les simulations de Raup et al. La distance verticale (mais non horizontale) correspond au déroulement chronologique. En haut: lignées regroupées selon un seuil en quatre groupes dont trois sont encerclés par des lignes discontinues. Les lignes qui restent forment le quatrième groupe. En bas: Les aires foncées (voir le graphique 4) correspondent aux groupes de lignées. L'épaisseur correspond au nombre actuel de lignées dans le groupe. Les flèches indiquent la direction de l'évolution.

En fait, si nous essayons de modifier les trois méthodes de regroupement de la section 2, pour qu'elles satisfassent à la fois la condition forte et le principe du seuil, nous nous rendrions vite compte qu'elles ne se combineraient pas bien ensemble, puisque le plus souvent elles produiraient des partitions peu intéressantes, consistituées d'un seul groupe, ou alors elles laisseraient beaucoup d'espèces non classées. Donc pour obtenir des regroupements intéressants, formés de plusieurs groupes contenant un minimum de membres, il faut affaiblir la condition forte. Comment le faire sans complètement négliger la structure de l'arbre?

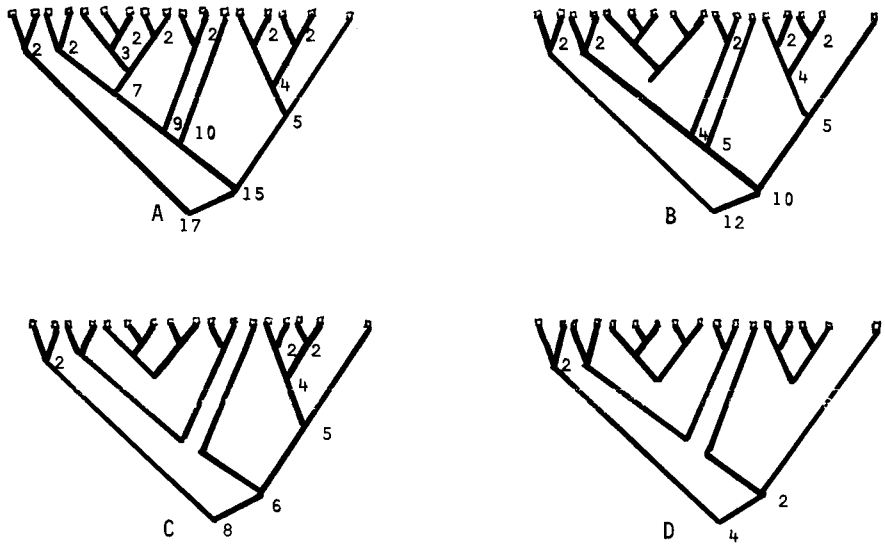
IV LA CONDITION FAIBLE ET UN ALGORITHME

La condition forte permettait qu'un sommet non terminal domine une partie d'un groupe ou bien un ou plusieurs groupes entiers. La condition faible que nous proposons ici remplace le "ou bien" par "et/ou". Donc, il n'y a qu'une restriction à savoir qu'un sommet ne peut dominer seulement une partie de deux (ou plusieurs) groupes. En examinant le graphique 5, on voit tout de suite que là où la condition forte est violée, la condition faible ne l'est pas.

Ainsi nous avons une condition de regroupement qui est compatible avec un seuil pour la taille des groupes. Définissons maintenant un algorithme pour trouver de telles partitions à partir d'un arbre donné. Il y a plusieurs possibilités, surtout parce que le seuil et la condition faible, même imposés simultanément, ne déterminent pas une partition unique. La méthode que nous décrirons diffère quelque peu de celle de Raup et al., mais nous l'adoptons parce qu'elle est comparable avec les méthodes (i)-(iii) de la section 2, en ce qu'elle est directement applicable aux arbres représentés aux graphiques 1 à 3, ainsi qu'à l'arbre du graphique 5 moyennant une légère modification de sa définition. Néammoins les conclusions que nous en tirerons nous semblent pertinentes à toute partition basée sur un seuil.

