Matrices AVL pour la classification et l'alignement de séquences protéïques

Israël-César Lerman, Philippe Peter, J.L. Risler

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

MATRICES AVL POUR LA CLASSIFICATION ET L'ALIGNEMENT DE SEQUENCES PROTEIQUES.

I.C. Lerman , Ph. Peter , J.L. Risler

N° 2466

septembre 1994

PROGRAMME 3

ISSN 0249-6399

MATRICES AVL POUR LA CLASSIFICATION ET L'ALIGNEMENT DE SEQUENCES PROTEIQUES.

I.C. Lerman , Ph. Peter*, J.L. Risler**

Programme 3 — Intelligence artificielle, systèmes cognitifs et interaction homme-machine
Projet Repco

Rapport de recherche n° 2466 — septembre 1994 — 41 pages

Résumé : La matrice de Dayhoff de comparaison par paires entre acides aminés, a joué un rôle crucial pour alimenter la recherche bioinformatique. Il peut s'agir d'aligner une famille de séquences protéïques, via des algorithmes de type "programmation dynamique". Il peut également s'agir de l'information "Similarité" pour la classification de telles séquences, après ou sans un alignement préalable. Si une nouvelle matrice dite BLOSUM 62 est maintenant également considérée par les biologistes, sa nature statistique est la même que celle de Dayhoff; dont le principe d'obtention est plus ambitieux, mais moins robuste. Un premier objectif de ce travail est d'expliciter en les comparant les conceptions de chacune des deux matrices. Nous proposons aussi deux nouvelles matrices qui sont respectivement associées; mais qui sont conformes à la méthodologie de l'*A*nalyse de la *V*raisemblance des *L*iens (AVL). Les quatre matrices sont comparées à travers la classification AVL, sur un ensemble aligné de 89 séquences protéïques de cytochrome C. Nous donnons au codage de l'information Similarité deux formes : "graphe valué" et "préordonnance". Si l'alignement mentionné est réalisé à partir de considérations structurelles, les nouvelles matrices permettent un alignement multiple via des algorithmes de type programmation dynamique et aggrégation selon le plus proche voisin.

Mots-clé : Similarité, Classification, Séquences génétiques

(Abstract: pto)

*. IRESTE, La Chantrerie, CP 3003, 44087 Nantes cedex.
**. Centre de Génétique Moléculaire, 91198 Gif-sur-Yvette cedex.

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : (33) 99 84 71 00 – Télécopie : (33) 99 84 71 71

LLA MATRICES FOR CLASSIFICATION AND ALIGNMENT OF PROTEIN SEQUENCES.

Abstract: The Dayhoff matrix of amino acid pairwise comparison has played a crucial part in computing biology research. Two main and related problems are concerned : Alignment and Classification. A new matrix, called BLOSUM 62, is now equally considered by the Biologists. However, its statistical nature is exactly the same as that of Dayhoff matrix. Nevertheless, the respective elaborations of both matrices, are fundamentally distinct. Our first aim in this paper, is to make clear the formal comparison between the respective conceptions of the two matrices. An important objective consists also of proposing two new matrices associated with the previous ones, according Likelihood Linkage Analysis (LLA) methodology. The four matrices are evaluated through LLA classification, on a set of aligned genetic sequences. The training set is provided by 89 protein sequences, aligned according structural considerations. Two representations are considered for the Similarity coding :"weighted graph" and "preordonnance". Both new matrices enable new results in Classification and multiple Alignment, through dynamic programming and clustering algorithms.

Key-words: Similarity, Classification, Genetic sequences

Table des matières

1 INTRODUCTION

Dans le domaine de la biologie moléculaire, la classification de séquences protéïques conformément à leurs "proximités" respectives, est crucial pour induire de la connaissance [7].

Formellement, les éléments qui sont à organiser selon un schéma classificatoire sont des suites de lettres provenant d'un alphabet fini A. Ces suites sont de longueurs respectives différentes ; d'autre part, l'alphabet A comprend 20 lettres, représentant chacune un acide aminé.

Les différences respectives de longueur sont justifiées, dans la théorie de l'évolution, par un phénomène de délétions ou d'insertions d'acides aminés dans une même séquence protéïque, au cours du temps. En introduisant un symbole spécial noté - et représentant un "saut", les séquences protéïques d'une même famille (e.g. le cytochrome C) peuvent être potentiellement, mutuellement alignées (alignement multiple) ; l'alphabet utilisé $(A \cup \{-\})$ comprenant ainsi, 21 lettres. Précisèment, un des objectifs majeurs de la recherche en biologie moléculaire consiste à valider des alignements multiples.

Un alignement multiple peut être obtenu de différentes façons. Pour chacune ; et si on se limite à la comparaison d'une seule paire de séquences, ce qui intervient de façon fondamentale c'est une *comparaison d'un terme résidu de la première séquence à un terme résidu de la seconde.* Dans ces conditions, deux problèmes essentiels et liés se présentent : comment appareiller les deux termes et quelle notion de similarité utiliser pour la comparaison ?

A ce dernier égard, différentes notions se référant à différents points de vue peuvent être considérées [3]. Signalons tout de suite que notre expérience de classification rapportée ici porte sur un ensemble de 89 séquences protéïques du cytochrome C, préalignées à partir de considérations liées à la structure. Conformément à ce point de vue, on établit bien dans [16] une matrice d'association entre acides aminés ; mais, que nous n'utiliserons pas ici. De façon précise l'alignement multiple a été réalisé suivant la méthode de Sander [C. Sander and R. Schneider 1991] et extrait de la banque de données HSSP distribuée par l'EMBL. De la sorte, l'évaluation de nos matrices de comparaison entre acides aminés, via la classification hiérarchique AVL (Analyse de la Vraisemblance des Liens) [10], [14], sera parfaitement indépendante de la méthode de construction de l'alignement multiple.

En effet, la donnée d'une matrice de similarités numériques entre acides aminés permet, via la programmation dynamique et un algorithme de groupement par proximité [cf. références 6 à 12 de [7]], de réaliser - à partir d'un principe de parcimonie - un alignement multiple. C'est ainsi que la célèbre matrice de Dayhoff a particulièrement nourri les algorithmes d'alignement.

Précisèment, la première matrice que nous proposerons sera déduite à partir de celle, conforme à la méthode de construction de Dayhoff, correspondante à 250 PAM et obtenue par Jones, Taylor et Thornton [6]. Nous la désignerons par D. D'autre part, une nouvelle famille de matrices, dite BLOSUM et dûe à [5], est de plus en plus considérée par les biologistes. Si l'élaboration d'une matrice BLOSUM diffère sensiblement de celle d'une matrice Dayhoff, sa nature statistique est la même. Un même élément représente en effet dans chacun des cas le logarithme d'une densité de probabilité. Dans cette deuxième famille, la matrice standard est BLOSUM 62 ; qui donc, conduira à la deuxième matrice que nous proposerons. C'est au paragraphe II que nous y étudierons les conceptions respectives de la matrice de Dayhoff et de celle des Henikoffs. Puis, après une analyse comparative, nous montrerons comment nous déduisons les matrices, respectivement assoicées, d'indices probabilistes que nous pouvons appeler AVL (D) et AVL (B) ; ou celles, de "dissimilarité informationnelle", directement déduites.

Après avoir, à partir d'une même matrice, inféré les valeurs des similarités entre les 20 acides aminés et une déletion, ainsi qu'entre deux déletions, il nous sera possible de construire un indice de similarité entre séquences protéïques mutuellement alignées, conforme à la méthode AVL.

A cette fin, deux approches sont envisagées. La première est numérique et consiste à considérer que la variable associée à un site est un graphe valué complet sur l'ensemble A' des 21 lettres (représentant les 20 acides aminés et la déletion). La seconde approche est ordinale et consiste à ne retenir de la valuation numérique que le préordre total qu'elle induit sur l'ensemble des couples de lettres ; il s'agit d'une "préordonnance" sur A'. C'est au paragraphe III que nous décrirons cette construction.

Etant donné une famille de séquences protéïques, déjà préalignée et "au mieux", un premier objectif majeur de la recherche est de trouver le codage de l'information similarité - pour une comparaison site par site - qui conduise par la méthode AVL, aux résultats les plus cohérents, compte tenu de la connaissance acquise en phylogénie. En effet, un bon niveau de cohérence permettra à l'outil de constituer une aide à l'inférence dans des situations moins familières.

D'autre part et de façon liée, on peut comparer différents alignements de par la qualité des classifications associées.

Le point de départ de la construction d'une matrice de type Dayhoff ou BLOSUM ; et donc également AVL est un ensemble de séquences déjà alignées, ou bien, une famille de tels ensembles. De tels alignements sont effectués

à partir d'une connaissance a priori, mixée avec des aspects algorithmiques de comparaisons simples. Mais, ces mêmes matrices peuvent servir - via la programmation dynamique - à aligner un ensemble de séquences ; qu'il s'agisse de l'un de ceux dont on est parti ou d'un nouveau. Et, à cet égard, on a pu reprocher une certaine circularité du raisonnement. Nous discuterons ce point en conclusion ; surtout, en y faisant intervenir l'apport et l'éclairage que fournit l'outil classification (paragraphe V).

Le paragraphe V quant à lui, est consacré à l'analyse des différents traitements effectués et à la comparaison des résultats. Ce qui nous permettra de fixer une stratégie AVL pour la classification d'un ensemble de séquences protéïques alignées, avec une comparaison site par site.

Ce n'est pas ce seul type de comparaison qui est envisagé dans [12], qui constitue une première expérience, ayant motivé la présente étude.

C'est après les réfèrences bibliographiques que nous porterons les figures, tables et annexes.

2 Matrices d'Association entre Acides Aminés

2.1 Conception de la matrice de Dayhoff

Le mathématicien abordant le sujet, mais aussi, le biologiste, peut éprouver une difficulté certaine à dégager - à partir des publications - de façon claire et précise, le principe, pourtant simple, des calculs conduisant à la matrice de Dayhoff. Dans ces conditions, nous allons proposer des *formules* qui ne permettront plus la moindre ambiguïté.

La donnée de base est une famille de séquences protéïques d'un même type (e.g. le cytochrome C) et, plus ou moins contemporaines quand à leur évolution. On suppose ; et c'est une condition nécessaire pour le calcul, que les séquences sont alignées deux à deux (par paires). Cependant, il n'est pas nécessaire que la longueur - supposée commune - d'une paire alignée, soit la même pour les différentes paires. D'autre part, plusieurs familles de séquences pourront contribuer au calcul des indices statistiques qui seront considérés ; puisque dans ces derniers, l'opération fondamentale est une sommation sur un ensemble de paires de séquences. Nous y reviendrons.

Pour simplifier, se fixer les idées et rendre plus claire la nature des calculs, nous supposons qu'on dispose d'une famille

$$S = \{s_k / 1 \leq k \leq K\} \tag{1}$$

de K séquences toutes de même longueur L et préalignée (alignement multiple). Plus précisèment, nous considérons ici que chaque séquence est un mot comprenant L lettres prises dans un alphabet

$$A = \{Z_i / 1 \leq i \leq 20\} \tag{2}$$

représentant les 20 acides aminés. C'est sans difficulté majeure que nous pourrons déduire le cas le plus général ci-dessus mentionné et où, de plus, pour l'alignement d'une même paire de séquences, on peut être conduit à introduire un 21ème symbole dans l'alphabet, noté -, et représentant une déletion.

Ainsi, dans les conditions où nous nous sommes mises, une même séquence protéïque s_k peut être représentée par une application que nous noterons également s_k de l'ensemble.

$$L = \{1, 2, ..., l, ..., L\} \tag{3}$$

des indices ou étiquettes dans A :

$$\begin{array}{rccl} s_k & : & L & \longrightarrow A \\ & & l & \longrightarrow s_k(l), \end{array} \tag{4}$$

où $s_k(l)$ est l'acide aminé occupant le site l de la k-ème séquence, $1 \leq k \leq K$.

Maintenant, le support du calcul est l'ensemble $P_2(S)$ des parties à deux éléments de S; c'est à dire, des paires de séquences :

$$P_2(S) = \{\{s_k, s'_k\} / 1 \leq k < k' \leq K\} \tag{5}$$

Pour s appartenant à S, on peut associer la distribution de fréquences absolues de A sur S :

$$\{n_s(i) = card[(s^{-1}(Z_i)]/1 \leq i \leq 20\} \tag{6}$$

$n_s(i)$ est le nombre de fois où la lettre Z_i apparait dans la séquence s. On a bien sûr

$$\sum_{1 \leq i \leq 20} n_s(i) = L \tag{7}$$

Relativement à l'ensemble S des K séquences, introduisons les fréquences relatives

$$f_i = \frac{N(i)}{KL}, 1 \leq i \leq 20;$$

où

$$N(i) = \sum_{s \epsilon f} n_s(i) \tag{8}$$

est le nombre total de fois où l'acide aminé Z_i (la lettre Z_i) apparaît dans l'ensemble des séquences qui regroupent bien KL, lettres.

On peut établir que les proportions f_i sont les mêmes si on prenait comme support du calcul l'ensemble $P_2(S)$ des paires de séquences (5) ci-dessus. On a bien en effet :

$$f_i = \frac{\sum\{[n_s(i) + n_{s'}(i)]/\{s, s'\} \epsilon P_2(S)\}}{[K(K-1)L]} \tag{9}$$

$1 \leq i \leq 20$; le dénominateur étant le nombre total de lettres qu'on trouve dans $P_2(S)$.

Ce qui intervient de façon fondamentale dans l'élaboration de la matrice de Dayhoff c'est en premier lieu la notion quantifiée de *mutabilité relative d'un acide aminé à travers un ensemble S de séquences deux à deux alignées.*

Commençons par introduire, en nous aidant d'un exemple, la notion de mutabilité relative *non normalisée* d'un acide aminé Z_i, correspondante à une paire $\{s, s'\}$ de séquences alignées ; et que nous noterons $\mu_i(\{s, s'\})$. Ce nombre est le rapport entre le nombre de fois (de sites) où Z_i est présent dans l'une seulement des deux séquences ; et le nombre de fois où Z_i est présent dans l'une ou l'autre des deux séquences s ou s'. Ainsi, si on considère les deux séquences

$$ACDEFL$$

$$AGDEAL,$$

alors la mutabilité relative non normalisée de A est 1/3, celle de G (ou d'ailleurs de C) est 1, celle de D est 0,... Il s'agit de toute façon d'un nombre compris entre 0 et 1.

Plus généralement, désignons par $d_i(\{s, s'\})$ et par $e_i(\{s, s'\})$ (d comme *d*oublé et e comme *e*sseulé), respectivement, le nombre de sites où face à Z_i, il y a le même acide aminé Z_i ; et le nombre de sites, où face à Z_i, il y a un autre acide aminé que Z_i. On a alors

$$\mu_i(\{s, s'\}) = \frac{e_i(\{s, s'\})}{2 \times d_i(\{s, s'\}) + e_i(\{s, s'\})} \tag{10}$$

$1 \leq i \leq 20$.

On se rend compte que $\mu_i(\{s, s'\})$ est un indice par trop brut de la mutabilité de l'acide aminé Z_i ; en effet, le taux $\mu_i(\{s, s'\})$ n'a pas la même signification selon qu'il y a un grand pourcentage de mutations entre s et s', ou un petit. Dans ces conditions, introduisons, relativement à une même paire $\{s, s'\}$ de séquences alignées, le pourcentage $p(\{s, s'\})$ de mutations ; de façon précise

$$p(\{s, s'\}) = \frac{L_m(\{s, s'\})}{L} \tag{11}$$

où L est la longueur commune des deux séquences et où $L_m(\{s, s'\})$ est le nombre de sites où il y a mutation ; c'est à dire, formellement, où on ne trouve pas la même lettre, à la fois dans s et dans s'. La *mutabilité normalisée* ou *mutabilité tout court*, de l'acide aminé Z_i à travers la paire $\{s, s'\}$ de séquences se définit par :

$$m_i(\{s,s'\}) = \frac{\mu_i(\{s,s'\})}{p(\{s,s'\})} = \frac{e_i(\{s,s'\})}{[2d_i(\{s,s'\}) + e_i(\{s,s'\})]p(\{s,s'\})}, \quad (12)$$

$1 \leq i \leq 20$.

Maintenant, pour quantifier la mutabilité de l'acide aminé Z_i à travers une famille S de séquences ; il y a lieu de totaliser les effets de l'ensemble $P_2(S)$ des paires de séquences. Ce qui est proposé dans la littérature [2], [6] consiste à condidérer le rapport entre deux totalisations sur $P_2(S)$; le première concerne le numérateur de $m_i(\{s,s'\})$ et la seconde, le dénominateur de $m_i(\{s,s'\})$ (12). Très précisément, on a :

$$m_i = (\sum \{e_i(\{s,s'\})/\{s,s'\} \in P_2(S)\}/ \sum \{[2d_i(\{s,s'\}) + e_i(\{s,s'\})]p(s,s') \{s,s'\} \in P_2(S)\}) \quad (13)$$

Nous voulons mentionner ici une autre façon qui nous parait plus précise de définir la mutabilité relative d'un acide aminé Z_i à travers une famille S de séquences. Il s'agit de la moyenne sur l'ensemble des paires de séquences où l'acide aminé Z_i apparait des mutabilités relatives $m_i(\{s,s'\})$ (12), respectivement associées.

