Chapitre 3

Exploration de la notion de ressemblance

La lune? Ce n'est plus la même,
Le printemps? Ce n'est plus
Le printemps d'autrefois.
Moi seul
N'est pas changé.

Ise monogatari, 4. Ariwara no Narihira.
Anthologie de la poésie japonaise classique.
G. Renondeau (Ed. et trad.). Gallimard, 1971.

Ce chapitre réalise une exploration des diverses façons de définir et de mesurer la ressemblance d'un ensemble d'objets. Rappelons que dans notre cas, ceux-ci sont soit des chaînes de caractères sur un certain alphabet, soit des mots présents dans ces chaînes. Par ailleurs, ces chaînes représentent le plus souvent des séquences biologiques, c'est-à-dire des structures primaires de macromolécules, mais peuvent aussi être un codage linéaire de leur structure tridimensionnelle. La notion de ressemblance dont nous allons parler et les concepts que nous allons introduire pour la définir et la mesurer font ainsi référence à certains objets biologiques bien précis, mais il est important d'observer qu'ils sont généralisables à d'autres types d'objets dans d'autres domaines. En fait, tout ce que nous allons présenter ici et dans les chapitres suivants peut s'appliquer à des chaînes dénotant n'importe quel objet décomposable en unités élémentaires sur lequel un ordre de succession a pu être défini. Typiquement un texte est un tel objet, l'ordre en question porte sur les lettres qui le composent et est alors spatial : les lettres se suivent dans le texte. Mais l'ordre peut aussi être temporel et l'objet par exemple une suite d'événements qui se succèdent dans le temps.

Ce chapitre se divise en quatre parties. La première s'attache à fournir une idée intuitive de comment évaluer la ressemblance d'objets lorsqu'une contrainte d'ordre linéaire joue sur les éléments les composant. Nous introduisons ainsi le concept d'un alignement de deux chaînes permettant d'apparier un certain nombre d'éléments de ces objets. Dans le cas des macromolécules, ces éléments sont des nucléotides ou des acides aminés et nous sommes ainsi amenés à examiner les critères de ressemblance portant sur ces monomères, en tant que symboles ou en tant qu'objets 3D. Dans cette première partie, la comparaison porte toujours uniquement sur deux objets à la fois et ceux-ci sont comparés directement entre eux. Il a d'autres façons de procéder et la seconde partie est consacrée à une présentation générale de diverses méthodes de comparaison, globale ou locale, deux à deux ou multiple, directe ou indirecte. Nous montrons en particulier tout l'intérêt, dans certains cas, d'une comparaison de plus de deux objets simultanément et ce travail porte essentiellement sur les comparaisons multiples. Dans ce but, la distinction entre méthodes directes et méthodes indirectes devient particulièrement importante (bien que des liens existent entre les deux comme nous l'indiquons). Les deux dernières parties de ce chapitre sont donc réservées à ces méthodes, les directes d'abord puis les indirectes, et à leurs façons respectives de représenter des ensembles d'objets similaires. Ces représentations constituent d'une certaine manière un résumé plus ou moins compact de l'information contenue dans les ensembles. Nous montrons enfin que dans le cas des mesures indirectes de la ressemblance, ces représentations servent à l'identification même de ces ensembles.

Observons pour conclure que la division entre mesures directe et indirecte reflète la séparation entre le travail réalisé par d'autres avant nous et le notre.

3.1 Définition de base et conventions de notation

Avant toute chose, il convient d'établir une définition plus formelle des éléments avec lesquels nous allons travailler et d'introduire certaines conventions qui sont utilisées dans tout le reste du travail.

Définition 3.1.1 *Étant donné un alphabet Σ et une chaîne s, c'est-à-dire un élément de Σ^*, nous appelons un mot u dans s un élément de Σ^k pour $k \geq 1$ tel que $s = xuy$ avec $x, y \in \Sigma^*$. Nous disons dans ce cas également que u est présent dans s. Nous désignons par $\mid u \mid = k$ la longueur de u et par λ le mot vide (par convention $\mid \lambda \mid = 0$).*

Définition 3.1.2 *Étant donné un mot u dans une chaîne s tel que $s = xuy$, la position de u dans s est égale à $\mid x \mid + 1$ (une chaîne est ainsi indexée de 1 jusqu'à $\mid s \mid$). Nous admettons que la position d'un mot dans une chaîne est totalement associée au mot et qu'un mot u dont on connaît la longueur est ainsi désigné de manière non ambiguë par cette position.*

Bien entendu, un mot u dans une chaîne s est lui-même une chaîne — notons-la s_u (voir la figure 3.1). Nous disons alors que u est une occurrence de la chaîne s_u dans une autre chaîne, ici s. Observons que la chaîne s_u n'est pas associée à une position dans une chaîne comme l'est le mot u bien que les deux objets soient identiques en tant qu'éléments de Σ^k où $k = \mid u \mid$.

Pour simplifier, et lorsqu'il n'y aura pas danger de confusion, nous confondrons parfois ces deux entités, mot u et chaîne s_u associée à u, et désignerons les deux par le terme mot.

Finalement, dans tout ce qui suit :

- $s1$,$s2$,...,sN représente N chaînes;
- $u1$,$u2$,...,uM indique M mots;
- $s_1, s_2, ..., s_n$ et $u_1, u_2, ..., u_m$ respectivement n et m symboles dans une chaîne s ou un mot u.