Reformulons la notion de seuil pour qu'elle s'applique rigoureusement à un arbre tel que ceux qui figurent dans les graphiques 1 à 3.

Une partition des sommets terminaux et non terminaux satisfait le principe de seuil si chaque groupe contient un nombre de sommets terminaux plus grand que ou égal au seuil, et si ces sommets sont tous reliés à un ancêtre commun uniquement par l'intermédiaire des sommets non terminaux appartenant au même groupe. De plus, le groupe ne peut contenir un sous-groupe qui satisfasse ces mêmes conditions.



GRAPHIQUE 6

A. Arbre où chaque sommet non terminal est indexé selon le nombre de sommets terminaux qu'il domine.

B. Après cinq cycles de l'algorithme, un groupe est isolé.

C. Un deuxième groupe est isolé.

D. Un troisième groupe est formé, et l'algorithme se termine parce que l'indice de la racine ne peut plus baisser sans violer le seuil. Les sommets qui restent constituent un dernier groupe.

Avant d'appliquer l'algorithme, il faut établir un seuil. Nous illustrons l'application de l'algorithme au graphique 6 avec un seuil de quatre sommets terminaux par groupe. On établit une convention selon laquelle le sommet non-terminal R à la racine de l'arbre est dans le groupe G_1 et on initialise $i = 1$.

(1) La première étape consiste à indexer chaque sommet non-terminal selon le nombre de sommets terminaux qu'il domine.

(2) Dans la deuxième étape, on choisit n'importe lequel des sommets terminaux T qui ne sont pas encore classés. (Lorsqu'il ne reste aucun sommet non classé, l'algorithme est terminé.) On examine les ancêtres de T un par un à partir de son ancêtre immédiat. On s'arrête aussitôt qu'on rencontre ou un ancêtre A déjà classé dans un groupe, ou un ancêtre B dont l'indice est plus grand

que ou égal au seuil. Dans le premier cas, on passe à la troisième étape, dans le deuxième cas à la quatrième étape.

(3) Lorsqu'on rencontre un ancêtre A déjà classé dans le groupe G_j où $j \leq i$, on assigne T ainsi que tous les sommets non terminaux entre T et A à ce groupe. On retourne à l'étape 2.

(4) Si aucun des descendants immédiats non-classés de B n'a un indice plus grand que ou égal au seuil, nous allons directement à l'étape 5. Si par contre un des descendants immédiats non-classé de B a un indice plus grand que ou égal au seuil, nous ré-étiquettons ce descendant "B" et nous recommençons la quatrième étape.

(5) Nous augmentons i à i+1 et nous créons un nouveau groupe G_i où B est le membre initiateur. On ré-indexe tous les ancêtres de B, en soustrayant des anciennes valeurs l'indice de B, sauf si cette opération entraîne une réduction de l'indice de R en deçà du seuil, auquel cas on annule le nouveau groupe, on remet i à i-1, et on regroupe B ainsi que tous ses ancêtres, dans G_{i-1}. On retourne à l'étape 2.