Plus précisément en désignant par S_i l'ensemble des séquences où l'acide aminé Z_i est présent et par T_i l'ensemble complémentaire ; on a :

$$S = S_i + T_i(\text{somme ensembliste}) \quad (14)$$

L'ensemble des paires de séquences où l'acide aminé Z_i intervient peut se mettre sous la forme suivante :

$$P_2(i) = P_2(S_i) + S_i * T_i(\text{somme ensembliste}) \quad (15)$$

où $P_2(S_i)$ $(resp. S_i * T_i)$ est l'ensemble des paires de séquences dont les deux composants appartiennent à S_i (resp. dont l'une des deux composantes appartient à S_i et l'autre, à T_i). En notant K_i la taille de S_i, on a :

$$card[P_2(i)] = \frac{1}{2}K_i(K_i - 1) + K_i(K - K_i) = \frac{1}{2}K_i(2K - K_i - 1). \quad (16)$$

En désignant par m'_i cette interprétation nouvelle de la mutabilité relative globale, on a :

$$m'_i = \frac{2}{K_i(2K - K_i - 1)} \sum \{m_i(\{s,s'\})/\{s,s'\} \in P_2(i)\} \quad (17)$$

((12) ci-dessus)

On se rend compte que le calcul de $m_i(\{s,s'\})$ sur lequel il y a lieu de se focaliser et qui permet la détermination de m_i ou de m'_i, nécessite seulement que - relativement à la famille S - les séquences soient deux à deux alignées. La longueur commune d'une paire de séquences alignées n'est pas un paramètre qui intervient. D'autre part, si des déletions doivent intervenir dans l'alignement d'une même paire de séquences, on peut envisager l'une des solutions suivantes :

(*i*) comme nous l'avons déjà mentionné, introduire un 21ème symbole dans l'alphabet A (2), noté -, et considéré au même titre que chacun des 20 acides aminés ;

(*ii*) ignorer complètement les sites de la paire $\{s,s'\}$ de séquences alignées où une déletion intervient dans s ou (non exclusif) dans s' ;

(*iii*) ignorer les sites où une déletion apparait pour le calcul des paramètres $e_i(\{s,s'\})$ et $d_i(\{s,s'\})$ (12) ; mais, retenir ces sites de la même façon que ci-dessus cf. (i), pour le simple calcul du pourcentage $p(\{s,s'\})$.

d'autre part, on voit que rien n'empêche à ce que plusieurs familles de séquences contribuent à la détermination de $m_i(resp.m'_i)$. Si $\{S^t/1 \leq t \leq u\}$ désigne un ensemble de familles, chacune homogène, on remplacera dans (13), $P_2(S)$ par

$$\sum \{P_2(S^t)/1 \leq t \leq u\} \quad (18)$$

(somme ensembliste). Alors que dans (17), on commencera à associer à chaque ensemble S^t, son ensemble $P_2^t(i)$, de la même façon que $P_2(i)$ a été associé à S. Dans ces conditions, m_i' sera défini par un rapport dont le numérateur a la forme suivante :

$$\sum_{1\leq t\leq u}\sum\{m_i(\{s,s'\})/\{s,s'\}\in P_2^t(i)\} \tag{19}$$

Le dénominateur est la somme des cardinaux des ensembles $P_2^t(i)$; c'est à dire,

$$\sum_{1\leq t\leq u} K_i^t(2K^t - K_i^t - 1)/2 \tag{20}$$

où K^t est le cardinal de S^t et où K_i^t est le cardinal du sous ensemble S_i^t de S^t, où Z_i est présent.

Reprenons la mutabilité relative globale m_i (15) ; on aurait pu -encore une fois- considérer de façon plus précise m_i' cf. (19), d'un acide aminé Z_i à travers une famille S de séquences, $1 \leq i \leq 20$.

Ici intervient une hypothèse essentielle modélisant la *probabilité de mutation* de i, pour "une période unitaire d'évolution". Nous notons $M_{\bar{i}i}$ cette probabilité de passage de i vers non i. On pose ; et *c'est crucial* pour la suite de la construction, que $M_{\bar{i}i}$ est proportionnelle à la mutabilité relative m_i :

$$M_{\bar{i}i} = \lambda m_i, \tag{21}$$

$1 \leq i \leq 20$

Dans ces conditions, la probabilité de persistance de i, s'écrit :

$$M_{ii} = 1 - M_{\bar{i}i} = 1 - \lambda m_i, \tag{22}$$

$1 \leq i \leq 20$

Maintenant, considérons

$$p_S(i) = f_i M_{ii}; \tag{23}$$

il s'agit de la probabilité de rencontre d'un acide aminé Z_i qui persiste dans le contexte de S. Plus précisèment, il s'agit de la *probabilité pour une paire $\{s,s'\}$ de séquences alignées, prise uniformément au hasard dans $P_2(S)$ (5) et pour un site choisi au hasard, de rencontrer l'acide aminé Z_i à la fois dans les séquences s et s', au site choisi.*

Dans ces conditions, la probabilité globale de la persistance à travers S est :

$$p_S = \sum_{1\leq i\leq 20} p_S(i) = \sum_{1\leq i\leq 20} f_i M_{ii}. \tag{24}$$

Par rapport à l'expression précédente, il s'agit de la probabilité de rencontrer un acide aminé qui ne mute pas entre s et s' ; quel que soit ce dernier.

On impose précisément à cette probabilité d'être égale à 0.99 :

$$p_S = \sum_{1\leq i\leq 20} f_i M_{ii} = 0.99; \tag{25}$$

ce qui correspond à une mutation pour 100 sites, pour une période unitaire d'évolution. La relation (25) qui s'écrit

$$\sum_{1\leq i\leq 20} f_i M_{\bar{i}i} = \lambda \sum_{1\leq i\leq 20} f_i m_i = 0.01, \tag{26}$$

permet de fixer la valeur de λ :

$$\lambda = 0.01/(\sum_{1\leq i\leq 20} f_i m_i) \tag{27}$$

Ici, on évalue par comptage la probabilité conditionnelle que Z_i mute en Z_j, sachant que i mute. On a pour cette dernière :

$$P_{ji}^i = A_{ji}/(\sum_{\{k/k\neq i\}} A_{ki}) \tag{28}$$

où A_{li} $(l \neq i)$ est le nombre de fois, calculé sur l'ensemble des paires de séquences, où face à l'acide aminé Z_i, on trouve l'acide aminé Z_l, $1 \leq i \neq l \leq 20$.

On en déduit alors la probabilité de transition i à j :

$$\begin{aligned} M_{ji} &= M_{\bar{i}i} \times P^i_{ji} \\ &= \lambda m_i \times A_{ji} / (\sum_{k \neq i} A_{ki}) \ ; \end{aligned} \tag{29}$$

et ; on a bien

$$\sum_{1 \leq j \leq 20} M_{ji} = 1 \tag{30}$$

De la même façon que nous avons défini la probabilité $p_S(i)$ de rencontrer un acide aminé Z_i qui persiste, par rapport au modèle aléatoire de choix ci-dessus souligné ; on peut exprimer la probabilité $s_S(i)$ de rencontrer un acide aminé Z_i qui mute. On a :

$$s_S(i) = f_i M_{\bar{i}i}, \tag{31}$$

$1 \leq i \leq 20$

On vérifie bien sûr que

$$p_S(i) + s_S(i) = f_i, \tag{32}$$

$1 \leq i \leq 20$

Et que,

$$\sum_{1 \leq i \leq 20} [p_S(i) + s_S(i)] = 1 \tag{33}$$

Résumons nous sur le principe de la méthode employée. On commence par fixer une probabilité (ou pourcentage) de mutation globale égale à 0.01. On considère alors un ensemble d'apprentissage formé d'une famille S de séquences, plus ou moins contemporaines et deux à deux alignées. On répartit alors cette probabilité globale sur les 20 acides aminés, proportionnellement, d'une part, à leurs mutabilités relatives respectives (il s'agit des $m_i, 1 \leq i \leq 20$) et d'autre part, à leurs fréquences relatives respectives d'exposition (il s'agit des $f_i, 1 \leq i \leq 20$) (21) et (26). Maintenant, la probabilité de mutation de l'acide aminé Z_i, $M_{\bar{i}i}$, se trouve répartie de façon proportionnelle par rapport à l'ensemble $\{j/j \neq i\}$ des 19 possibilités de mutation, $1 \leq i \leq 20$. Cette proportionnalité est calculée conformément à l'observation de S (29).

La matrice obtenue

$$\{M_{ji}/1 \leq i, j \leq 20\} \tag{34}$$

est de transition et correspond à une période unitaire d'évolution, caractérisée par 1 % de mutation globale. Pour une période d'évolution de longueur k dite k PAM ("Point Accepted Mutation"), on aura à élever à la puissance k la matrice ci-dessus. Comme nous l'avons exprimé dans l'introduction, nous travaillerons avec un matrice 250 PAM issue de [6]. Nous continuerons - sans ambiguïté - à la noter comme ci-dessus (34).

Quelle que soit la période d'évolution concernée, nous allons - dans le cadre du modèle aléatoire souligné ci-dessus - déterminer la probabilité que dans un site choisi au hasard d'une paire $\{s, s'\}$ de séquences alignées, on rencontre l'acide Z_i dans l'une des séquences et Z_j dans l'autre, $1 \leq i, j \leq 20$.

Le site étant fixé, imaginons d'abord le choix de s puis le choix de s'. Si i est différent de j, l'évènement qui nous intéresse est :"(rencontrer Z_i en s et Z_i mute en Z_j dans s') ou (rencontrer Z_j en s et Z_j mute en Z_i dans s')".La probabilité de cet évènement est

$$(f_i M_{ji} + f_j M_{ij}) \tag{35}$$

Si j est identique à i, l'évènement qui nous intéresse est :

"rencontrer Z_i en s et Z_i persiste dans s'".

La probabilité de cet évènement est

$$f_i M_{ii} \tag{36}$$

Maintenant, la probabilité du premier évènement dans l'hypothèse d'indépendance totale est

$$f_i f_j + f_j f_i = 2 f_i f_j \ ; \tag{37}$$

et celle du second :

$$f_i^2 \tag{38}$$

Dans ces conditions, les densités de probabilité du premier et second évènements se mettent respectivement sous la forme :

$$R_{ji} = \frac{f_i M_{ji} + f_j M_{ij}}{2 f_i f_j} \ et \ R_{ii} = \frac{f_i M_{ii}}{f_i^2}, \tag{39}$$

$1 \leq i, j \leq 20$
On se rend compte que la matrice

$$\{R_{ij} / 1 \leq i, j \leq 20\} \tag{40}$$

est symétrique. En fait et même plus, on admet généralement l'égalité :

$$f_i M_{ji} = f_j M_{ij}, \tag{41}$$

$1 \leq i, j \leq 20$; de sorte qu'on se contente de définir R_{ji} par la formule :

$$R_{ji} = \frac{f_i M_{ji}}{f_i f_j} = \frac{M_{ji}}{f_j}, \tag{42}$$

$1 \leq i, j \leq 20$; mais rien n'aurait empêché d'utiliser la première expression (39) que est plus précise.
Au lieu d'une paire, considérons ici un couple (s, s') de séquences distinctes, choisies uniformément au hasard et alignées. Pour un site choisi également au hasard, désignons par O_{ij} la probabilité de rencontrer Z_i dans s et Z_j dans s' ; O_{ij} est donné par le premier membre de (41) ; alors que O_{ji} par le second. et, on admet l'égalité entre O_{ij} et O_{ji} ; de sorte que :

$$R_{ij} = \frac{O_{ij}}{f_i f_j} \tag{43}$$

$1 \leq i, j \leq 20$.
La matrice de Dayhoff est la matrice des logarithmes à base 10 des nombres R_{ij} ; soit

$$D = \{S_{ij} = Log_{10}(R_{ij}) / 1 \leq i \leq j \leq 20\}, \tag{44}$$

qui est une demi matrice supérieure, diagonale comprise.
D peut être interprétée comme une matrice de similarités entre acides aminés ; mais où la similarité entre un acide aminé et lui même, n'est pas constante quel que soit ce dernier.

2.2 Conception de la matrice des Henikoffs et Comparaison

Pour introduire cette matrice [5], nous allons, pour des raisons d'homogénéïté, chercher à utiliser les notations qui ressemblent le plus à celles considérées ci-dessus.

Cette matrice s'obtient à partir d'une famille S de séquences alignées ; que cet alignement soit déjà donné ou construit de façon récursive. Nous allons considérer le cas direct où les différents éléments de S doivent avoir la même représentation dans le calcul. Les auteurs considèrent -pour des raisons biologiques- un cas plus complexe que nous évoquerons ci-dessous. Cependant, fondamentalement, la nature du calcul statistique est la même.

Relativement à S (1), la matrice des Henikoffs s'obtient au moyen d'un *comptage* dont le support est l'ensemble $P_2(S)$ des paires de séquences.

Reprenons ici les entiers

$$\{A_{ij} / 1 \leq j \leq i \leq 20\} \tag{45}$$

(28), où, rappelons le, A_{ij} est le nombre de sites dans l'ensemble des paires de séquences où les deux acides aminés Z_i et Z_j se retrouvent face à face. Plus précisèment, en notant A_{ij} $(\{s, s'\})$ le nombre de sites où Z_i et Z_j se retrouvent face à face, l'un dans s et l'autre dans s', on a :

$$A_{ij} = \sum \{A_{ij}(\{s, s'\})/_{\{}s, s'\} \in P_2(S)\}, \tag{46}$$

$1 \leq j \leq i \leq 20$.

On obtient alors la probabilité empirique d'une rencontre entre Z_i et Z_j au moyen de l'expression :

$$Q_{ij} = \frac{A_{ij}}{\sum \{A_{i'j'}/1 \leq j' \leq i' \leq 20\}}, \tag{47}$$

$1 \leq j \leq i \leq 20$.

On remarquera que la valeur du dénominateur est $LK(K-1)/2$ (1) ; puisque, pour chaque paire de séquences, on finit par calculer le nombre de sites qui est L ; en effet, chaque site intervient pour incrémenter d'une unité l'un des $A_{i'j'}, 1 \leq j' \leq i' \leq 20$. D'autre part, à la différence de l'équation (28), il s'agit ici d'une probabilité totale et non d'une probabilité conditionnelle. Q_{ij} jouera le même rôle que (35) dans le cas où $i \neq j$, respectivement (36) dans le cas où $i = j$. Précisèment, reprenons le modèle aléatoire ayant conduit à ces dernières expressions. Si i est différent de j ; il s'agit de la probabilité de rencontrer Z_i en s et Z_j en s', ou bien Z_j en s et Z_i en s'. Si i est identique à j ; il s'agit de la probabilité de rencontrer Z_i, à la fois en s et s'. Par ailleurs, on a déjà en (37) et (38) les probabilités respectives de ces deux évènements dans l'hypothèse d'indépendance totale. Nous notons

$$E_{ij} = \begin{cases} 2f_i f_j & \text{si } i \neq j; \\ f_i^2 & \text{si } j = i. \end{cases} \tag{48}$$

Pour nous rapprocher au mieux de l'expression calcul des auteurs mentionnés, on constatera que le numérateur de f_i (9) peut se mettre sous la forme :

$$2A_{ii} + \sum_{\{j/j \neq i\}} A_{ij}, \ ; \tag{49}$$

de sorte que f_i s'écrit :

$$f_i = Q_{ii} + \frac{1}{2} \sum_{\{j/j \neq i\}} Q_{ij}, \tag{50}$$

$1 \leq i \leq 20$.

On introduit ensuite la densité en (i, j) :

$$r_{ij} = \frac{Q_{ij}}{E_{ij}} \ ; \tag{51}$$

et, comme dans le cas de la matrice de Dayhoff, le logarithme, mais maintenant à base 2, de r_{ij} :

$$s_{ij} = log_2(r_{ij}), \tag{52}$$

$1 \leq i \leq j \leq 20$.

La matrice

$$H = \{h_{ij} = 2s_{ij}/_1 \leq i \leq j \leq 20\} \tag{53}$$

est alors considérée comme évaluant les degrés respectifs de substitution entre les différents acides aminés.

Cherchons maintenant à préciser clairement ce qui différencie la conception de la nouvelle matrice de celle, plus élaborée, plus fine mais plus hypothétique, de Dayhoff. On désignera ci-dessous par (D) le cas "Dayhoff" et par (H), le cas "Henikoff".

Alors que dans (H), on constate - au niveau de l'ensemble d'apprentissage S - une probabilité empirique de *persistance* sous la forme :

$$\sum_{1 \leq i \leq 20} Q_{ii}, \tag{54}$$

on impose dans (D) qu'elle soit égale à 0.99 :

$$\sum_{1\leq i\leq 20} O_{ii} = \sum_{1\leq i\leq 20} f_i M_{ii} = 0.99 \tag{55}$$

La notion originale de mutabilité relative d'un acide aminé (13) permet de distribuer de façon proportionnelle (21), la probabilité complémentaire de mutation (26). La probabilité de mutation $M_{\bar{i}i}$ d'un acide aminé i est répartie sur les différentes possibilités de mutation, conformément à l'observation de l'ensemble d'apprentissage S (28) et (29). La matrice des probabilités d'occurrences

$$\{O_{ij}/1 \leq i \leq j \leq 20\} \tag{56}$$

est alors déduite de façon conséquente de la matrice de transition (34) et des probabilités empiriques a priori $f_i, 1 \leq i \leq 20$, (8), (9), (41) et (43).