Les nucléotides et les acides aminés sont désignés par des lettres majuscules mais quand une définition, un exemple ou un algorithme s'applique aux deux, des lettres minuscules sont employées, éventuellement indexées. Les lettres grecques sont réservées pour le codage en coordonnées internes des structures de protéines.

s: $a\ g\ \boxed{a\ t\ g\ t\ g\ g}\ a\ g$ ... chaîne associée à u: $s_u = atgtgg$

mot u occurrence de la chaîne $atgtgg$ à la position 3 dans s

Figure 3.1: Illustration des définitions de base.

3.2 De la ressemblance des colliers de perles

3.2.1 Alignement de colliers

Une jolie image utilisée à propos des molécules biologiques est celle d'un collier de perles de couleurs différentes, les perles étant les nucléotides (4 couleurs) ou les acides aminés (20 couleurs) composant la molécule et le fil du collier les liaisons chimiques reliant un des atomes d'un monomère à un des atomes du monomère suivant le long de la chaîne nucléique ou polypeptidique.

Cette image suggère tout de suite une façon de comparer deux de ces molécules. Elle consiste à les placer allongées l'une en dessous de l'autre et à examiner chaque paire de perles mises ainsi face à face pour voir lesquelles sont identiques (ont la même couleur) et lesquelles diffèrent. Bien sûr les facteurs de variabilité étudiés au chapitre précédent font que les choses sont plus complexes. Dans le cas le plus simple correspondant aux mutations ponctuelles, certaines perles d'un collier peuvent avoir été remplacées par des perles d'autres couleurs, ou encore perdues ou rajoutées. Pour comparer les colliers, il faut donc parfois étendre le fil localement entre deux perles sur un des colliers et ne rien mettre face à une perle de l'autre. Il faut aussi considérer de déplacer un des colliers par rapport à l'autre (si ce sont les perles d'une des extrémités qui ont été égarées).

Ce processus qui consiste à placer deux objets ordonnés linéairement l'un en dessous de l'autre et à mettre certains éléments d'un objet face à certains éléments de l'autre est ce que l'on appelle réaliser un alignement. La ressemblance entre les deux objets peut alors être mesurée tout simplement en comptant le nombre de paires d'éléments mis face à face qui sont identiques, ou inversement le nombre d'éléments d'un de ces objets qui est mis face soit à un élément de l'autre ayant une nature différente soit à rien du tout. On appelle ce 'rien du tout' un trou, qui correspond ainsi pour nous exclusivement à ce qui est appelé en anglais un 'gap'. Cette terminologie est un peu trompeuse puisqu'un trou donne toujours l'idée de l'absence de quelque chose. S'il est vrai que par rapport à un alignement une insertion d'un élément dans un objet n'est qu'une délétion d'un élément dans celui qui lui fait face, en termes d'évolution les deux ne sont pas symétriques. Pour simplifier toutefois, et parce que ce que nous voyons n'est après tout que l'alignement, nous employons le terme de trou pour désigner les deux phénomènes simultanément. Il faut noter également qu'il existe en fait deux mots en anglais qui font référence à des délétions ou insertions d'éléments dans un alignement : 'indel' et 'gap'. Le premier désigne une seule délétion ou insertion tandis que le second dénote une suite contiguë de délétions. Dans l'usage courant cependant, le terme de 'gap' est utilisé dans les deux situations. Cela est le cas également pour nous en ce qui concerne le mot 'trou' qui fait ainsi référence à une ou plusieurs délétions ou insertions dans un alignement.

Il y a bien sûr un très grand nombre de possibilités pour aligner un objet par rapport à un autre, et celle donnant lieu au plus grand nombre de paires alignées qui sont identiques (ou au plus petit nombre qui sont différentes) est alors une façon de caractériser le degré de ressemblance des deux objets. C'est la façon la plus classiquement utilisée en informatique lorsque l'on veut comparer deux chaînes de caractères puisque la distance, appelée distance d'édition, entre ces chaînes correspond au plus petit nombre d'opérations de substitutions, délétions et insertions nécessaires pour passer d'une chaîne à l'autre.

Ce n'est pas la seule manière de procéder, nous en verrons d'autres. Nous reprendrons aussi de manière plus formelle celle que nous venons de présenter. L'important pour le moment est qu'à la base de toutes les autres façons de mesurer cette ressemblance d'objets totalement ordonnés se trouve toujours, implicitement ou explicitement, un alignement de ces objets.

Une question que nous pouvons nous poser par rapport à la caractérisation de la ressemblance présentée plus haut concerne la valeur attribuée à la substitution d'un monomère par un autre : elle est en effet la même quels que soient les monomères en question. Deux d'entre eux sont ainsi toujours soit identiques, soit différents. Or nous avons vu que ces monomères sont en fait des entités complexes pouvant posséder plusieurs attributs dont certains en commun. La prochaine section est donc dédiée à une exploration de la ressemblance au niveau des perles du collier — les nucléotides et les acides aminés — et aux diverses façons d'utiliser les mesures de similarité que l'on a pu établir entre eux.