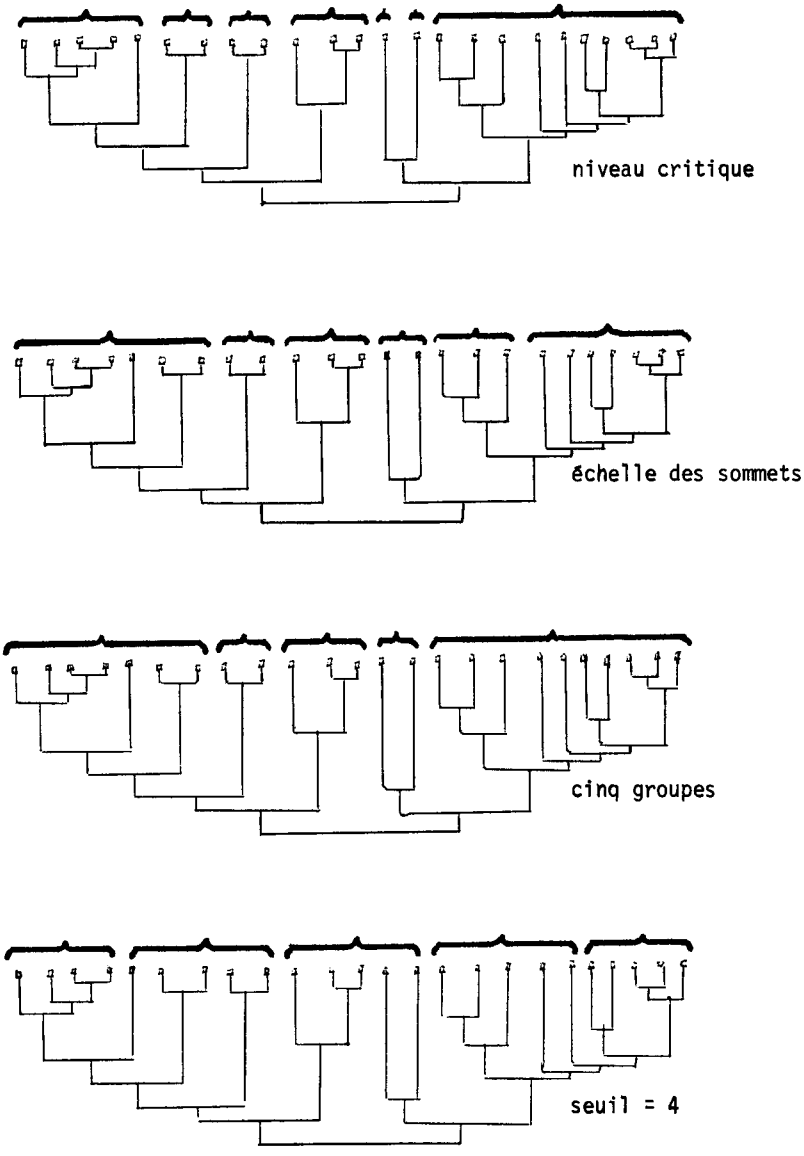
Cet algorithme respecte la condition faible et il ne crée que des partitions où chaque groupe contient au moins le nombre de membres spécifiés par le seuil. Pour l'appliquer à un arbre comme au graphique 5, au lieu d'indexer chaque sommet non terminal par le nombre de sommets terminaux qu'il domine, on l'indexe selon la longueur totale des lignées qu'il donne.

Le graphique 7, à la page suivante, montre comment les résultats de cet algorithme peuvent différer de ceux obtenus par d'autres méthodes.

V CONSIDERATIONS DE STABILITE

Au nombre des propriétés désirables des critères et méthodes de regroupement figure celle de stabilité par rapport aux données. Si on change légèrement l'arbre auquel on applique la méthode, on préfère que la partition qui en résulte ne change pas terriblement. Afin d'évaluer notre nouvelle méthode par rapport aux autres, nous avons entrepris des essais de simulation.

A partir d'un processus Yule d'embranchement (Karlin et Taylor 1975,ch.4), nous avons engendré 50 arbres de 24 sommets terminaux chacun, sous la condition que le même intervalle de temps s'écoule pour chaque arbre entre le premier embranchement (la racine) et la 24ième occurence d'embranchement. Les 24 sommets terminaux sont attachés aux 24 lignées existantes à ce dernier stade. Les quatre méthodes de



GRAPHIQUE 7

Résultats des quatre méthodes sur un arbre simulé

regroupement ont été créés à chacun des 50 arbres ainsi construits. Ensuite, 50 nouveaux arbres ont été créés en adjoignant à chacun des arbres originaux un nouveau sommet terminal, attaché à un nouveau sommet non terminal, ce dernier situé aléatoirement sur des branches de l'arbre original. Les quatre méthodes de regroupement ont été appliquées de nouveau avec arbres ainsi modifiés. Pour évaluer la stabilité de chaque méthode, il s'agissait donc de comparer les regroupements qu'elle produirait sur chaque paire d'arbres, avant et après la modification.