Alors que la matrice des probabilités d'occurences

$$\{Q_{ij}/1 \leq i \leq j \leq 20\} \tag{57}$$

qui peut correspondre à (56) est calculée dans (H) directement à partir de l'ensemble S d'apprentissage (47).

On peut signaler ici que la matrice carrée

$$\{O_{ij}/1 \leq i, j \leq 20\}, \tag{58}$$

qui est symétrique par condition, est celle d'une distribution jointe de probabilité dont la distribution marginale commune est donnée par

$$\{f_i/1 \leq i \leq 20\}\cdot \tag{59}$$

Si on veut à partir de (57) obtenir une matrice même type que (58), on aura à reprendre le même argument entre les expressions (42) et (43) et on sera conduit à poser :

$$(\forall 1 \leq i < j \leq 20), O'_{ij} = O'_{ji} = \frac{1}{2} Q_{ij}$$

et

$$(\forall\ 1 \leq i \leq 20), O'_{ii} = Q_{ii}\cdot \tag{60}$$

La matrice

$$\{O'_{ij}/1 \leq i, j \leq 20\} \tag{61}$$

qui est symétrique par construction, est alors celle d'une distribution jointe sur $I \times I$ (où $I = 1, 2, ..., i, ..., 20$) dont la distribution marginale commune est donnée par (59).

Cependant, le calcul direct des Q_{ij} dans (H) (47) peut être biaisé dans sa signification biologique si on peut décomposer S en B blocs :

$$S = \sum \{S_b/1 \leq b \leq B\} \tag{62}$$

(somme ensembliste),

où les tailles des différents blocs sont par trop hétérogènes et où chacun des blocs S_b est formé de séquences protéïques fortement homologues, $1 \leq b \leq B$. On remarquera qu'une méthode de classification automatique permet aisément d'obtenir la décomposition (62). Toutefois, les auteurs l'obtiennent par un algorithme d'enchaînement successifs. Ainsi, imaginons qu'on soit à la r-ème étape de la formation de S_b qui comprend donc l'ensemble des séquences alignées

$$\{s_{b1}, s_{b2}, ..., s_{bq}, ..., s_{br}\}\cdot \tag{63}$$

On y introduira la $(r+1)$ème séquence $s_{b(r+1)}$, si dans l'ensemble précédent des séquences, on peut trouver au moins une séquence s_{bq} qui peut être alignée avec $s_{b(r+1)}$, avec un taux d'au moins p % de sites où le résidu est conservé d'une séquence à l'autre. BLOSUM 62 correspond précisèment à $p = 62$.

Relativement à la décomposition (62), les classes $S_b, 1 \leq b \leq B$, vont être équipondérées ; chacune des classes comptant comme une seule séquence. Plus précisèment, on remplacera (46) par

$$A_{ij} = \sum\{A_{ij}(b,b')/1 \leq b < b' \leq B\}, \tag{64}$$

où

$$A_{ij}(b,b') = \frac{1}{k_b \times k_{b'}} \sum\{A_{ij}(\{s,s'\})/(s,s') \in S_b \times S_{b'}\}, \tag{65}$$

où

$$k_b = card(S_b) \text{ et } k_{b'} = card(S_{b'})$$

A la matrice des s_{ij} $(1 \leq i \leq j \leq 20)$ (52), les auteurs associent deux indices qui correspondent à deux types de moyennes. Les deux concernent la distribution des nombres s_{ij}. Le premier est une moyenne par rapport à la loi de probabilité (57) et le second, se veut par rapport au carré de la loi marginale (59). Le premier indice appelé "entropie relative" s'écrit

$$H = \sum\{Q_{ij} \times s_{ij}/1 \leq i \leq j \leq 20\} \tag{66}$$

et le second - il y a à cet égard une erreur dans l'article - doit s'écrire :

$$E = \sum\{f_i \times f_j \times s_{ij}/1 \leq i,j \leq 20\} \tag{67}$$

où on aura posé

$$s_{ij} = s_{ji}$$

pour tout $i \neq j, 1 \leq i,j \leq 20$. De cette façon, la somme des pondérations probabilistes est bien égale à 1.

C'est sur la base de l'indice H qui est trop global, donc imprécis, que les auteurs précités comparent leurs matrices $BLOSUM$ à celles de Dayhoff, pour différentes valeurs de PAM.

2.3 Les matrices AVL

Les ingrédients pour former une matrice AVL sur l'ensemble des 20 acides aminés sont d'une part, la distribution de probabilité empirique jointe (58) ou (61) ; d'autre part, la distribution marginale (59).

Le point de départ de la construction de cette matrice est le coefficient d'association que nous avons établi entre attributs booléens ; en nous référant à une hypothèses d'absence de liaison à caractère Poissonnien [8], [9], [11]. Si (a_i, b_j) est un couple de tels attributs booléens observés sur un ensemble E d'objets de taille n, la forme corrélationnelle d'un tel coefficient, où on neutralise n, est :

$$\rho_{ij} = \frac{f(i \wedge j) - f_i g_i}{\sqrt{f_i g_j}}, \tag{68}$$

où $f(i \wedge j)$ est la proportion des objets où la conjonction des deux attributs $a_i \wedge b_j$ est à 'vrai' (a_i et b_j sont à la fois présents) et où f_i (resp. g_j) est la proportion des objets où a_i (resp. b_j) est à 'vrai' [a_i (resp. b_j) est présent].

On rappelle que - pourvu que nf_ig_j ne soit pas petit- $\sqrt{n}\rho_{ij}$ est, dans l'hypothèse d'indépendance statistique entre a_i et b_j, considéré comme une réalisation d'une variable aléatoire $N(0,1)$, normale, centrée et réduite.

Maintenant, supposons que $\{a_i/1 \leq i \leq I\}$ (resp.$\{b_j/1 \leq j \leq J\}$) est l'ensemble des attributs-modalités (ou valeurs) d'une variable qualitative (on dit encore catégorielle) a (resp. b). Le croisement entre a et b conduit à un tableau de contingence auquel on peut associer le tableau des fréquences relatives

$$\{f(i \wedge j)/1 \leq i \leq I, 1 \leq j \leq J\}, \tag{69}$$

dont la marge colonne est définie par la distribution $\{f_i/1 \leq i \leq I\}$.

Si n est la taille de l'ensemble des objets sur lequel se trouve bâti le tableau de contingence ; alors $\sqrt{n}\rho_{ij}$ (68) est ce que nous appelons : "la contribution orientée de la case (i,j) à la statistique du χ^2 attachée au tableau de contingence". ρ_{ij} est donc la contribution de la case (i,j) à l'indice

$$\frac{\chi^2}{n} = \sum \left\{ \left[\frac{f(i \wedge j) - f_i g_j}{\sqrt{f_i g_j}} \right]^2 /1 \leq i \leq I, 1 \leq j \leq J \right\}, \tag{70}$$

qui ne dépend que du tableau des proportions (69).

Dans le cas qui nous intéresse, la variable b est identique à la variable a qui représente la variable "acide aminé". Cette dernière présente un ensemble de 20 valeurs ou modalités et le tableau (69) devient celui (58) ou (61). Formellement, il s'agit d'une situation analogue à celle du tableau des fréquences relatives associé à un tableau de Burt. Une situation comparable peut également être fournie par ce que l'on appelle "une matrice de confusion". La matrice (69) devient symétrique.

Pour la comparaison quantifiée deux à deux de l'ensemble A des 20 acides aminés, l'indice (68) devient

$$\rho_{ij} = \frac{f(i \wedge j) - f_i f_j}{\sqrt{f_i f_j}}, \tag{71}$$

$1 \leq i, j \leq 20$. Comme celle (69), la matrice des coefficients (71) est symétrique.

Une telle matrice conduit à une matrice AVL d'indices qui se réfèrent à une échelle de probabilité et que nous appelons "probabilistes". La transformation $-Log_2$ permet alors de transformer cette matrice en une matrice d'indices de dissimilarité "informationnels".

On substitue à la matrice des indices

$$\{\rho_{ij} / 1 \leq i, j \leq 20\}, \tag{72}$$

une matrice d'indices "globalement" normalisés

$$\{\rho^g_{ij} / 1 \leq i, j \leq 20\} \tag{73}$$

où ρ^g_{ij} se déduit de ρ_{ij} après centrage et réduction relativement à une distribution qu'il y a lieu de considérer. A cette fin, on peut prendre la distribution de probabilité jointe

$$\{f(i \wedge j) / 1 \leq i, j \leq 20\}; \tag{74}$$

comme si, ρ_{ij} a été observé avec la probabilité $f(i \wedge j)$, $1 \leq i, j \leq 20$. Dans ces conditions,

$$\rho^g_{ij} = \frac{\rho_{ij} - m_1(\rho)}{\sqrt{\mathrm{var}_1(\rho)}}, \tag{75}$$

$1 \leq i, j \leq 20$; où

$$m_1(\rho) = \sum \{f(i \wedge j)\rho_{ij} / 1 \leq i, j \leq 20\} \tag{76}$$

et

$$\begin{aligned} \mathrm{var}_1(\rho) &= \sum \left\{ f(i \wedge j)[\rho_{ij} - m_1(\rho)]^2 / 1 \leq i, j \leq 20 \right\} \\ &= \sum_{i,j} f(i \wedge j)\rho^2_{ij} - [m_1(\rho)]^2; \end{aligned} \tag{77}$$

où, rappelons le encore une fois ici, le rôle de la matrice (74) peut être joué par (58) si on adopte "Dayhoff" ou bien par (61), si on adopte "Henikoff".

Une autre réduction globale des similarités (12) peut être effectuée par rapport au carré de la loi marginale $\{f_i / 1 \leq i \leq 20\}$ (59). Ainsi, au lieu de (74), on considère

$$\{f_i f_j / 1 \leq i, j \leq 20\} \cdot \tag{78}$$

Le coefficient normalisé s'écrit ici

$$\rho^h_{ij} = \frac{\rho_{ij} - m_2(\rho)}{\sqrt{var_2(\rho)}}, \tag{79}$$

$1 \leq i, j \leq 20$,
où

$$m_2(\rho) = \sum \{f_i f_j \rho_{ij} / 1 \leq i, j \leq 20\} \tag{80}$$

et

$$var_2(\rho) = \sum\{f_i f_j[\rho_{ij} - m_2(\rho)]^2 / 1 \leq i,j \leq 20\}$$
$$= \sum_{(i,j)} f_i f_j \rho_{ij}^2 - [m_2(\rho)]^2 \quad (81)$$

Désignons par

$$\{\sigma_{ij} / 1 \leq i,j \leq 20\} \quad (82)$$

la matrice adoptée des coefficients globalement normalisés tels que (75) ou (79). La matrice des indices probabilistes de la "vraisemblance du lien", s'écrit alors

$$\{P_{ij} = \Phi(\sigma_{ij}) / 1 \leq i,j \leq 20\} \quad (83)$$

où ϕ est la fonction de répartition de la loi normale $N(0,1)$, centrée et réduite.
La matrice des indices de dissimilarité "informationnels" est alors :

$$\{\delta_{ij} = -Log_2(P_{ij}) / 1 \leq i,j \leq 20\} \quad (84)$$

δ_{ij} est la quantité d'information de l'évènement de probabilité $P_{ij}, 1 \leq i,j \leq 20$.
S'il s'agit de travailler avec une matrice de similarités entre acides aminés, nous travaillerons, selon la nature du problème, avec la matrice (83) ; ou bien, celle qui s'en déduit de façon croissante :

$$\{\sigma_{ij} = (\delta_{\max} - \delta_{ij}) / 1 \leq i,j \leq 20\} \cdot \quad (85)$$

3 SIMILARITE AVL SUR UN ENSEMBLE ALIGNE DE SEQUENCES

3.1 Introduction

Nous allons commencer par expliciter concrètement comment nous obtenons les cœfficients de la forme (75) respectivement associés à la matrice de Dayhoff et à celle des Henikoffs. Nous noterons ici, $QG(D)$ et $QG(B)$ les matrices des indices (75) qui correspondent, respectivement, à celle de Dayhoff (pour 250 PAM) et à celle, dite BLOSUM 62, des Henikoffs :

$$QG(D) = \{QG_d(i,j) / 1 \leq j \leq i \leq 20\} \quad (86)$$

et

$$QG(B) = \{QG_b(i,j) / 1 \leq j \leq i \leq 20\} \quad (87)$$

La matrice de Dayhoff adoptée est très précisèment celle sous diagonale (comprise) de la table I de la page 279 de [6]; et celle, BLOSUM 62, est fournie à partir de la matrice sous diagonale (comprise) de la figure 2 de la page 10917 de [5]. Il y a lieu également de signaler que c'est à partir de la 3ème colonne de la table III du premier article cité que nous recueillons la distribution marginale des fréquences relatives $\{f_i / 1 \leq i \leq 20\}$ (59). Cette distribution servira aussi bien pour réduire la matrice de Dayhoff que celle des Henikoff. Ainsi - comme nous l'avons exprimé - c'est la fréquence jointe $f(i \wedge j)$ qui se trouve déterminée d'une façon ou d'une autre, selon que l'on adopte "Dayhoff" ou "Henikoff" (71), (74), (58) et (61), $1 \leq i,j \leq 20$.

Les deux matrices que nous venons de référencer seront désignées par $\tau(D)$ pour celle de Dayhoff-250 PAM ; et par $\tau(B)$ pour celle BLOSUM 62.

Après avoir montré comment explicitement obtenir $QG(D)$ et $QG(B)$ (86) et (87) à partir respectivement de $\tau(D)$ et $\tau(B)$, nous indiquerons une fois de plus (84), les matrices de dissimilarité "informationelle", respectivement notées $\delta(D)$ et $\delta(B)$; et enfin, celles qui s'en déduisent immédiatemment, $\sigma(D)$ et $\sigma(B)$, de similarité "informationnelle". C'est au niveau de ces matrices que nous proposerons des valeurs de comparaison entre un acide aminé fixé et une déletion, ainsi qu'entre deux déletions. Nous pourrons également considérer le cas où la comparaison concerne un acide aminé qui n'a pas pu être identifié. Mais ce cas est de plus en plus rare.

Nous montrerons enfin, comment - à partir de $\sigma(D)$ ou de $\sigma(B)$ - on induit une similarité AVL entre séquences protéïques.

3.2 Matrices $QG(D)$ et $QG(B)$

La matrice $\tau(D)$ représente 10 fois la matrice (44) de terme général noté S_{ij}. Par conséquent

$$\tau(D) = \{S^{'}(i,j)/1 \leq j \leq i \leq 20\} \tag{88}$$

donne la matrice des R_{ij} (43) à partir de

$$R_{ij} = 10^{0.1 \times S^{'}(i,j)}, \tag{89}$$

$1 \leq j \leq i \leq 20$. Dans ces conditions,

$$O_{ij} = f_i f_j R_{ij} = f_i f_j 10^{0.1 S^{'}(i,j)} \tag{90}$$

$1 \leq j \leq i \leq 20$. Ayant les O_{ij}, on obtient aisément, en tenant compte de la symétrie en (i,j), les coefficients $\rho_{ij}(D)$ (71) ; puisque $f(i \wedge j)$ est alors identique à O_{ij}, $1 \leq i,j \leq 20$. La normalisation statistique indiquée autour des expressions (73) à (77), conduit alors aux indices de la forme (75), que nous avons ici indiqués par $QG_d(i,j), 1 \leq j \leq i \leq 20$; et ce, afin de nous rapprocher des notations adoptées dans AVL.

La matrice $\tau(B)$ est celle déjà notée H (53). Le $f(i \wedge j)$ de $\rho_{ij}(B)$ (71), correspond ici à $O^{'}_{ij}$ (60) et (61), $1 \leq j \leq i \leq 20$. Compte tenu des expressions (51) et (52), on a pour tout (i,j), $1 \leq j \leq i \leq 20$,

$$O^{'}_{ij} = f_i f_j \times 2^{0.5 h_{ij}}, \tag{91}$$

où h_{ij} est le (i,j)ème terme de la matrice $\tau(B)$ (53).

Ayant les $O^{'}_{ij}$, on a les ρ_{ij} (B) (71), à partir desquels on obtient les ρ^{g}_{ij} (B) (75), notés ici $QG_b(i,j)$; où, encore une fois, les $f(i \wedge j)$ sont remplacés par les $O^{'}_{ij}$, $1 \leq i,j \leq 20$. On obtient ainsi $QG(B)$.