3.2.2 De la ressemblance des perles

3.2.2.1 Les nucléotides

Dans le chapitre 2, nous avons indiqué que les nucléotides sont composés de trois éléments : une base azotée, un phosphate et un sucre (ribose ou désoxyribose suivant qu'il s'agit d'ARN ou d'ADN). C'est la base azotée qui permet de différencier les nucléotides d'un même type d'acide nucléique.

Ces bases possèdent cependant certaines propriétés chimiques et structurales qui rendent possible de les grouper par classes. Nous avons mentionné un de ces groupements, composé de la classe des purines (Adénine et Guanine) et de celle des pyrimidines (Cytosine et Thymine ou Cytosine et Uracile) distinguables entre elles par le nombre de cycles que comportent les bases (respectivement 2 et 1) (voir la figure 3.2).

La capacité de chacune de ces bases à former un nombre différent de liaisons hydrogènes avec une autre permet une seconde division de celles-ci en deux nouvelles classes correspondant aux bases complémentaires Adénine-Thymine (ou Uracile) et Cytosine-Guanine, la première paire pouvant en effet établir deux liaisons et la seconde trois.

Malgré ces quelques traits en communs, il est usuel de considérer que le rôle joué par chacune de ces bases dans une chaîne nucléique lui est tout à fait spécifique et qu'une base ne ressemble ainsi qu'à elle-même. Il ne faut pas en déduire de cela cependant que les modifications pouvant survenir dans une molécule d'ADN ou d'ARN ont toutes les mêmes conséquences. Prenons le cas d'une partie de l'ADN correspondant à un gène et donc codant pour une protéine. Nous avons vu dans le chapitre précédent que la dégénérescence du code fait que certaines substitutions d'un nucléotide le long de ce gène ne changent pas le produit final, c'est-à-dire la protéine (bien que ces substitutions puissent avoir une influence plus subtile sur l'efficacité de la traduction [Lewin, 1994]). C'est ce qui a été appelé une mutation silencieuse. Une comparaison de gènes codant pour la même protéine peut ainsi amener à considérer comme semblables deux nucléotides différents mais situés à une même position relative le long du gène.

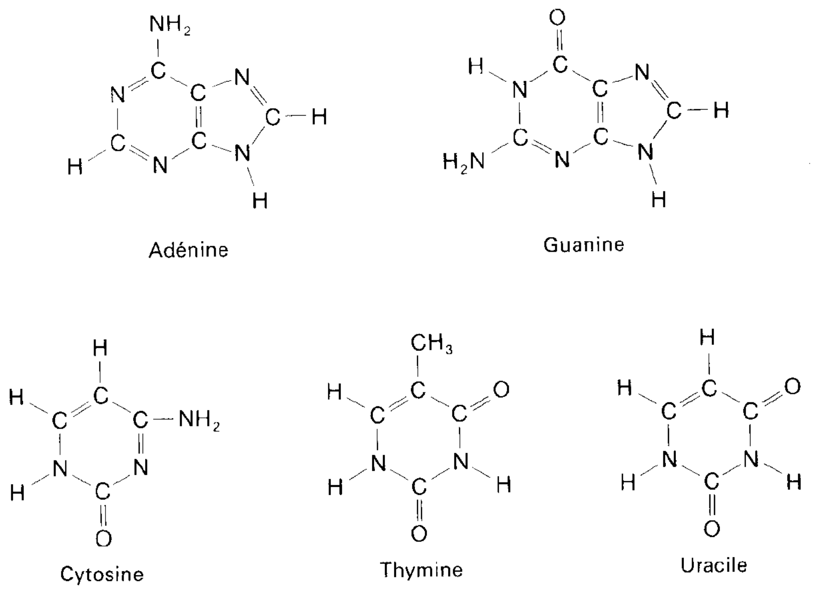


Figure 3.2: Cycles des bases nucléiques.

Une mise en correspondance d'un autre type intervient sur les nucléotides impliqués dans la formation des structures secondaires d'ARN. Ces structures sont essentiellement formées par une combinatoire de palindromes biologiques (voir la figure 2.9) et la substitution par exemple d'un nucléotide dans le brin d'une tige peut ne pas affecter la structure si elle est accompagnée d'une substitution sur l'autre brin permettant de maintenir la complémentarité des bases placées face à face (covariation). L'identification entre nucléotides est ici à la fois positionnelle et contextuelle : elle a lieu à un endroit précis et doit être accompagnée d'une autre substitution également très précise intervenant à un autre endroit structuralement lié au premier.

Enfin, une mise en correspondance plus subtile et bien plus complexe à établir met en jeu les nucléotides présents dans les régions de l'ADN ou de l'ARN impliquées dans des activités régulatrices. Ces régions sont souvent composées de suites contiguës de monomères, c'est-à-dire de mots, mais ces mots ne sont pas uniques pour un même signal (chaque mot en particulier peut dénoter une force variable de ce signal). La question est alors de savoir si l'on peut dans ce cas parler d'identification de nucléotides à certaines positions, ou si la mise en correspondance ne porte que sur le mot lui-même, et quelle est la nature de cette correspondance. En particulier, est-ce qu'elle obéit à une règle clairement définissable, dépent-elle de facteurs exclusivement internes aux signaux ou situés ailleurs? Aucune réponse définitive n'a encore été apportée à ces questions. De toute façon, il paraît assez clair que si une identification entre nucléotides est établie dans ce cas aussi, elle ne peut être que de nature positionnelle et contextuelle.