Pour comparer le regroupement des 24 sommets originaux et le regroupement des 25 sommets obtenus après la modification, nous avons considéré le nombre de différences entre les matrices d'affinité des 24 sommets avant et après. L'entrée (i,j) d'une telle matrice est 1 si $i < j$ et si le $i^{\text{ième}}$ et $j^{\text{ième}}$ sommets sont dans le même groupe, et 0 sinon.

Un autre critère consiste simplement à voir si la matrice reste inchangée ou non.

Avant de présenter les résultats de cette expérience, il faut mentionner quelques détails techniques. La méthode basée sur le seuil risque de paraître instable simplement à cause de la non-unicité de la partition qui satisfait le critère non-unicité qui se réalise dans l'algorithme aux étapes (2) et (4) où l'on choisit n'importe lequel des sommets terminaux non classés, et où on choisit d'examiner un des descendants non classés de B. Pour éviter cela, nous avons numéroté les sommets et chaque fois qu'il y avait un choix à faire, l'algorithme le faisait en ordre numérique. La même numérotation servait avant et après la modification.

Afin d'assurer un maximum de comparabilité des quatre méthodes, nous avons ajusté leurs paramètres (niveaux de regroupements, nombre de groupes, seuil) pour qu'elles donnent toutes à peu près cinq groupes par partition.

	Nombre de groupes	Pourcentage de regroupements changés	Nombre moyen de changements† à la matrice d'affinité
Niveau critique	5.48	0*	0*
Echelle des sommets	4.44	10	1.44
Cinq groupes	5.00*	16	2.80
Seuil = 4	5.30	40	7.12

* par définition

† maximum $\binom{24}{2} = 276$

TABLEAU 1

Stabilité des quatre méthodes.

Les résultats sont présentés au Tableau 1. Remarquons que le nombre moyen de groupes est plus ou moins cinq pour chaque méthode. Il est évident que la méthode basée sur le seuil est le moins stable des quatre, selon les deux critères.

VI CONCLUSIONS

Résumons les propriétés de la méthode que nous avons proposée.

(i) Les groupes tendent à être de taille uniforme.

(ii) Les regroupements ne respectent pas nécessairement la condition forte. Ainsi ils peuvent permettre les rapports entre groupes qu'on retrouve dans les études de paléontologie. En respectant la condition faible, ils conservent des éléments de la structure de l'arbre.

(iii) Les regroupements sont relativement instables.

Nous pouvons donc conclure que l'imposition d'un seuil, soit explicitement, soit implicitement (comme Raup et al. le perçoivent dans la tradition de la nomenclature en paléontologie), mène à une certaine instabilité. Les classifications ainsi constantes risquent d'être reformulées sensiblement chaque fois qu'on découvre de nouvelles données et qu'on essaie de les incorporer. Voilà peut-être ce qui explique partiellement pourquoi il y a tant de controverse en paléontologie!

REFERENCES

KARLIN,S. et TAYLOR,H.M., *A first course in stochastic processes*. Second edition. New York, Academic Press, 1975.

RAUP,D.M., GOULD,S.J., SCHOPF,T.J.M. et SIMBERLOFF,D.S., Stochastic models of phylogeny and the evolution of diversity. Journal of Geology 81, (1973), 525-542.

SNEATH,P.H.A. et SOKAL,R.R., *Numerical taxonomy*, W.H.Freeman, San Francisco, 1973.

Centre de Recherches Mathématiques
Université de Montréal
Montréal, Québec, Canada H3C 3J7