3.3 Les indices de similarités "probabilistes" ou "informationnels" sur l'ensemble des acides aminés.

Conformément à (83) les indices de similarité probabilistes respectivement associés à $QG(D)$ et $QG(B)$, sont définis dans les matrices

$$P(D) = \{P_d(i,j) = \Phi[QG_d(i,j)]/1 \leq i,j \leq 20\} \tag{92}$$

et

$$P(B) = \{P_b(i,j) = \Phi[QG_b(i,j)]/1 \leq i,j \leq 20\}. \tag{93}$$

Maintenant, conformément à (84), les matrices des indices de dissimilarité "informationnels" s'écrivent :

$$\delta(D) = \{\delta_d(i,j) = -Log_2[P_d(i,j)]/1 \leq i,j \leq 20\} \tag{94}$$

et

$$\delta(B) = \{\delta_b(i,j) = -Log_2[P_b(i,j)]/1 \leq i,j \leq 20\} \tag{95}$$

Considérons à présent, conformément à (85), les valeurs maximales $\delta_{max}(D)$ et $\delta_{max}(B)$ de chacune des deux matrices $\delta(D)$ et $\delta(B)$ (94) et (95). Nous substituerons à chacune des ces deux dernières matrices, les matrices, respectivement associées, de similarité "informationnelle".

$$\sigma(D) = \{\sigma_d(i,j) = [\delta_{max}(D) - \delta_d(i,j)]/1 \leq i,j \leq 20\} \tag{96}$$

et

$$\sigma(B) = \{\sigma_b(i,j) = [\delta_{max}(B) - \delta_b(i,j)]/1 \leq i,j \leq 20\} \tag{97}$$

C'est sur la base de ces matrices que nous établirons nos matrices d' indices de similarité entre séquences protéïques. Mais, il s'agit d'abord de pouvoir *comparer un acide aminé donné avec une déletion (notée -) ainsi que, deux déletions entre elles.* Il s'agit aussi, bien que cela soit moins crucial, de pouvoir intégrer dans la comparaison le cas d'un acide aminé non identifié que nous notons X. En d'autres termes, il s'agit d'inférer à partir de $\sigma(D)$ [resp. $\sigma(B)$] une matrice de similarité sur un alphabet $A^{''}$ de 22 lettres :

$$A^{''} = A \cup \{-, X\} \tag{98}$$

Désignons par σ l'indice adopté ; il peut s'agir de l'indice σ_d (96) ou bien σ_b (97). D'autre part, indiquons par Z_i l'un des acides aminés, $1 \leq i \leq 20$. Pour répondre au problème d'inférence posé, nous avons à considérer les cas suivants de comparaison :

$(Z_i, -), (-, -), (Z_i, X), (-, X)$ et (X, X).

Comme nous avons pu l'exprimer, le cas le plus important est celui de la déletion. L'évaluation des associations dans ce cas reposera sur l'interprétation suivante, de la substitution d'un acide aminé Z par une déletion, en tant qu'évènement :

"Z allait se transformer en un *autre* acide aminé mais cela ne s'est pas produit"

Dans ces conditions, nous proposerons pour la comparaison entre l'acide aminé Z_i et une déletion-, la moyenne *pondérée* des indices d'association entre Z_i et les *autres* acides aminés. Cette *pondération* tient précisèment compte des probabilités de passage entre Z_i et les différents *autres* acides aminés. On a ainsi,

$$\sigma(Z_i, -) = \frac{1}{f_i} \sum_{\{j/j \neq i\}} f(i \wedge j)\sigma(i, j), \tag{99}$$

où $f(i \wedge j)$ aura à être remplacé par O_{ij} (90) si on se réfère à σ (D) (96) et par $O^{'}_{ij}$ (91) si on se réfère à σ (B) (97). On peut signaler que cette optique est fondamentalement différente de celle adoptée dans [4].

Maintenant, conformément à l'interprétation ci-dessus, la comparaison d' une déletion avec elle même, est du même ordre que la comparaison d'un acide aminé avec lui même. Mais, ce dernier n'étant pas spécifié, on prendra la moyenne - pondérée par les fréquences d'exposition - des indices de similarité(σ_d ou σ_b) entre un acide aminé et lui même ; ainsi

$$\sigma(-, -) = \sum_{1 \leq i \leq 20} f_i \sigma(i, i) \tag{100}$$

X est un acide aminé non identifié, nous considèrerons qu'il peut être Z_j avec la probabilité $f_j, 1 \leq j \leq 20$. Dans ces conditions, la valeur de la similarité à proposer entre un acide aminé Z_i et X est

$$\sigma(Z_i, X) = \sum_{1 \leq j \leq 20} f_j \sigma(Z_i, Z_j), \tag{101}$$

$1 \leq i \leq 20$.

Compte tenu de l'interprétation de $X, \sigma(X, -)$ est la moyenne de $\{\sigma(Z_i, -)/1 \leq i \leq 20\}$ (99), par rapport à la distribution $\{f_i/1 \leq i \leq 20\}$. On a donc

$$\sigma(X, -) = \sum_{1 \leq i \leq 20} f_i \sigma(Z_i, -). \tag{102}$$

Il reste enfin le cas de la comparaison (X, X). De façon cohérente, nous proposons :

$$\sigma(X, X) = \sum \{f(i \wedge j)\sigma(i, j)/1 \leq i, j \leq 20\}, \tag{103}$$

avec le sens que l'on sait de $f(i \wedge j)$ selon qu'on se réfère à $\sigma(D)$ (96) où à $\sigma(B)$ (97).

Concrètement, dans la famille des 89 séquences, sur laquelle nous allons proposer nos classifications, tous les acides aminés sont reconnus ; de sorte que nous travaillerons avec un alphabet $A^{'}$ comprenant 21 symboles. Très précisèment, on a :

$$A^{'} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, -\}. \tag{104}$$

Compte tenu des équations (99) et (100), à partir d'une matrice de similarité entre acides aminés {il peut s'agir de celle de Dayhoff, de celle des Henikoffs ; ou bien de celles, σ (D) et σ (B), respectivement associées (96) et (97)}, on établira la matrice sous diagonale (comprise) concernant $A^{'}$ et comprenant par conséquent $21 \times 22/2 = 231$ termes. La symétrie de l'association fait qu'une telle matrice détermine un graphe valué complet sur $A^{'}$. Il s'agira de l'un des codages de la variable "site".

Le deuxième codage que nous envisagerons est ordinal. $\{Z_i/1 \leq i \leq 20\}$ représentant $A^{'}$ (104), la valuation définie par une matrice de similarité sur $A^{'}$ du type de celles ci-dessus considérées, induit un préordre total sur l'ensemble C des couples

$$C = \{(Z_i, Z_j)/1 \leq j \leq i \leq 20\}; \tag{105}$$

et que nous appelons "préordonnance" sur A'.

En rangeant de gauche à droite, la position d'un même couple est d'autant plus portée à droite que la ressemblance entre ses deux composantes est plus grande. Nous coderons cette position à partir de la notion de "rang moyen" [13] ; en effet, le cas des exaequos se présente dans un tel rangement.

Il est à remarquer que la valuation numérique ou ordinale, concernant la comparaison d'un acide aminé avec lui même, n'est pas constante. Elle n'est pas non plus nécessairement, toujours supérieure à la comparaison entre deux acides aminés différents. Mais cela n'influence en rien notre méthode d' association entre séquences protéïques.

3.4 Indices de similarités entre séquences protéïques pour la classification AVL.

Nous allons considérer huit traitements qui sont deux à deux appareillés. Chaque paire de traitements sera associée à une même matrice. Pour une matrice donnée, le premier traitement concerne la valuation ou codage numérique ; et le second, le codage ordinal en termes de "préordonnance" sur A' [cf.ci-dessus]. Les matrices que nous allons considérer sont d'une part, $\tau(D)$ et $\tau(B)$ [cf. § III.1, (44) et (53)] et d'autre part, $\sigma(D)$ et $\sigma(B)$. Ainsi, en indiquant NM pour "numérique" et PO pour "préordonnance", les quatre paires de traitements seront - avec des notations que l'on comprend - notés :

$\{NM[\tau(D)], PO[\tau(D)]\}$,
$\{NM[\tau(B)], PO[\tau(B)]\}$,
$\{NM[\sigma(D)], PO[\sigma(D)]\}$
et
$\{NM[\sigma(B)], PO[\sigma(B)]\}$.

Un traitement de type PO ne diffère d'un traitement de type NM que parce que la valuation numérique attachée à un élément de C, est remplacée par son rang, à partir de la notion de "rang moyen" que nous avons déjà évoquée. Il s'agit donc aussi d'une valuation numérique d'un type cardinal. L'intérêt de ce deuxième codage ordinal consiste à se rendre compte de la stabilité des résultats. Et, dans le cas où il y a différence, il s'agit de choisir entre les deux codages.

L'ensemble des séquences protéïques est considéré ici comme un ensemble d'objets élémentaires. L'alignement multiple permet d'associer à chaque site une variable descriptive. L'ensemble des valeurs de la variable est toujours le même (104). Comme nous venons de le signaler, la sémantique que nous adoptons sur l'ensemble A' des valeurs correspond soit à un graphe valué, soit à une préordonnance. C'est - bien sûr - la différence de comportement entre les différents sites, à travers l'ensemble des séquences protéïques, qui fera la classification.

Bien que les différents sites n'aient pas la même importance biologiques et qu'il existe entre ces derniers des liens complexes, notre coefficient de similarité attribuera à chaque variable-site, le même pouvoir global de discrimination. C'est a priori ; c'est à dire, en cas d'ignorance, la solution la plus logique. Ce sont les interliaisons statistiques cachées entre variables qui vont permettre l'émergence de la classification.

Désignons par

$$\{\sigma(i,j)/1 \leq j \leq i \leq 21\} \tag{106}$$

l'une des similarité sur A' (104) issue de $\tau(D)$, $\tau(B)$, $\sigma(D)$ ou $\sigma(B)$ et qui détermine une valuation sur C (105). Indiquons aussi par

$$\{r_\sigma(i,j)/1 \leq j \leq i \leq 21\} \tag{107}$$

le codage "rang moyen" de la préordonnance associée. Comme nous venons de l'exprimer, l'usage de (106) correspond à un traitement NM et celui (107), à un traitement PO. De sorte que nous désignerons par

$$\{u(i,j)/1 \leq j \leq i \leq 21\} \tag{108}$$

l'une ou l'autre des deux matrices (106) ou (107).

Considérons à présent une paire $\{s_k, s_{k'}\}$ de séquences dans l'ensemble S des séquences (alignées) (5). Nous allons préciser la contribution du l-ème site dans la comparaison entre s_k et $s_{k'}$. Conformément aux notations introduites (4), $\{s_k(l), s_{k'}(l)\}$ est la paire d'acides aminés qu'on rencontre. Nous introduisons alors la contribution brute de l à la comparaison entre les deux séquences s_k et $s_{k'}$, sous la forme :

$$c^l(k,k') = u[s_k(l), s_{k'}(l)] \tag{109}$$

Cet indice brut est centré et réduit sur l'ensemble des couples de séquences ; ce qui permet de définir la contribution *normalisée* du site l, à la comparaison des deux séquences k et k' :

$$C^l(k,k') = \frac{c^l(k,k') - moy_e(c^l)}{\sqrt{var_e(c^l)}}, \tag{110}$$

où $moy_e(c^l))$ et $var_e(c^l)$ sont respectivement la moyenne et la variance de $c^l(k,k')$ (109) sur l'ensemble de tous les couples de séquences ; soit sur $\underline{K} \times \underline{K}$, où $\underline{K} = \{1, 2, ..., k, ..., K\}$. Nous avons bien une expression élégante du second membre de (110) qui intègre la distribution de la variable site l, sur l'ensemble des séquences [13].

Maintenant - à cette étape - c'est de façon additive qu'on tient compte de l'ensemble des L sites (3), pour définir

$$C(k,k') = \sum_{1 \le l \le L} C^l(k,k') \tag{111}$$

qui est la somme des contributions des différentes variables sites, $1 \le k < k' \le K$.

Le précédent coefficient est globalement statistiquement normalisé sur l'ensemble des paires de séquences (5), pour obtenir

$$Q_s(k,k') = \frac{C^l(k,k') - moy_e(C^l)}{\sqrt{var_e(C^l)}}, \tag{112}$$

où $moy_e(C^l)$ et $var_e(C^l)$ sont la moyenne et la variance empiriques de C^l sur l'ensemble $P_2(\underline{K})$, $1 \le k < k' \le K$.

Dans ces conditions, l'indice probabiliste de la vraisemblance du lien se met sous la forme

$$P(k,k') = \phi[Q_s(k,k')], \tag{113}$$

où ϕ est la fonction de répartition de la loi normale centrée réduite, $1 \le k < k' \le K$

C'est la table des valeurs (113) qui est donnée comme argument au critère de la vraisemblance du lien maximal dans la méthode AVL de Classification Ascendante Hiérarchique.

Dans une telle méthode, on prévoit un paramétrage du critère de formation ascendante de l'arbre des classifications par un nombre ε, compris entre 0 et 1. c'est la valeur $\varepsilon = 0.5$ soit $AVL_{0.5}$ que nous appliquons avec le plus de bonheur dans nos expériences. [10]

4 ANALYSE COMPARATIVE DES DIFFERENTS TRAITEMENTS

Nous allons dans ce paragraphe analyser les résultats obtenus suivant trois traitements différents : i) Analyse Factorielle des Correspondances (AFC), entre les 20 acides aminés, telles que définies à partir des matrices O_{ij} et O'_{ij} ; ii) classification AVL des acides aminés à partir des matrices $QG(D)$ et $QG(B)$; iii) classifications AVL de 89 cytochromes c. Enfin, nous comparerons la méthode AVL adoptée avec une classification phylogénique issue d'un ensemble de programmes, appelé PHYLIP et considéré comme l'un des meilleurs par les biologistes.

4.1 Analyse factorielle de la correspondance entre acides aminés.

Une première différence saute aux yeux entre les matrices O_{ij} (Dayhoff) et O'_{ij} (Blosum) : dans le premier cas (cf. figure 1), les valeurs propres décroissent assez rapidement puisque les 5 premières expliquent 83 % de l'inertie totale, alors que pour la matrice Blosum (cf. figure 2), il faut les 9 premières valeurs propres pour obtenir le même résultat. En soi, cette observation ne donne aucune indication sur la qualité des matrices de départ, elle indique simplement qu'il faut plus de paramètres (inconnus) pour décrire un acide aminé suivant Blosum que suivant Dayhoff. L'examen des premiers plans factoriels montre bien, en effet, que la matrice de Dayhoff conduit à des regroupements facilement identifiables (A, P, T) (I, L, M, V) (D, E) ... alors que cet exercice est plus difficile pour Blosum. *Intuitivement*, le biologiste qui a l'habitude de regarder des alignements de séquences aura tendance à préférer la matrice de Dayhoff.

4.2 Classification AVL des acides aminés

Les classifications obtenues à partir des matrices $QG(D)$ et $QG(B)$ (cf. figures 3 et 4), diffèrent en certains points. Ici encore, nous manquons de critères objectifs pour faire un meilleur choix qui ne sera guidé que par l'habitude du biologiste. Celle obtenue à partir de $QG(B)$ présente des regroupements "inhabituels", en particulier (P,T) et (H,N). A l'inverse, il n'y a rien de "choquant" dans la classification issue de Dayhoff.

En conclusion de ces deux traitements : aucun critère vraiment objectif ne permet de décider quelle matrice, de Dayhoff ou de Blosum, est la meilleure. Cependant, l'habitude et l'intuition du biologiste font pencher la balance en faveur de la matrice de Dayhoff.

4.3 Classification des cytochromes c

Afin de tester à la fois et la méthode AVL sur des objets biologiques, et les matrices de similarité issues des traitements décrits dans les paragraphes précédents, nous avons procédé à la classification des séquences de 89 cytochromes c issus d'organismes différents. Le choix de cette protéïne résulte du fait qu'elle est présente chez la plupart des organismes vivants, depuis les algues et les bactéries jusqu'aux mammifères, et que l'on dispose de plus de 100 séquences différentes. On peut donc, à partir des séquences du cytochrome c, réaliser une classification AVL que l'on comparera à la classification taxonomique classique. Il n'est pas question ici de réaliser une "phylogénie" des espèces, puisque la longueur des branches de l'arbre AVL n'a aucune raison d'être proportionnelle au temps. En revanche, on doit s'attendre à ce qu'une méthode de classification sélective regroupe ensemble les séquences apparentées : les bactéries doivent former (au moins) une classe, les plantes une autre, les insectes une troisième, etc...Si, en outre, la méthode est sensible, on peut s'attendre à observer des sous-classes - par exemple, chez les animaux, les oiseaux devraient être séparés des mammifères.

Notons tout de suite que les séquences des cytochromes c sont très conservées : il n'y a, sur 104 acides aminés, que 16 différences entre l'homme et la tortue, et encore la grande majorité de ces substitutions concerne-t'elle des acides aminés "voisins". Par ailleurs on ne peut pas s'attendre à obtenir un arbre "parfait", car certaines espèces sont sous-représentées. Ainsi, on ne dispose que d'une séquence pour les nématodes (sorte de petits vers). Il s'ensuit que LA séquence du nématode risque de ne pas faire une classe à elle seule, mais de s'agglomérer au groupe qui lui est le plus proche dans le lot d'exemples.