Cette étroite relation avec une position qui semble intervenir dans toute identification de nucléotides fait que l'on ne peut pas réellement parler d'une ressemblance entre deux quelconques d'entre eux en considérant uniquement leur nature chimique. En général donc, la valeur attribuée à un alignement de deux nucléotides différents est malheureusement indépendante de leurs positions dans la chaîne.

3.2.2.2 Les acides aminés

Le tableau 2.2 présenté dans le chapitre 2 donnait une idée des propriétés physico-chimiques que possèdent les 20 acides aminés présents naturellement dans les protéines. Il montrait aussi que certaines chaînes latérales présentent des propriétés en commun. Celles-ci concernent, par exemple, leur taille, leur caractère hydrophobe ou hydrophile, ou leur charge. Tous ces facteurs sont importants pour évaluer la ressemblance de ces monomères en relation avec la structure tridimensionnelle de la protéine. Cette structure doit sa forme non pas tant à l'identité exacte des acides aminés qui la composent mais plutôt au volume qu'ils occupent dans la configuration finale et aux liens qu'ils peuvent établir avec d'autres acides aminés de la chaîne ou avec le milieu aqueux dans lequel évolue la molécule. Une similarité au niveau de quelques-unes de ces propriétés signifie ainsi que certains résidus peuvent se remplacer mutuellement sans que la structure de la protéine en soit trop affectée et par conséquent sans que soit trop modifiée son activité métabolique ou catalytique (cette activité peut éventuellement s'en trouver modulée). Au contraire des nucléotides, il est donc possible de parler d'une réelle ressemblance entre acides aminés. Le problème est de mesurer cette ressemblance.

Il y a essentiellement deux façons de le faire, reflétant deux philosophies différentes, mais produisant des résultats souvent comparables. Dans les deux cas aussi, ce qui est obtenu en dernière instance est une matrice, appelée matrice de substitution, de 20 lignes et 20 colonnes qui attribue à chaque paire de résidus une valeur numérique dont on peut dire pour l'instant de manière simplifiée qu'elle indique leur degré de similarité.

La première façon de procéder pour calculer ces valeurs consiste à réaliser une comparaison des acides aminés directement entre eux sur la base de certaines de leurs caractéristiques. Celles-ci peuvent porter tout simplement sur l'identité des acides aminés, la matrice obtenue est alors la matrice identité. Ces caractéristiques peuvent concerner également le nombre maximum de nucléotides que les codons de deux acides aminés ont en commun. L'idée est que la substitution d'un acide aminé par un autre dans une protéine se produit d'autant plus rarement que le nombre de mutations correspondantes sur l'ADN est grand. Les matrices obtenues par cette méthode de comparaison sont les matrices dites génétiques [Feng *et al.*, 1985] [Fitch, 1966] [Fitch, 1967]. Les valeurs varient donc entre 0 (les acides aminés n'ont aucun nucléotide en commun) et 3 (les acides aminés sont identiques). Enfin, les caractéristiques les plus couramment utilisées pour mesurer la ressemblance entre acides aminés portent sur leurs propriétés physico-chimiques. Les trois matrices établies de cette façon utilisent ainsi de 2 à 5 propriétés pour calculer une distance euclidienne entre les acides aminés [Grantham, 1974] [Miyata *et al.*, 1979] [Rao, 1987].

La seconde façon de mesurer la ressemblance entre acides aminés est plus indirecte et nous allons nous y attarder plus longuement. Elle consiste à examiner avec quelle fréquence les résidus se substituent les uns aux autres dans un jeu de protéines homologues, c'est-à-dire ayant un ancêtre commun. L'idée ici est que deux acides aminés se ressemblent d'autant plus qu'est grande la fréquence observée de substitution de l'un par l'autre (conservant la fonction initiale). Une telle mutation est appelée une mutation acceptée (par l'évolution). Toutes les méthodes de cette seconde catégorie commencent donc par un alignement des protéines sélectionnées. Cet alignement est établi soit d'après une comparaison directe des séquences en considérant les protéines globalement [Benner *et al.*, 1994] [Dayhoff *et al.*, 1983] [Dayhoff and Eck, 1972] [Dayhoff *et al.*, 1978] [Gonnet *et al.*, 1992] [Jones *et al.*, 1992] [McLachlan, 1971] ou localement [Henikoff and Henikoff, 1992], soit d'après les structures secondaires [Levin *et al.*, 1986], ou tertiaires [Johnson and Overington, 1993] [Miyazawa and Jernigan, 1993] [Overington *et al.*, 1992]

[Overington *et al.*, 1990] [Risler *et al.*, 1988], soit encore d'après la comparaison de plusieurs protéines dont seules les séquences sont connues avec une (ou plusieurs) protéine dont la structure aussi est connue [Luthy *et al.*, 1991]. Dans certains cas les protéines sont alignées par paires [Benner *et al.*, 1994] [Dayhoff and Eck, 1972] [Dayhoff *et al.*, 1978] [Dayhoff *et al.*, 1983] [Gonnet *et al.*, 1992] [Jones *et al.*, 1992] [Levin *et al.*, 1986] [McLachlan, 1971] [Risler *et al.*, 1988], dans d'autres l'alignement concerne un ensemble de plus de deux protéines à la fois [Henikoff and Henikoff, 1992] [Johnson and Overington, 1993] [Luthy *et al.*, 1991] [Overington *et al.*, 1992] [Overington *et al.*, 1990]. Il n'est pas essentiel pour notre propos de savoir comment précisément sont calculées les valeurs de chacune de ces matrices, mais plutôt de comprendre le principe de base de leur construction sur un exemple particulier. Nous choisissons pour cela la plus ancienne et la plus connue de toutes : la matrice de Dayhoff.