4.3.1 Comparaison des traitements "numériques" (notés NM) et "préordonnances" (notés PO) avec les matrices τ et σ.

- d'une manière générale, quelle que soit la matrice et le traitement, la classification AVL regroupe correctement en 3 classes distinctes les séquences des micro-organismes, des plantes et des insectes. Les différences essentielles - qui permettront de choisir "le meilleur" arbre, se situent au niveau des mammifères et des oiseaux.

- $NM[\tau(D)]$ et $PO[\tau(D)]$: PO se comporte moins bien que NM dans la mesure où une partie des oiseaux se regroupe avec une partie des mammifères. Avantage à NM.

- $NM[\tau(B)]$ et $PO[\tau(B)]$: dans les deux cas les oiseaux forment une classe unique. PO classe correctement la tortue avec les oiseaux, ce que ne fait pas NM, mais PO regroupe le phoque et la tortue, l'étoile de mer et le cheval. Avantage à NM.

- $NM[\sigma(D)]$ et $PO[\sigma(D)]$: mauvais résultat de PO qui sépare trop le boeuf du cheval, regroupe la tortue et le boeuf. Très bonne classification de NM qui regroupe correctement les oiseaux, la tortue et le varan d'une part et les mammifères d'autre part, sans rapprocher - comme le font tous les autres - le crotale de l'homme. Net avantage à NM.

- $NM[\sigma(B)]$ et $PO[\sigma(B)]$: NM regroupe la tortue et la chauve-souris, sépare trop le cheval du boeuf. PO regroupe la tortue et le phoque, l'étoile de mer et le cheval. Mauvais score dans les deux cas.

  Donc : très clairement, NM se comporte mieux que PO.

- $NM[\tau(D)]$ et $PO[\tau(B)]$: les résultats sont comparables mais peu satisfaisants dans les deux cas. Les mammifères sont éclatés en deux classes dont une est liée aux oiseaux, la tortue est séparée des oiseaux.

- $NM[\sigma(D)]$ et $PO[\sigma(B)]$: avec $\sigma(B)$, rapprochement d'une partie des mammifères et des oiseaux, regroupement de la tortue avec la chauve-souris et du crotale avec l'homme. Score presque parfait avec $\sigma(D)$ où la seule grosse "erreur" est le rapprochement de la tortue et de la grenouille.

Conclusion générale :

a) le traitement NM est meilleur que le traitement PO

b) les matrices σ sont meilleurs que les matrices τ

c) la matrice $\sigma(D)$ est meilleure que la matrice $\sigma(B)$

d) **meilleur score :** $NM[\sigma(D)]$

Dans ces conditions, nous nous limiterons à présenter l'arbre $AVL(0.5)$ issu de $NM[\sigma(D)]$ (cf. figure 5).

4.4 Examen détaillé de la classification obtenue avec $NM[\sigma(D)]$

La figure 6 est un extrait de l'arbre précédent obtenu avec les 89 séquences de cytochrome c, duquel nous avons éliminé les micro-organismes et les plantes.

- dans le détail, il arrive souvent que la classification obtenue ne corresponde pas exactement à ce qu'on attendrait d'un arbre phylogénétique. Par exemple, le varan serait plus éloigné du poulet qu'il ne l'est ici. Mais la méthode AVL ne prétend en aucun cas être une méthode phylogénétique !

- on observe deux classes bien distinctes, l'une comprenant les oiseaux et l'autre les mammifères. Le varan (lézard) et la tortue sont correctement rattachés aux oiseaux. La grenouille ne devrait pas se trouver dans ce groupe et le kangourou devrait être plus nettement séparé des mammifères. Néanmoins, cette séparation en 2 classes bien distinctes est tout à fait satisfaisante.

- une troisième classe - qui comporte les cytochromes c2, et non les cytochromes c, du rat et de la souris, regroupe en fait les organismes pluricellulaires variés faisant partie du lot de séquences, et qui ne sont ni des oiseaux ni des mammifères. On s'attendrait simplement à ce que le crotale ait été le plus proche du varan et de la tortue.

4.5 Comparaison avec un programme de phylogénie

A titre de comparaison, nous avons procédé à une classification phylogénétique des mêmes séquences de cytochrome c au moyen d'un ensemble de programmes appelé PHYLIP, très employé dans le milieu et considéré comme l'un des meilleurs. Ici, la longueur des branches de l'arbre est censée être proportionnelle au temps. Les distances initiales entre les séquences sont estimées à partir de la matrice $PAM001$ de Dayhoff. Après calcul des distances et estimation des temps de divergence, les arbres ont été construits par les méthodes classiques du "neighbor joining" et "UPGMA".

a) Neighbor joining : peu satisfaisant, l'homme et le singe sont liés au crotale, proche des oiseaux et beaucoup trop éloignés des autres mammifères. La tortue est liée à tort à la grenouille et devrait être plus proche des oiseaux.

b) UPGMA : d'une part les mammifères ne sont pas regroupés, d'autre part l'homme et les singes sont plus éloignés de la souris, par exemple, que des oiseaux.

4.6 Conclusion

Du point de vue du biologiste, la classification obtenue par AVL $+NM[\sigma(D)]$ sur les séquences du cytochrome c est extrêmement satisfaisante. Certes, nous n'avons pas poussé le programme PHYLIP dans ces derniers retranchements, mais à coup sûr ce programme utilisé "normalement" donne de moins bons résultats sur notre lot d'exemple.

Afin de sérier les problèmes, c'est volontairement que les premiers essais ont été effectués sur des séquences qui se ressemblent beaucoup et comportent peu -ou pas - d'insertions/délétions. La méthode AVL appliquée à ces séquences, couplée à la matrice $NM[\sigma(D)]$, fournit des résultats qui montrent qu'elle est à la fois sensible et sélective, et qu'elle devrait pouvoir être utilisée avec profit par les biologistes.

La suite naturelle de ce travail est maintenant l'étude d'un lot de séquences plus hétérogènes, de longueurs différentes, et comportant beaucoup plus d'insertions/délétions. Nous disposons d'un tel lot d'exemple, constitué par les aminoacyl-tRNA synthétases dont nous avons réalisé un alignement multiple fiable. Leurs séquences sont tellement différentes que les programmes classiques de phylogénie ne peuvent les traiter (il y a "saturation").

5 CONCLUSION ET PERSPECTIVES

Un premier intérêt de ce travail concerne l'analyse de la nature de la matrice de Dayhoff, de celle des Henikoffs ; ainsi que de leur comparaison (cf. § II. 1 et II.2). Un point d'importance pour l'application de la méthode impliquée ici, est d'en déduire, respectivement, des matrices conformes à l'optique AVL (cf. § II.3). Mais, l'intérêt de ces matrices pourra aller bien au delà de l' objectif "classification".

Dans la démarche que nous avons entreprise, nous avons comparé dans le cadre de la classification AVL, les comportements des différentes matrices d'association entre acides aminés, pour élaborer un indice de similarité sur un ensemble aligné, de séquences protéïques. Et, à cet égard, une solution spécifique est considérée pour la comparaison entre un acide aminé et une déletion, ou bien, entre deux déletions.

Un alignement multiple se fondant sur des aspects structurels a un intérêt propre. Il a pour nous - comme nous l'avons exprimé en introduction - un intérêt supplémentaire qui consiste à rendre autonome la comparaison.

Pour une matrice donnée, la contribution d'un site est un graphe valué sur l' ensemble des séquences. Nous avons pu dans ces conditions nous rendre compte d'une certaine stabilité globale des résultats de la classification lorsqu'on remplace la valuation numérique par une valuation ordinale. Cette stabilité est d'une part, une qualité de la méthode ; mais d'autre part, caractérise l'aptitude de la matrice adoptée d'association (entre acides aminés), compte tenu des résultats obtenus par la classification.

La méthode AVL peut prendre en compte les formes les plus diverses de l'information similarité; et en particulier, n'importe laquelle des matrices d'association entre acides aminés. Cependant on peut espérer une cohérence accrue des résultats, si précisèment cette dernière matrice est conforme à l'optique d'AVL.

Ce travail en constitue précisèment la démonstration. Et alors, c'est le codage numérique qui serre au plus près la nature de l'information similarité, qui s'est avéré le plus performant.

L'une quelconque des matrices obtenues permet - via des algorithmes de type "programmation dynamique" - un alignement multiple [1]. Ce qui permet une classification de l'ensemble des séquences, avec une comparaison site par site. La pertinence des résultats obtenus est certes, fonction de la méthode de classification ; mais aussi, de la qualité de l'alignement multiple. Précisèment, un de nos immédiats projets futurs consiste à procéder à un alignement multiple avec une matrice AVL ; et, à faire suivre ce dernier par une classification AVL. Nous espérons ainsi, une cohérence accrue dans l'organisation classificatoire par rapport à la connaissance phylogénétique.

Signalons ici que l'optique que nous avons adoptée consiste à considérer chaque séquence protéïque comme un objet et chaque site comme une variable. Le point de vue dual est également très fortement considéré en AVL. Dans ce dernier cas, chaque protéïne est considérée comme un variable définissant sur l'ensemble des sites (ou objets), un graphe valué de la forme :

$$\{\lambda_k(\ell,\ell^{'}) = \sigma[s_k(\ell), s_k(\ell^{'}]/1 \leq \ell,\ell^{'} \leq L\} \tag{114}$$

où σ résulte de la matrice adoptée d'association entre acides aminés et où $s_k(\ell)$, relativement à la protéïne k, est précisé en (4). On se retrouve alors face à un problème de classification de graphes valués, où l' indice brut de comparaison de graphes entre k et $k^{'}$ se met sous la forme

$$\sum\{\lambda_k(\ell,\ell^{'})\lambda_{k^{'}}(\ell,\ell^{'})/1 \leq \ell,\ell^{'} \leq L\} \tag{115}$$

et où, le support du calcul est - en plus de la matrice d'association σ - une table de contingence 21×21, conformément à la taille de l'alphabet $A^{'}$ (104) [9], [15].

Reprenons ici un point évoqué dans l'introduction et qui revient souvent dans la littérature (voir par exemple dans [3], à propos de la circularité logique du raisonnement, lorsqu'il s'agit de procéder à un alignement multiple à partir d'une matrice d'association ou de subtitution entre acides aminés. En effet, nous l'avons vu (cf. § II), il faut bien partir d'un ensemble de séquences alignées, à tout le moins par paires, pour pouvoir déduire une telle matrice ; qui, à son tour, doit servir pour aligner.

Cependant, la démarche scientifique est classique et procède de ce que nous pourrons appeler la méthode des "approximations successives". Nous voulons imaginer que l'état initial est fourni par un ensemble $S^{(0)}$ de séquences aligné au mieux, à partir d'un ensemble S de séquences non aligné. Cet alignement d'origine est supposé obtenu à partir de la connaissance et d'indicateurs simples. Nous lui faisons correspondre, d'une part, la matrice $\sigma^{(0)}$ d'association entre acides aminés ; et d'autre part, la classification $Cl^{(0)}$ qui utilise $\sigma^{(0)}$ (cf.par exemple § III et § IV ci-dessus). $Cl^{(0)}$ permettra de juger de la cohérence relative de l'alignement. On obtient ainsi le triplet :

$$(S^{(0)}, \sigma^{(0)}, Cl^{(0)}) \tag{116}$$

A partir de $\sigma^{(0)}$; et en utilisant $Cl^{(0)}$, on obtient

$$S^{(1)} = (\text{alignement par } \sigma^{(0)})(S) \ ; \tag{117}$$

ce qui permet d'obtenir une nouvelle matrice $\sigma^{(1)}$; laquelle donnera la classification $Cl^{(1)}$ de $S^{(1)}$, en utilisant $\sigma^{(1)}$. D'où obtention du triplet :

$$(S^{(1)}, \sigma^{(1)}, Cl^{(1)}) \tag{118}$$

De nouveau, en utilisant $Cl^{(1)}$, on a

$$S^{(2)} = (\text{alignement par } \sigma^{(1)})(S^{(1)}) \ ; \tag{119}$$

et, ainsi de suite ... jusqu'à arriver à un état stable

$$(S^{(t)}, \sigma^{(t)}, Cl^{(t)}), \tag{120}$$

qui suppose la convergence du processus.

On peut alors grossir l'ensemble S et affiner ainsi, $\sigma^{(t)}$ et $Cl^{(t)}$.

Nous espérons précisèment appliquer une telle démarche par rapport à une matrice σ d'association et une classification hiérarchique telles que nous les avons ci-dessus développées.

Références

[1] Day W.H.E. and Mc Morris F.R. (1993) : Alignement, Comparison, and Consensus of Molecular Sequences : A Bibliography. *Proceed. of the International Federation of Classification Societies*, 1-4 Sept, Springer Verlag.

[2] Dayhoff M.O., Eck R.V. and Park C.M. (1972) : In *Atlas of Protein Sequence and Structure.* vol. 5 pp. 89-99.

[3] George D.G., Barker W.C. and Hunt L.T. (1990) : Mutation Data Matrix and Itg Uses. In *Methods in Enzymology,* vol. 183, Academic Press, pp. 313- 330.

[4] Gonnet G.H. (1992) : *A tutorial introduction to Computational Biochemistry using Darwin.* Research Report, Informatik E.T.H., Zurich, Switzerland, Nov. 24, 180 pages.

[5] Henikoff S. and Henikoff J.G. (1992) : Amino acid substitution matrices from protein blocks, *Proceed. Natl. Acad. Sci.* USA, vol. 89, pp. 10915- 10919, Nov. 1992, Biochemistry.

[6] Jones D.T., Taylor W.R. and Thornton (1992) : The rapid generation of mutation data matrices from protein sequences, *Cabios*, vol. 8, n^o3, pp. 275-282.

[7] Landès C. Hénault A. and Risler J.L. (1992) : A comparaison of several similarity indices used in the classification of protein sequences : a multivariate analysis. *Nucleic Acids Research*, 20, pp. 3631-3637.

[8] Lerman I.C. (1981) : *Classification et analyse ordinale des données.* Dunod, Paris, 760 pages.

[9] Lerman I.C. (1992) : Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles I et II : *Rev. Math. Infor. & Sci. Hum.*, 30è année ; I : n118, pp. 35-52, II : n119, pp. 75-100, Paris.

[10] Lerman I.C. (1993) : Likelihood linkage analysis (LLA) classification method : an example treated by hand. *Biochimie*, 75, Elsevier editions, pp. 379-397.

[11] Lerman I.C., Gras R. and Rostam H. (1981) : Elaboration et évaluation d'un indice d'implication pour des données binaires I et II : *Rev. Math. Infor. & Sci. Hum.*, 19è année ; I : n74, pp. 5-35, II : n 75, pp. 5-47, Paris.

[12] Lerman I.C., Nicolas J., Tallur B. and Peter Ph. (1993) : Classification of aligned biological sequences. *Proceed. of the International Federation of Classification Societies,* 1-4 Sept, Springer Verlag.

[13] Lerman I.C. and Peter Ph. (1985) : *Elaboration et logiciel d'un indice de similarité entre objets d'un type quelconque. Application à la recherche d'un consensus en classification.* Publ. Int. IRISA, n 262, Juillet 1985, 72 pages.

[14] Lerman I.C., Peter Ph and Leredde H. (1993) : Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hièrarchique par Analyse de la Vraisemblance des Liens) I et II : *La Revue de Modulad*, I : Déc. 1993, numéro 12, pp. 33-70, II : 1994, numéro 13, INRIA.

[15] Ouali-Allah M. (1991) : *Analyse en préordonnances, des données qualitatives. Applications aux données numériques et symboliques.* Thèse de doctorat de l'Université de Rennes I, 5 Déc. 1991.

[16] Risler J.L., Delorme M.O. Delacroix H. and Henaut A. (1988) : Amino acid substitutions in structurally related proteins : a pattern recognition approach. Determination of a new and efficient scoring matrix. *journal of Molecular Biology*, 204, pp. 1019-1029.