Cette matrice a été élaborée sur la base d'un certain nombre de mutations ponctuelles acceptées observées entre des protéines très proches les unes des autres (moins de 15% de résidus différents entre chaque paire). Par ailleurs, Dayhoff avait établi deux hypothèses concernant le processus d'évolution des protéines. La première était que ce processus suit un modèle Markovien de substitution d'ordre 0 des acides aminés. Ceci veut dire que l'on considère la probabilité de mutation d'un résidu à n'importe laquelle des positions de la protéine comme étant indépendante de ce qui a pu se produire dans le passé à cet endroit. Par exemple, la probabilité qu'a une cystéine d'une protéine donnée d'être transformée en une sérine est la même pour tous les résidus de cystéine de la protéine en question, quelle que soit l'identité des résidus qui ont pu occuper le site auparavant. La seconde hypothèse faite par Dayhoff était que les événements mutationnels sont indépendants du contexte, c'est-à-dire de la position et des résidus adjacents sur la chaîne.

La matrice de Dayhoff a alors été calculée à partir d'une matrice de transition dont chaque élément fournit la probabilité pour qu'un acide aminé X soit remplacé par un acide aminé Y en une unité d'évolution. Les éléments diagonaux d'une telle matrice indiquent la probabilité qu'a un acide aminé de demeurer inchangé, et leur somme la probabilité qu'il n'y ait aucune substitution dans l'intervalle d'évolution considéré. Cette matrice a été normalisée de telle façon que cette somme corresponde à une probabilité de 99%. Ainsi, l'unité d'évolution représentée par la matrice de transition correspond à une mutation ponctuelle (substitution) acceptée par 100 acides aminés (d'où le nom de PAM1). Il faut remarquer que le temps n'est pas explicitement pris en compte dans cette construction, les probabilités de substitutions sont supposées être constantes par rapport à l'unité d'évolution. Elles ne sont donc pas fonction du laps de temps nécessaire pour que cette substitution ait lieu. C'est la raison pour laquelle Dayhoff considère valable de comparer deux protéines qui ont subi des mutations sur des échelles de temps différentes, à condition que l'on puisse supposer que cette échelle est constante pour une protéine donnée. En d'autres termes, l'unité d'évolution est supposée être proportionnelle au temps avec un facteur de proportionnalité différent d'une famille à l'autre mais, pour une famille donnée, constant sur toute son histoire évolutive.

Chaque élément de cette matrice de transition indique en fait la probabilité conditionnelle pour que l'acide aminé X soit remplacé par l'acide aminé Y, étant donné que X est remplacé, multiplié par la probabilité avec laquelle il est remplacé. Cette dernière probabilité représente la 'mutabilité' (en anglais 'mutability') relative de X et est estimée par la fréquence relative du nombre de mutations subies par X. Pour chaque paire de protéines, l'exposition au changement d'un acide aminé X est la fréquence d'occurrence de X multipliée par le nombre de toutes les mutations observées à chaque 100 sites de cette paire. L'exposition totale au changement est alors la somme sur toutes les paires de ces expositions pour chaque paire. Les probabilités conditionnelles quant à elles sont évaluées sur la base des fréquences observées de substitutions entre acides aminés des protéines sélectionnées. Celles-ci ont été choisies très proches afin d'éviter que les différences observées soient le résultat de mutations multiples, c'est-à-dire où un résidu X est remplacé par un résidu Y, qui lui-même est remplacé par un résidu Z donnant lieu à l'observation du couple XZ. De la même façon, ce choix garantit que lorsqu'aucune différence n'est observée (les acides aminés sont identiques), il n'y a pas eu de mutation (au niveau de la protéine) à cet endroit. La formule pour les éléments non diagonaux de la matrice de transition est alors la suivante :

$$\mathcal{M}(X,Y) = K.m_X.\frac{n_Y}{\sum_{X,Y} n_{XY}}$$

où n_{XY} est le nombre de substitutions observées pour la paire (X,Y), m_X est la mutabilité de X et K est une constante de pondération. La formule pour les éléments diagonaux est donnée par :

$$\mathcal{M}(X,X) = 1 - K.m_X.$$

Ainsi que nous l'avons vu, K est choisi de telle façon que $\sum_{X\in\Sigma} f_X.\mathcal{M}(X,X) = 99\%$ où f_X est la fréquence d'apparition de l'acide aminé X.

En accord avec le modèle Markovien, les matrices de probabilités correspondant à des distances évolutives plus grandes sont obtenues en multipliant la matrice PAM1 par elle-même le nombre de fois nécessaire. Ainsi, la matrice PAMn est égale à $(\text{PAM1})^n$ et représente la matrice des probabilités pour une séquence de protéines qui a subi n pour cent de mutations ponctuelles acceptées.

En fait, lorsque l'on veut effectuer des comparaisons entre protéines, il est plus utile d'employer une matrice dont les éléments indiquent le rapport de la probabilité qu'a un acide aminé d'être substitué par un autre à la fréquence de ces deux résidus dans des chaînes produites au hasard. Pour des raisons pratiques, les valeurs présentes dans les matrices PAM indiquent finalement les logarithmes de ces rapports (en anglais, 'log-odds'), multipliés par dix pour limiter dans les calculs en entiers les erreurs d'arrondis. Ces valeurs correspondent à des scores de substitution attribués à chaque paire de résidus. Ce score est ainsi nul, positif ou négatif suivant que les fréquences d'appariements observés dans les protéines alignées ayant même origine et exerçant une même activité sont égales, supérieures ou inférieures aux fréquences que le hasard laisserait prévoir.

Bien que parfois différentes dans leur approche, en particulier concernant les hypothèses faites sur le processus d'évolution des protéines, le mode de construction des autres matrices apparentées à celle de Dayhoff repose sur les mêmes principes de base et leurs scores indiquent ainsi toujours le logarithme du rapport pondéré de la probabilité qu'il y ait substitution d'un certain acide aminé X par un autre Y dans des protéines homologues (en anglais 'target frequency') par la probabilité que cette substitution se produise au hasard ('background frequency') qui est donnée dans l'hypothèse d'indépendance par le produit des fréquences d'apparition de X et Y [Altschul, 1991] [Karlin and Altschul, 1990] :

$$score(X,Y) = (\ln \frac{q_{XY}}{f_X.f_Y})/\lambda$$

où q_{XY} est la 'target frequency', f_X, f_Y sont les fréquences d'apparition de X et Y respectivement et λ est un facteur de normalisation.

La première question que l'on peut se poser quant à la fiabilité de toutes ces matrices concerne le fait que les protéines du jeu sélectionné doivent se trouver préalablement alignées. Il faut ainsi avoir une certaine confiance dans l'alignement initialement établi puisque de sa qualité va dépendre celle des scores de substitutions obtenus à la fin. Or la bonne réalisation d'un tel alignement implique déjà une certaine connaissance du poids à attribuer à l'appariement des acides aminés. Par ailleurs, les mutations observées entre séquences homologues ne mettent pas seulement en jeu les substitutions mais aussi des délétions et insertions. Il faut donc pouvoir placer correctement les trous dans l'alignement. Il y a eu diverses façons d'aborder ces problèmes. La plus simple, employée par Dayhoff, consiste ainsi que nous venons de le voir à travailler avec des séquences très proches en termes d'évolution. Dans ces circonstances, peu de mutations sont survenues (en particulier très peu de délétions et insertions) et il est relativement aisé d'obtenir un alignement exact par minimisation du nombre de ces mutations observées entre chaque séquence et un ancêtre commun présumé. Une autre façon de procéder consiste à compter les substitutions dans les segments alignés sans trous qui correspondent aux régions les mieux conservées du jeu de protéines sélectionnées. C'est la méthode de Henikoff, qui de plus aligne plusieurs séquences simultanément et non plus deux seulement à chaque fois. Nous verrons plus tard pourquoi cela peut produire un alignement plus sûr. Une troisième façon de faire consiste à aligner des structures et non des séquences ce qui rend l'équivalence entre résidus plus facile à établir avec certitude. Plusieurs méthodes s'appuient ainsi, lorsque cela est possible, sur des comparaisons structurales pour établir leurs scores de substitutions. Enfin, quelle que soit la façon d'établir l'alignement initial, les erreurs que celui-ci présente éventuellement peuvent être rectifiées par une application itérative de la méthode servant à obtenir les valeurs de la matrice. Celle-ci peut en effet être utilisée pour produire un second alignement des séquences de départ, donnant lieu à une seconde matrice de substitution, qui est employée pour réaliser un troisième alignement et ainsi de suite jusqu'à convergence des valeurs de la matrice. Cette méthode de raffinements successifs est celle adoptée par Henikoff, Jones, Gonnet et Benner. Bien sûr elle ne peut garantir absolument la convergence sur des valeurs correctes si l'alignement initial est violemment faux.

On peut aussi s'interroger sur la sélection même du jeu initial. Toute mesure établie sur la base de fréquences observées reflète d'autant mieux la réalité du phénomène étudié que le nombre d'observations effectuées est grand. Dans le cas présent, il faut donc pouvoir disposer de nombreuses protéines. Le jeu avec lequel Dayhoff a initialement travaillé comportait 71 molécules. Gonnet et Benner [Benner *et al.*, 1994] [Gonnet *et al.*, 1992] ont refait des calculs très similaires à ceux de Dayhoff mais sur toute une banque de séquences cette fois (ce sont ainsi 1,7 millions de paires de séquences de la version 64 de la base MIPS [Gonnet *et al.*, 1992] qui ont été alignées et analysées). En particulier, ces séquences se trouvaient à toutes sortes de distances évolutives les unes par rapport aux autres et les matrices correspondant à des distances inférieures à 100 PAM ont été déterminées directement, sans extrapolation, contrairement aux matrices de Dayhoff. Les critères utilisés pour garantir une certaine fiabilité aux alignements nécessaires pour calculer des matrices à ces distances étaient essentiellement que les séquences impliquées devaient être suffisamment longues (comportant plus de 100 acides aminés) et le score final de l'alignement (obtenu par une technique de raffinements successifs) supérieur à un certain seuil. La valeur de ces critères a elle-même été établie par simulation sur des paires de séquences artificiellement engendrées à des distances PAM spécifiques l'une de l'autre.