Figures, tableaux et annexes

```
-----------------------------------------
 HISTOGRAMME DES PREMIERES VALEURS PROPRES
-----------------------------------------

     VALEUR-PROPRE   POURCENTAGE   POURCENTAGE
                                    CUMULE
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
   1     0.20483740     28.47        28.47      **************************************************************************
   2     0.17707236     24.61        53.08      ************************************************************
   3     0.10339913     14.37        67.46      ************************************
   4     0.06425097      8.93        76.39      **********************
   5     0.04672891      6.50        82.88      ****************
   6     0.03825214      5.32        88.20      **************
   7     0.02699681      3.75        91.95      **********
   8     0.02123458      2.95        94.90      ********
   9     0.01578050      2.19        97.10      ******
  10     0.00684798      0.95        98.05      ***
  11     0.00484959      0.67        98.72      **
  12     0.00294542      0.41        99.13      **
  13     0.00266657      0.37        99.50      *
  14     0.00135550      0.19        99.69      *
  15     0.00095611      0.13        99.82      *

-------------------------------------------------------------------------------------------------------------------------
                      PLAN DE PROJECTION DES   20 POINTS SUR LES AXES  1 ET  2
                      ------------------------------------------------------
                           ( AXE  1 - HORIZONTAL   /   AXE  2 - VERTICAL )
-------------------------------------------------------------------------------------------------------------------------

    ** ATTENTION **   LES POINTS CI-DESSOUS ETAIENT A PLUS DE  2.3 ECARTS-TYPES DU CENTRE.
                      ILS ONT ETE RAMENES SUR LE CADRE DU GRAPHIQUE.
                 ---------------------------------------------
                 I     POINT     I   ABSCISSE   I   ORDONNEE   I
                 I    DEPLACE    I    REELLE    I    REELLE    I
                 ---------------------------------------------
                 I  W           I       3.714 I       0.889 I
                 ---------------------------------------------


    POINTS MULTIPLES
   ------------------
-------------------------------------------------------------------------------------------------------------------------
      I          I          I NBRE DE I
 POINT I ABSCISSE  I ORDONNEE  I POINTS  I                          POINTS CACHES
  VU   I APPROCHEE I APPROCHEE I CACHES  I
-------------------------------------------------------------------------------------------------------------------------
  P    I   -0.10   I    0.01   I    1   I T
-------------------------------------------------------------------------------------------------------------------------
```

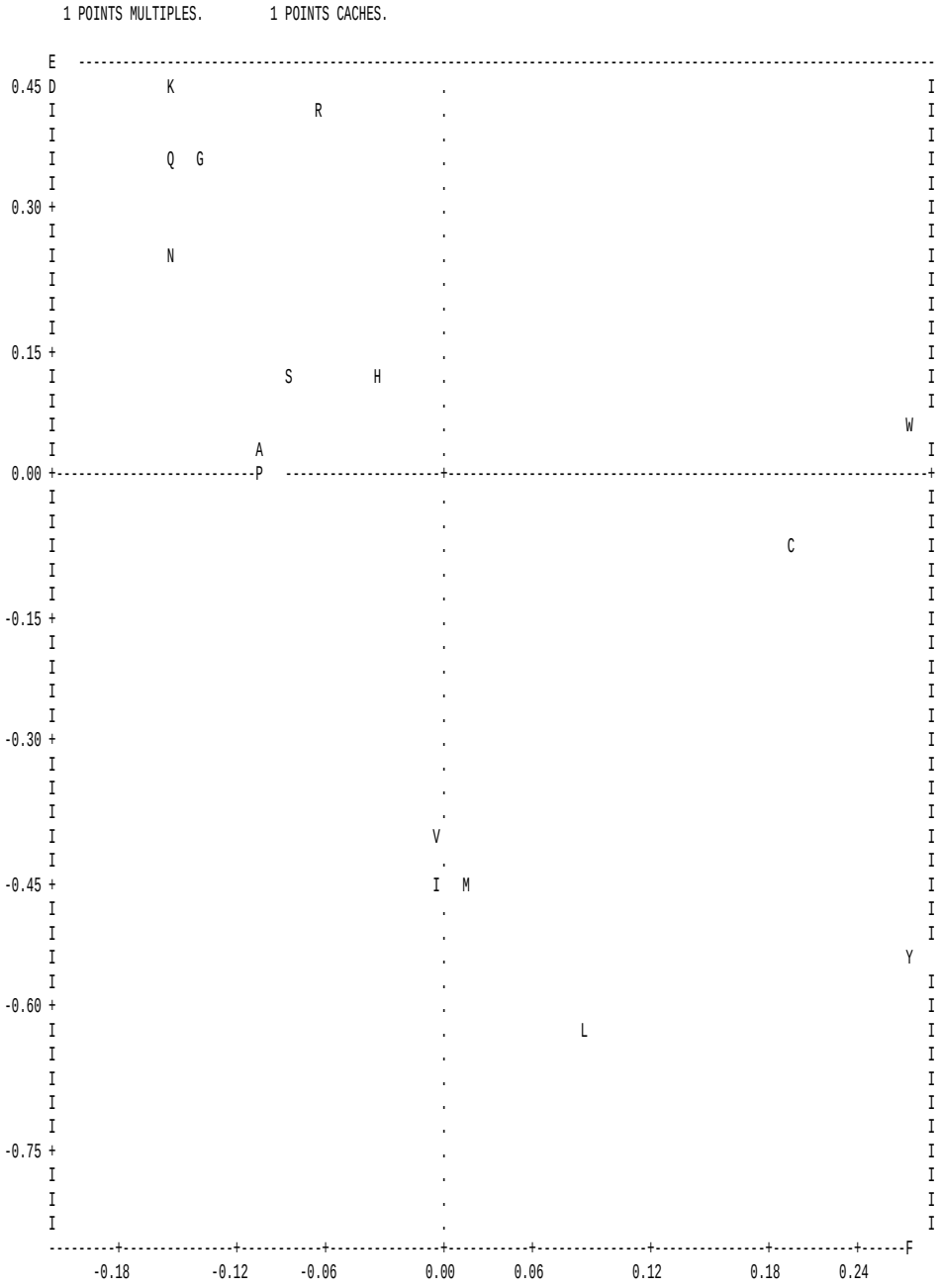


FIG. 1 - *Analyse des Correspondances du tableau des Oij*

HISTOGRAMME DES PREMIERES VALEURS PROPRES

	VALEUR-PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	0.33907497	15.24	15.24	**********
2	0.30107847	13.54	28.78	*********
3	0.25646183	11.53	40.31	********
4	0.23064519	10.37	50.68	*******
5	0.20066631	9.02	59.70	******
6	0.16753659	7.53	67.24	*****
7	0.13123132	5.90	73.14	****
8	0.10657902	4.79	77.93	****
9	0.09847586	4.43	82.36	****
10	0.07637283	3.43	85.79	***
11	0.05826006	2.62	88.41	***
12	0.05587424	2.51	90.92	***
13	0.05347573	2.40	93.33	***
14	0.04617719	2.08	95.40	**
15	0.03881841	1.75	97.15	**

PLAN DE PROJECTION DES 20 POINTS SUR LES AXES 1 ET 2

(AXE 1 - HORIZONTAL / AXE 2 - VERTICAL)

** ATTENTION ** LES POINTS CI-DESSOUS ETAIENT A PLUS DE 2.3 ECARTS-TYPES DU CENTRE.
ILS ONT ETE RAMENES SUR LE CADRE DU GRAPHIQUE.

POINT DEPLACE	ABSCISSE REELLE	ORDONNEE REELLE
W	3.683	2.528

POINTS MULTIPLES

POINT VU	ABSCISSE APPROCHEE	ORDONNEE APPROCHEE	NBRE DE POINTS CACHES	POINTS CACHES
S	-0.24	0.13	1	R
L	0.38	-0.72	1	I

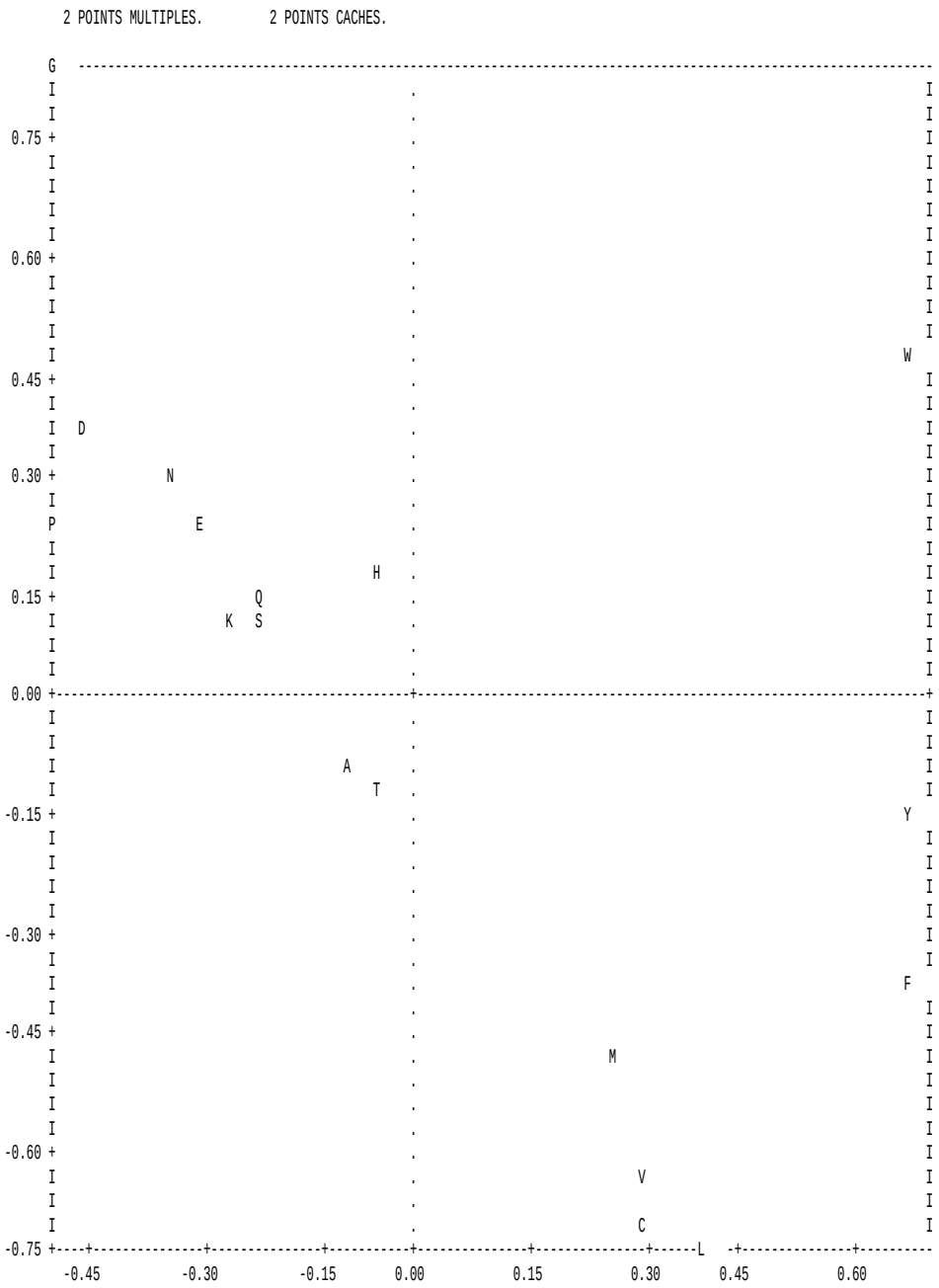


FIG. 2 - *Analyse des correspondances du tableau des O'ij*

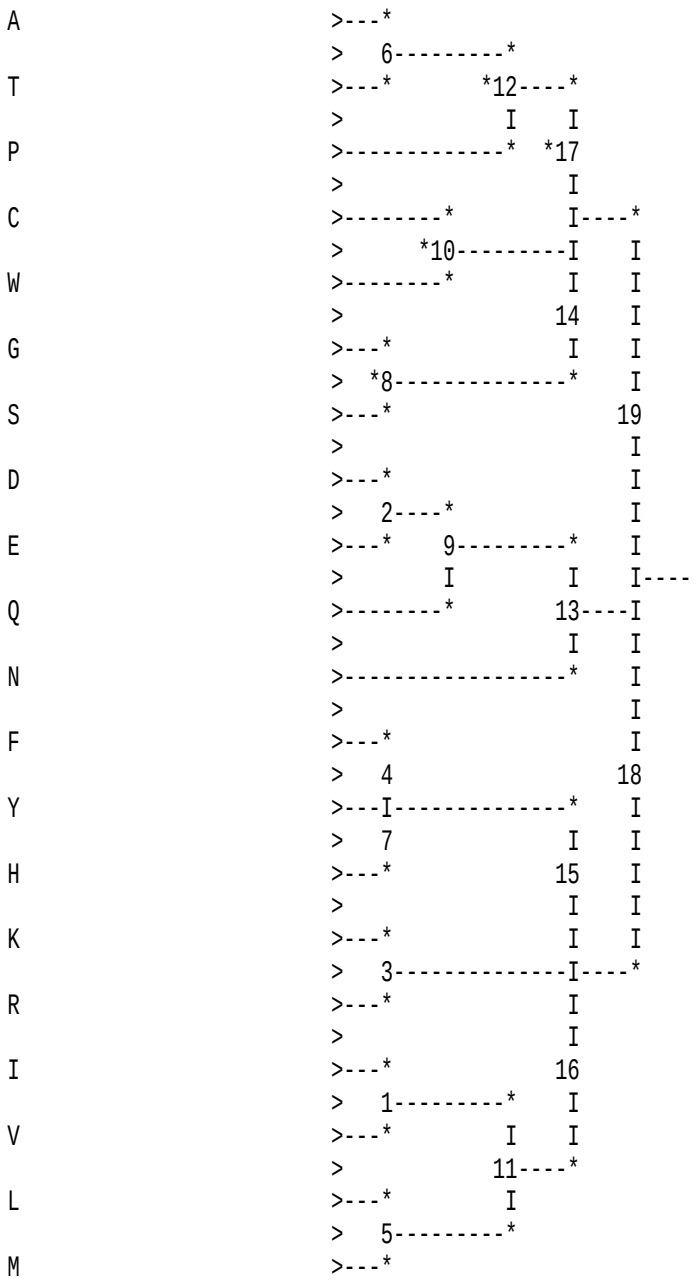


FIG. 3 - *Arbre des classifications AVL(0.5) relatif à QG(D)*

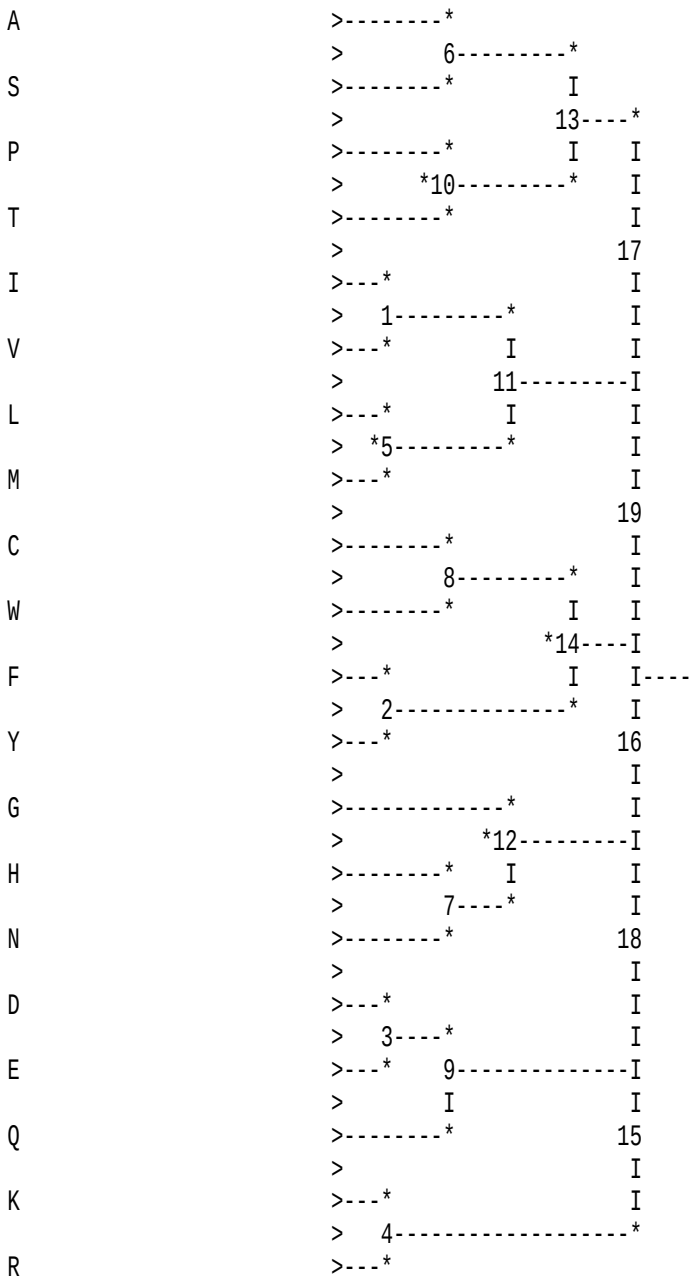
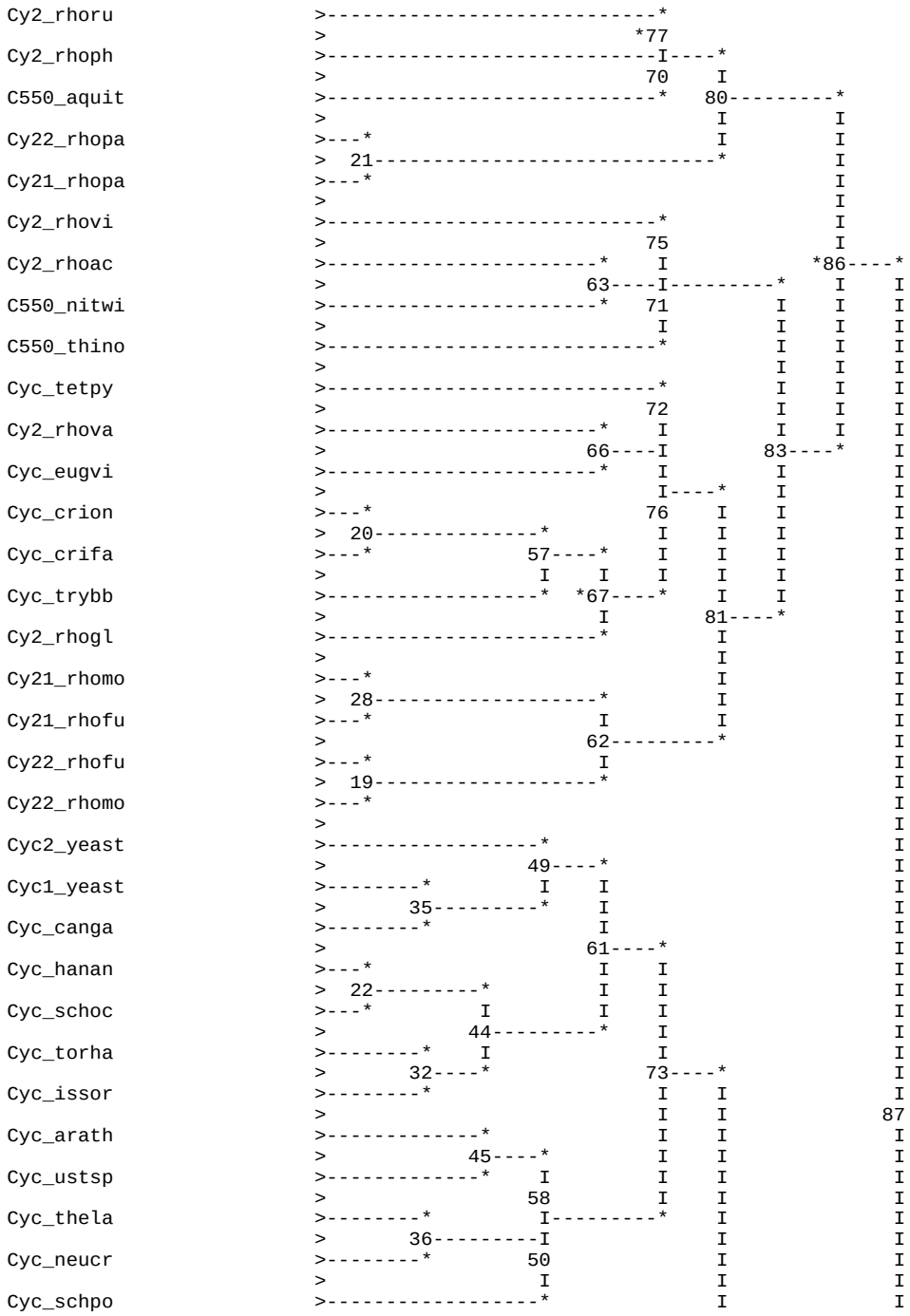
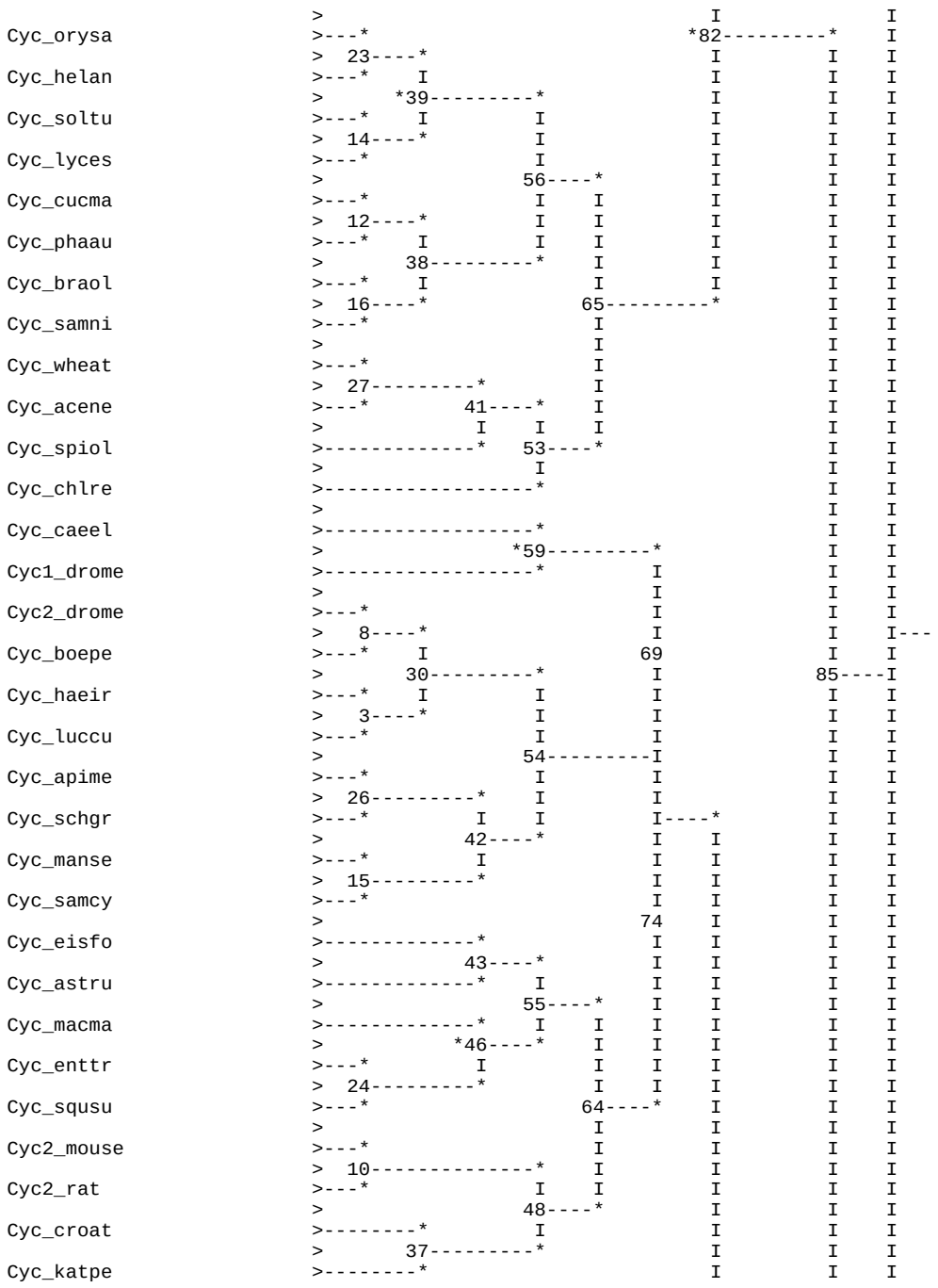


FIG. 4 - *Arbre des classifications AVL(0.5) relatif à QG(B)*

TAB. 1 - *Matrice NM(sigma(D)) sur* $\{(i,j)/21 \geq j \geq i \geq 1\}$

```
2.2785139
1.2574933 3.0485260
1.4461807 1.0511439 2.8774893
1.0902845 0.9095557 2.7170441 2.9560292
0.7150888 1.4552970 0.6127312 0.5225282 3.0453665
1.8570709 1.2616836 1.7922755 1.4450164 0.4210062 3.0091465
1.0631216 1.4570786 1.4536560 1.4528335 1.4547640 1.0715559 2.5683739
1.4460235 1.1587363 0.7603540 0.6872585 1.4506751 0.6052425 1.0143239 2.6716611
1.1000243 1.0221388 1.4482390 1.7846550 0.5489843 1.1076385 1.6645865 0.7086759
2.9363647
1.0048492 0.9048023 0.2785048 0.1478879 2.1351538 0.0000000 1.0255015 2.2156837
0.4351146 3.0401185
1.2370453 1.2620064 1.0085783 0.9630430 1.4545969 0.9122524 1.2468621 2.1428313
1.1082051 2.3037999 2.5975044
1.4476792 1.3108960 2.0106878 1.7405829 0.9218428 1.4479743 1.6361873 1.0048938
1.7342013 0.6063662 1.1630094 2.2519710
1.7953781 1.1648540 0.9655890 0.9153895 0.8685930 1.1341295 1.4537466 0.9603977
0.9300714 1.4449174 1.1346551 1.2159203 2.9969170
1.1637790 1.1003690 1.7128764 2.0415778 0.8171082 1.1699888 1.8362802 0.8474665
2.0294731 0.8631580 1.1704556 1.4517292 1.4507571 2.7392020
1.1271187 1.2969317 1.1900305 1.4480207 0.7333012 1.4466629 1.8750945 0.7679510
2.6944678 0.5182908 1.1346551 1.4504694 1.1927974 1.9950739 2.8674059
1.8447546 1.6662111 1.4470628 1.1114709 0.9546423 1.8379109 1.2542595 1.1395745
1.1206255 0.6603217 1.2495892 1.7548249 1.7789348 1.1805944 1.1461017 2.2118475
2.1980586 1.2839677 1.1682619 1.1395298 0.9961412 1.1076385 1.2705377 1.7620621
1.1479177 1.0660927 1.4528675 1.7342013 1.7567459 1.2028949 1.1712632 1.8009442
2.1229463
1.8371065 1.1212186 0.8901660 0.8316810 1.4492844 0.7662575 0.9575734 2.7432480
0.6087465 2.2825398 1.9342194 0.9468505 1.1534717 0.7675209 0.6766317 1.0995551
1.4464102 2.8104279
0.9503884 1.5566432 0.9884725 0.9407687 1.3409066 1.1623911 1.2413585 1.0443201
1.0989610 1.1276033 1.2363904 1.0348313 1.0528978 1.1629117 1.4557014 1.0671469
1.0193918 1.0764722 3.0506892
0.8029671 1.7753067 1.0762570 0.7446253 2.6495769 0.6692612 2.2285907 1.0722921
0.8934274 1.1762559 1.2061894 1.2691501 0.9371850 1.1223230 1.0802619 1.2150354
0.8934274 0.8570857 1.4571486 3.0472760
1.3585238 0.9379746 1.3880049 1.2226409 1.0659977 0.9782194 1.3493228 1.4569216
1.2125098 1.0078677 1.4522192 1.3398169 1.0926322 1.2838297 1.2843280 1.2411964
1.3399767 1.4014330 0.6181077 1.0512260 2.7385223
```


Cy2_rhoru
Cy2_rhoph
C550_aquit
Cy22_rhopa
Cy21_rhopa
Cy2_rhovi
Cy2_rhoac
C550_nitwi
C550_thino
Cyc_tetpy
Cy2_rhova
Cyc_eugvi
Cyc_crion
Cyc_crifa
Cyc_trybb
Cy2_rhogl
Cy21_rhomo
Cy21_rhofu
Cy22_rhofu
Cy22_rhomo
Cyc2_yeast
Cyc1_yeast
Cyc_canga
Cyc_hanan
Cyc_schoc
Cyc_torha
Cyc_issor
Cyc_arath
Cyc_ustsp
Cyc_thela
Cyc_neucr
Cyc_schpo
77
70
80
21
75
63
71
72
66
83
76
20
57
67
81
28
62
49
35
61
22
44
32
73
87
45
58
36
50


Cyc_orysa
Cyc_helan
Cyc_soltu
Cyc_lyces
Cyc_cucma
Cyc_phaau
Cyc_braol
Cyc_samni
Cyc_wheat
Cyc_acene
Cyc_spiol
Cyc_chlre
Cyc_caeel
Cyc1_drome
Cyc2_drome
Cyc_boepe
Cyc_haeir
Cyc_luccu
Cyc_apime
Cyc_schgr
Cyc_manse
Cyc_samcy
Cyc_eisfo
Cyc_astru
Cyc_macma
Cyc_enttr
Cyc_squsu
Cyc2_mouse
Cyc2_rat
Cyc_croat
Cyc_katpe

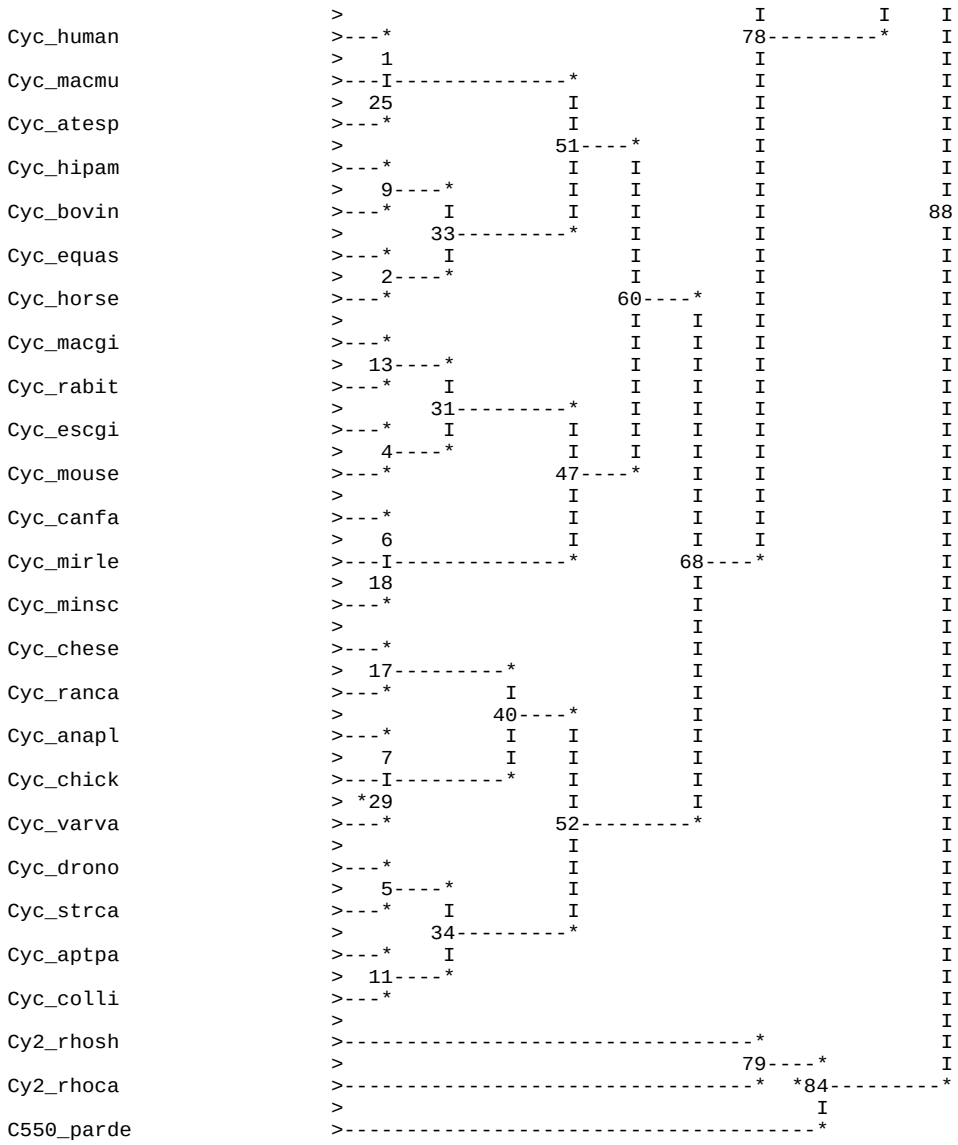


FIG. 5 - *AVL(0.5)(NM(sigma(D)))* sur l'ensemble des 89 sequences

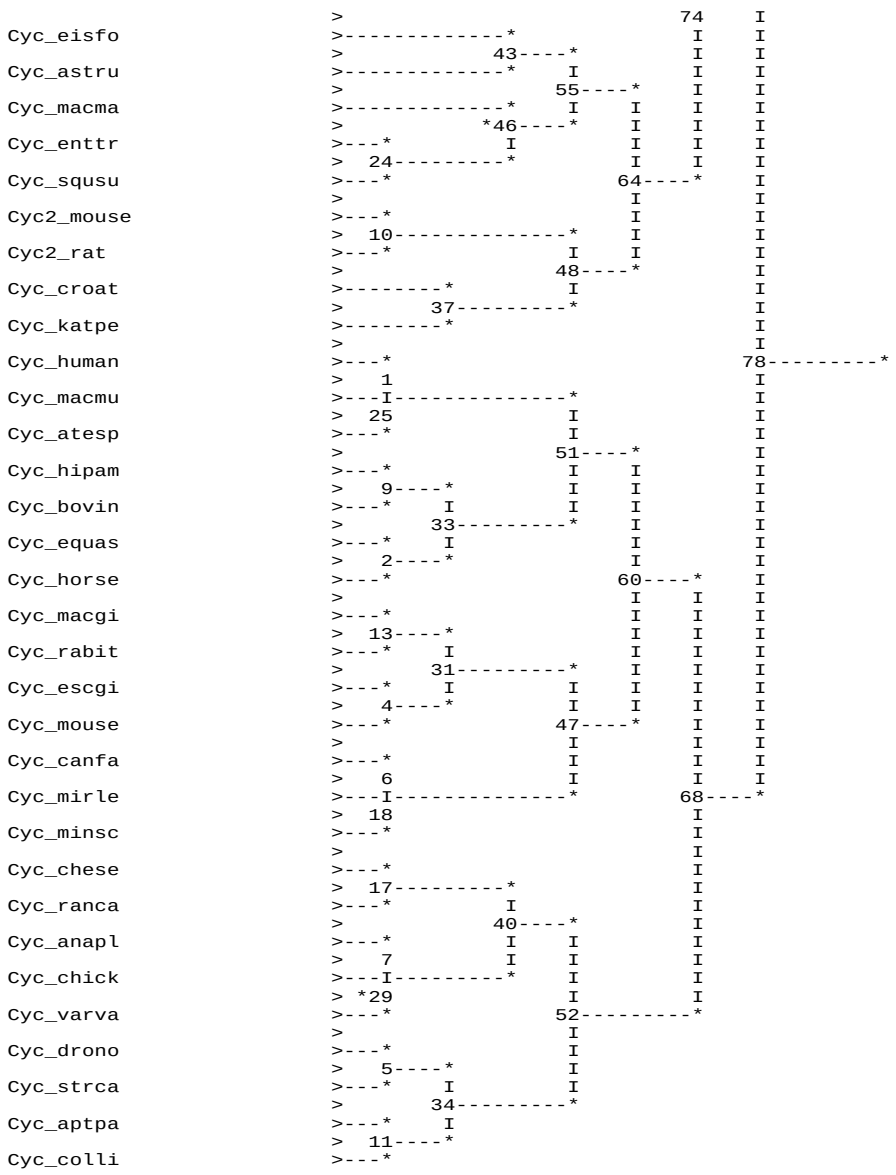


FIG. 6 - *AVL(0.5)(NM(sigma(D))) sur le sous ensemble formé des mamifères et des oiseaux*