Les résultats obtenus sont intéressants dans la mesure où ils montrent qu'il n'y aurait pas une simple différence dans la quantité de mutations observées entre séquences proches et séquences éloignées en termes d'évolution, mais aussi une différence dans le type de mutations produites. Ainsi pour de faibles distances évolutives, les valeurs de la matrice de substitution reflèteraient les proximités entre les codons des acides aminés, tandis que pour de grandes distances ces valeurs révèleraient plutôt les similarités physico-chimiques entre les acides aminés. Les travaux de Gonnet et Benner, ainsi que ceux de Wilbur [Wilbur, 1985] auparavant, semblent indiquer que l'hypothèse d'un modèle Markovien de substitution entre acides aminés est incorrecte (Wilbur argumente en outre que les fréquences de substitutions observées entre acides aminés qui sont la conséquence du remplacement de deux ou trois nucléotides sont plus grandes que celles que l'on pourrait déduire des fréquences de substitutions n'exigeant le remplacement que d'un seul nucléotide). Une matrice obtenue par comparaison de séquences très proches comme cela est le cas de Dayhoff ne serait en tout cas pas une bonne base de départ pour l'obtention de matrices pouvant servir à comparer des séquences plus distantes. Ces matrices sont sans doute construites de manière plus fiable à partir de l'observation directe de séquences évolutivement éloignées comme le font Gonnet et Benner, mais également Henikoff, Overington, Johnson et Lüthy. Dans le cas de Henikoff, plusieurs matrices sont calculées comme pour la série PAM, chacune appropriée à une certaine distance d'évolution (la matrice BLOSUM n correspond ainsi à une distance où au moins $n\%$ d'identité est observé entre séquences).

Un troisième problème qui surgit alors est que cette distance n'est bien sûr pas connue à l'avance pour les séquences que l'on veut comparer. Une façon d'éviter la difficulté consiste à utiliser plusieurs matrices à la suite, chacune correspondant à une distance différente [Altschul, 1993] [Collins *et al.*, 1988] et de choisir celle fournissant le 'meilleur résultat', en général celle rapprochant le plus les deux séquences comparées dans un sens qui sera rendu précis plus loin. Une autre hypothèse est cependant faite ici, concernant cette fois l'homogénéité des séquences par rapport au temps de l'évolution. Lorsque plusieurs séquences sont en jeu, comme cela va être le cas pour nous, celles-ci peuvent former divers groupes distincts en termes de distance évolutive. Aucune matrice ne peut alors être la plus appropriée simultanéement pour tous ces groupes.

Par ailleurs, nous avons vu dans le cas des nucléotides que toute comparaison de ces éléments est indissociable de leur contexte dans une séquence. L'idée intuitive que l'on se fait des acides aminés est qu'il est possible de dire quelque chose à propos de la ressemblance de deux d'entre eux dans l'absolu. D'une certaine façon cela est vrai, il est possible de mesurer une ressemblance entre eux soit directement, soit indirectement à travers des alignements comme nous venons de le montrer, et il est sans doute très significatif que les matrices qui ont été établies par cette dernière méthode reflètent souvent les propriétés physico-chimiques de ces résidus même lorsque ceux-ci n'ont pas du tout été pris en considération lors de leur élaboration. Cela est le cas en particulier de la matrice PAM250 de Dayhoff ainsi que le montre la figure 3.3. Mais les mesures fournies par ces matrices représentent seulement une moyenne de diverses caractéristiques. Or un acide aminé n'est qu'un élément dans une molécule et le rôle qu'il joue dans la structuration de celle-ci ou dans la fonction qu'elle exerce a une importance variable suivant la position qu'il y occupe et peut ne concerner à chaque fois qu'une partie de ses propriétés. La ressemblance entre acides aminés est donc elle aussi fortement dépendante du contexte. Ainsi, par exemple, l'arginine qui est normalement rangée parmi les acides aminés polaires (voir la figure 3.11 plus loin) peut présenter un comportement hydrophobe lorsque localisée dans une hélice α [Cornette *et al.*, 1987]. Des travaux récents, notamment de la part d'Overington [Overington *et al.*, 1992] [Overington *et al.*, 1990] et de Lüthy [Luthy *et al.*, 1991], ont porté sur l'observation des taux de substitution des résidus en fonction de ce contexte. Les valeurs numériques obtenues, et donc les scores, dépendent ainsi de l'emplacement de ces résidus dans la molécule et, en particulier, du type de structure secondaire dans laquelle ils se situent. Ces auteurs fournissent non pas une mais à nouveau plusieurs matrices, chacune donnant cette fois les scores des substitutions observées entre acides aminés faisant partie par exemple soit d'une hélice α, soit d'un feuillet β, soit encore d'un coude (en anglais 'turn'). Ces matrices 'contextuelles' sont très intéressantes mais posent une question importante : comment en effet s'en servir lorsque l'on veut comparer des macromolécules dont seules les séquences sont connues?