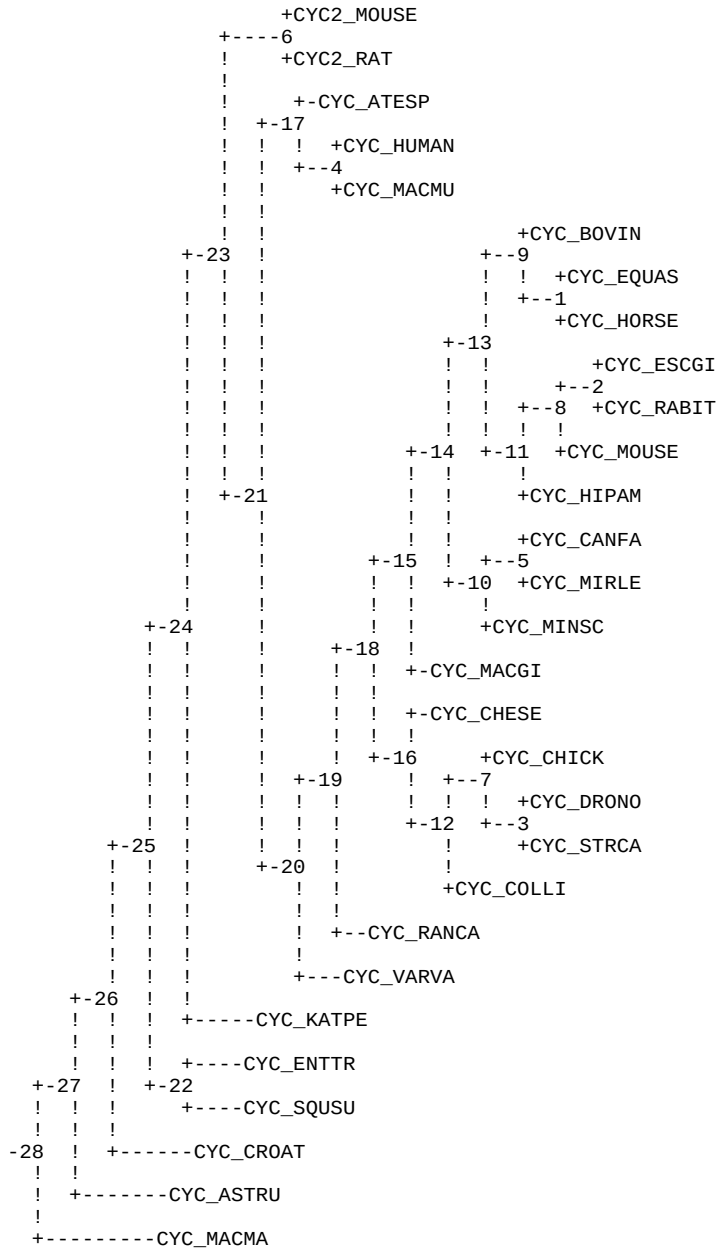


FIG. 7 - *La methode UPGMA extraite de PHYLIP*

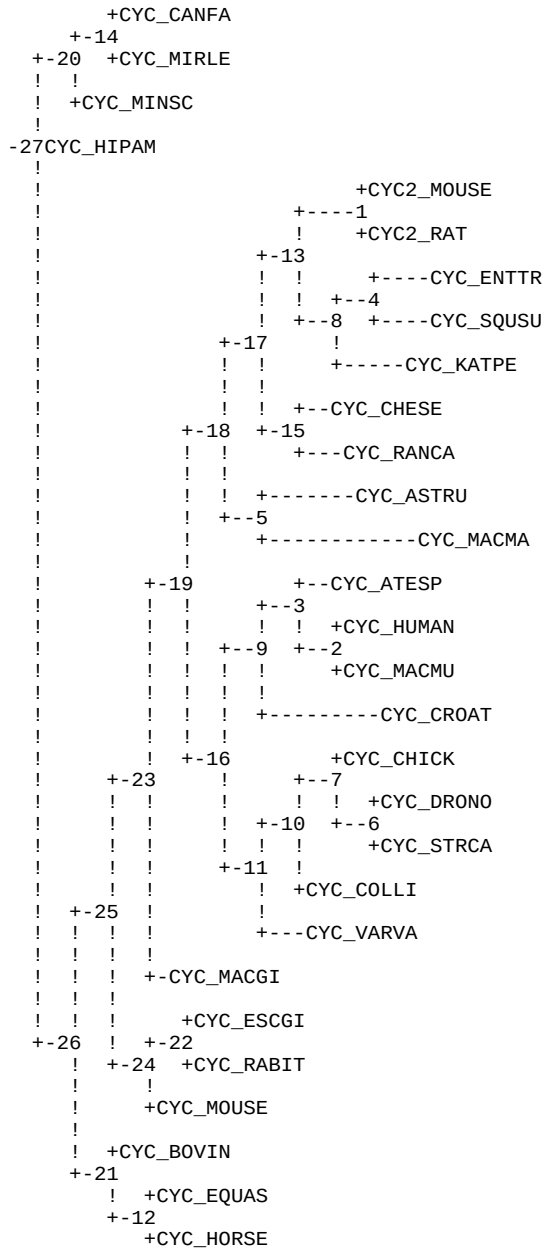


FIG. 8 - *La methode Neighbor Joining extraite de PHYLIP*

ANNEXE

Elements du programme CHAVL dans le guidage statiatique de l'interpretation de l'arbre [10],[14], sur l'ensemble des 89 sequences.

*** ETAPE INTRP ***

RANGEMENT DES ELEMENTS PAR VALEURS DE DISPERSIONS CROISSANTES
****

ELEMENT	1 : Cy2_rhoru	DISPERSION :	0.08040
ELEMENT	4 : Cy2_rhoca	DISPERSION :	0.08326
ELEMENT	2 : Cy2_rhosh	DISPERSION :	0.08813
ELEMENT	84 : C550_aquit	DISPERSION :	0.12099
ELEMENT	3 : C550_parde	DISPERSION :	0.14554
ELEMENT	46 : Cy2_rhoph	DISPERSION :	0.16780
ELEMENT	5 : Cy2_rhovi	DISPERSION :	0.20099
ELEMENT	83 : Cy21_rhopa	DISPERSION :	0.21136
ELEMENT	82 : Cy22_rhopa	DISPERSION :	0.21930
ELEMENT	6 : Cyc_tetpy	DISPERSION :	0.27221
ELEMENT	81 : C550_nitwi	DISPERSION :	0.29451
ELEMENT	86 : C550_thino	DISPERSION :	0.29638
ELEMENT	87 : Cy22_rhofu	DISPERSION :	0.30260
ELEMENT	89 : Cy22_rhomo	DISPERSION :	0.30335
ELEMENT	31 : Cy2_rhogl	DISPERSION :	0.30821
ELEMENT	88 : Cy21_rhofu	DISPERSION :	0.31530
ELEMENT	85 : Cy21_rhomo	DISPERSION :	0.31923
ELEMENT	32 : Cyc_trybb	DISPERSION :	0.32584
ELEMENT	45 : Cy2_rhoac	DISPERSION :	0.33315
ELEMENT	47 : Cy2_rhova	DISPERSION :	0.36176
ELEMENT	7 : Cyc_crion	DISPERSION :	0.36471
ELEMENT	9 : Cyc_crifa	DISPERSION :	0.37716
ELEMENT	79 : Cyc_eugvi	DISPERSION :	0.42487
ELEMENT	28 : Cyc_caeel	DISPERSION :	0.43993
ELEMENT	44 : Cyc1_drome	DISPERSION :	0.45221
ELEMENT	8 : Cyc2_yeast	DISPERSION :	0.48069
ELEMENT	26 : Cyc_torha	DISPERSION :	0.48781
ELEMENT	33 : Cyc_ustsp	DISPERSION :	0.49913
ELEMENT	10 : Cyc_arath	DISPERSION :	0.49929
ELEMENT	76 : Cyc_canga	DISPERSION :	0.50187
ELEMENT	11 : Cyc_thela	DISPERSION :	0.50594
ELEMENT	34 : Cyc_neucr	DISPERSION :	0.50627
ELEMENT	25 : Cyc_schoc	DISPERSION :	0.51218
ELEMENT	35 : Cyc_schpo	DISPERSION :	0.51250
ELEMENT	27 : Cyc_issor	DISPERSION :	0.51868
ELEMENT	23 : Cyc_chlre	DISPERSION :	0.52133
ELEMENT	29 : Cyc1_yeast	DISPERSION :	0.52230
ELEMENT	24 : Cyc_hanan	DISPERSION :	0.52459
ELEMENT	21 : Cyc_helan	DISPERSION :	0.56103
ELEMENT	22 : Cyc_spiol	DISPERSION :	0.56251
ELEMENT	30 : Cyc_eisfo	DISPERSION :	0.57513
ELEMENT	20 : Cyc_acene	DISPERSION :	0.57991
ELEMENT	18 : Cyc_wheat	DISPERSION :	0.58131

```
ELEMENT   19 : Cyc_samni              DISPERSION :    0.58342
ELEMENT   16 : Cyc_phaau              DISPERSION :    0.59625
ELEMENT   14 : Cyc_cucma              DISPERSION :    0.59974
ELEMENT   12 : Cyc_orysa              DISPERSION :    0.60144
ELEMENT   17 : Cyc_braol              DISPERSION :    0.60396
ELEMENT   13 : Cyc_soltu              DISPERSION :    0.60402
ELEMENT   15 : Cyc_lyces              DISPERSION :    0.60734
ELEMENT   74 : Cyc_macma              DISPERSION :    0.62825
ELEMENT   39 : Cyc_apime              DISPERSION :    0.62933
ELEMENT   42 : Cyc_samcy              DISPERSION :    0.65765
ELEMENT   73 : Cyc_katpe              DISPERSION :    0.65901
ELEMENT   70 : Cyc_croat              DISPERSION :    0.66278
ELEMENT   78 : Cyc_astru              DISPERSION :    0.66287
ELEMENT   36 : Cyc2_drome             DISPERSION :    0.66298
ELEMENT   37 : Cyc_haeir              DISPERSION :    0.67558
ELEMENT   41 : Cyc_manse              DISPERSION :    0.67619
ELEMENT   38 : Cyc_luccu              DISPERSION :    0.67930
ELEMENT   43 : Cyc_schgr              DISPERSION :    0.68195
ELEMENT   40 : Cyc_boepe              DISPERSION :    0.68301
ELEMENT   80 : Cyc_squsu              DISPERSION :    0.69927
ELEMENT   77 : Cyc_enttr              DISPERSION :    0.70482
ELEMENT   49 : Cyc2_rat               DISPERSION :    0.70886
ELEMENT   48 : Cyc2_mouse             DISPERSION :    0.71214
ELEMENT   75 : Cyc_ranca              DISPERSION :    0.71409
ELEMENT   72 : Cyc_atesp              DISPERSION :    0.72225
ELEMENT   71 : Cyc_varva              DISPERSION :    0.73669
ELEMENT   50 : Cyc_human              DISPERSION :    0.73746
ELEMENT   53 : Cyc_macmu              DISPERSION :    0.73778
ELEMENT   52 : Cyc_chese              DISPERSION :    0.74632
ELEMENT   67 : Cyc_aptpa              DISPERSION :    0.75191
ELEMENT   51 : Cyc_macgi              DISPERSION :    0.75404
ELEMENT   66 : Cyc_horse              DISPERSION :    0.75645
ELEMENT   56 : Cyc_minsc              DISPERSION :    0.75667
ELEMENT   69 : Cyc_colli              DISPERSION :    0.75811
ELEMENT   58 : Cyc_mirle              DISPERSION :    0.76258
ELEMENT   65 : Cyc_anapl              DISPERSION :    0.76331
ELEMENT   61 : Cyc_equas              DISPERSION :    0.76417
ELEMENT   60 : Cyc_strca              DISPERSION :    0.76513
ELEMENT   54 : Cyc_drono              DISPERSION :    0.76601
ELEMENT   68 : Cyc_chick              DISPERSION :    0.76855
ELEMENT   62 : Cyc_rabit              DISPERSION :    0.76908
ELEMENT   55 : Cyc_canfa              DISPERSION :    0.77474
ELEMENT   57 : Cyc_hipam              DISPERSION :    0.77505
ELEMENT   64 : Cyc_mouse              DISPERSION :    0.77930
ELEMENT   59 : Cyc_escgi              DISPERSION :    0.77991
ELEMENT   63 : Cyc_bovin              DISPERSION :    0.78167
```

STATISTIQUES DES NIVEAUX
****

	NIVEAU	STATISTIQUE GLOBALE	STATISTIQUE LOCALE
	1	1.4324	1.4324
	2	2.0456	0.6131
	3	2.5317	0.4861
	4	2.9517	0.4200
	5	3.3281	0.3763
	6	3.6792	0.3511
	7	4.0135	0.3343
	8	4.3350	0.3216
	9	4.6450	0.3100
	10	4.9497	0.3047
	11	5.2411	0.2914
	12	5.5233	0.2821
	13	5.8031	0.2798
	14	6.0848	0.2817
1 MAXIMUM	15	6.3666	0.2818
	16	6.6347	0.2681
	17	6.8976	0.2629
	18	6.7714	-0.1262
	19	7.0353	0.2639
	20	7.3039	0.2686
	21	7.5800	0.2761
	22	7.8643	0.2843
	23	8.1584	0.2940
2 MAXIMUM	24	8.4560	0.2977
	25	8.4240	-0.0320
	26	8.7168	0.2928
	27	9.0188	0.3020
3 MAXIMUM	28	9.3292	0.3104
	29	9.3395	0.0104
	30	8.6794	-0.6601
	31	8.4896	-0.1899
4 MAXIMUM	32	8.7300	0.2404
	33	8.7093	-0.0207
	34	8.8279	0.1186
	35	9.0547	0.2268
	36	9.2845	0.2298
5 MAXIMUM	37	9.5221	0.2376
	38	9.6543	0.1322
6 MAXIMUM	39	9.8117	0.1573
	40	9.6027	-0.2090
	41	9.8811	0.2784
7 MAXIMUM	42	10.1784	0.2972
	43	10.4025	0.2241
8 MAXIMUM	44	10.7034	0.3009
	45	10.9387	0.2353
9 MAXIMUM	46	11.2650	0.3262
	47	10.0490	-1.2160
10 MAXIMUM	48	10.4214	0.3725
	49	10.6920	0.2706
11 MAXIMUM	50	10.9807	0.2886
	51	10.5905	-0.3901

	52	9.7040	-0.8865
12 MAXIMUM	53	9.9336	0.2296
	54	10.1326	0.1990
	55	10.5087	0.3761
13 MAXIMUM	56	10.9210	0.4123
	57	11.1233	0.2023
14 MAXIMUM	58	11.5832	0.4599
	59	11.7065	0.1233
	60	9.8805	-1.8260
15 MAXIMUM	61	10.3649	0.4845
	62	10.6181	0.2531
	63	10.6882	0.0702
16 MAXIMUM	64	11.3233	0.6350
	65	11.7573	0.4340
	66	11.8361	0.0788
17 MAXIMUM	67	12.0318	0.1957
	68	9.4479	-2.5839
18 MAXIMUM	69	9.7884	0.3405
	70	9.8109	0.0225
19 MAXIMUM	71	9.8594	0.0485
	72	9.9033	0.0439
	73	10.6695	0.7661
20 MAXIMUM	74	11.5350	0.8655
	75	11.6428	0.1078
21 MAXIMUM	76	12.0551	0.4122
	77	12.1291	0.0740
	78	8.7918	-3.3373
	79	8.7923	0.0005
	80	8.8066	0.0143
	81	9.0000	0.1934
22 MAXIMUM	82	10.5795	1.5796
	83	11.1972	0.6177
	84	11.2072	0.0100
	85	8.9350	-2.2721
23 MAXIMUM	86	9.3881	0.4530
	87	7.0501	-2.3380
	88	0.9063	-6.1438

--- FIN DE L'ETAPE INTRP ---