Une autre approche du problème du contexte est celle adoptée par Gonnet [Gonnet *et al.*, 1994] qui a étudié les fréquences des substitutions observées pour les dipeptides cette fois,

		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
		Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp
C	Cys	12																			
S	Ser	0	2																		
T	Thr	-2	1	3																	
P	Pro	-3	1	0	6																
A	Ala	-2	1	1	1	2															
G	Gly	-3	1	0	-1	1	5														
N	Asn	-4	1	0	-1	0	0	2													
D	Asp	-5	0	0	-1	0	1	2	4												
E	Glu	-5	0	0	-1	0	0	1	3	4											
Q	Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H	His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

Sulfhydrile

Hydrophile de petite taille

Acide, Acide amine, Hydrophile

Basique

Hydrophobe de petite taille

Aromatique

Figure 3.3: Matrice PAM250 établie par Dayhoff.

où un dipeptide est une suite contiguë de 2 acides aminés. Les matrices obtenues ont ainsi 400 lignes et 400 colonnes. Gonnet montre que les fréquences trouvées sont souvent différentes du produit des fréquences des substitutions observées pour les acides aminés, rendant fausse l'autre hypothèse importante du modèle employé par Dayhoff et qui concernait l'indépendance des positions. Gonnet n'a cependant examiné que les interactions au voisinage immédiat d'un résidu, alors que d'autres plus lointaines existent sans doute aussi. Par ailleurs, même en ce qui concerne ces interactions très proches, on ne dispose pas d'assez de données pour que les matrices établies soient significativement éloignées de celles couramment utilisées.

Pour l'instant, il faut noter que, quelles que soient les limitations des valeurs fournies par ces matrices de substitutions, en pratique elles fournissent une bonne évaluation de la ressemblance entre acides aminés. Cela est à la fois positif et négatif. C'est positif car cela signifie que le choix d'une matrice est, dans une bonne partie des cas, moins crucial que l'on ne pourrait le craindre. Les matrices les plus connues et les mieux établies sont alors plus ou moins équivalentes dans leur pouvoir de mesurer cette ressemblance [Risler *et al.*, 1988] [Vogt *et al.*, 1995] (voir la figure 3.4) : il s'agit en particulier de la série PAM des matrices de Dayhoff, ou de ses dérivées [Benner *et al.*, 1994] [Gonnet *et al.*, 1992] [Jones *et al.*, 1992], et de la série BLOSUM des matrices de Henikoff. Cette bonne performance en pratique est négative dans le sens où elle est justement un reflet de cet effet 'moyenne' des mesures obtenues. Elles 'marchent bien' en général mais sont-elles capables d'évaluer correctement les ressemblances plus difficiles à distinguer, plus subtiles? Les éléments présentés dans ce chapitre et dans le précédent montrent ainsi que certains attributs des acides aminés, et cela vaut aussi pour les nucléotides, sont plus importants que d'autres pour mesurer leur similarité et que cette importance varie en particulier suivant l'endroit où se trouvent ces monomères et la fonction qu'ils exercent.

3.2.2.3 Les perles comme des objets 3D

Nous avons vu dans le chapitre précédent que l'activité exercée par les macromolécules dans l'organisme peut dépendre fortement de la forme géométrique que ces objets adoptent dans l'espace. Or la structure tridimensionnelle d'une macromolécule est déterminée par l'arrangement spatial des atomes des monomères la composant, et ceux-ci forment une chaîne. Une représentation de la structure qui préserve cet ordre linéaire nous permet ainsi de les considérer comme des chaînes de caractères. Avant de mesurer la ressemblance des monomères en tant qu'objets 3D, il nous faut voir comment les coder en symboles d'un alphabet géométrique.

Considérons par exemple le cas d'une protéine, et voyons comment établir le codage de sa structure en une chaîne. Tous les atomes n'ont pas besoin d'être pris en considération pour cette représentation en chaîne et il est possible de ne considérer en fait, pour chaque résidu, que son carbone α, l'atome d'azote du groupement amino NH_2 et l'atome de carbone du groupement carboxyl COOH. La chaîne de ces atomes pour tous les résidus constitue ce qui est appelé le squelette de la protéine. La conformation locale de ce squelette peut être définie à chaque résidu par trois coordonnées internes habituellement appelées angles dihédraux : Φ, Ψ and ω [Creighton, 1993] [Ramachandran *et al.*, 1963]. Φ et Ψ correspondent aux angles qu'établissent entre eux, d'un côté le plan où se situe la liaison peptidique, et d'un autre celui contenant soit l'atome de carbone soit celui d'azote respectivement, ainsi que l'atome de carbone α le plus proche et la chaîne latérale reliée à cet atome (voir la figure 3.5). L'angle ω quant à lui correspond à un angle de torsion autour de la liaison peptidique. Pour des raisons chimiques, ω est fixé (à 0 ou 180°) et peut être ignoré en première approximation, de sorte que les coordonnées internes du squelette sont constituées des seuls angles Φ et Ψ que l'on représente sur une carte planaire Φ, Ψ nommée carte de Ramachandran [Ramachandran *et al.*, 1963]. La structure du

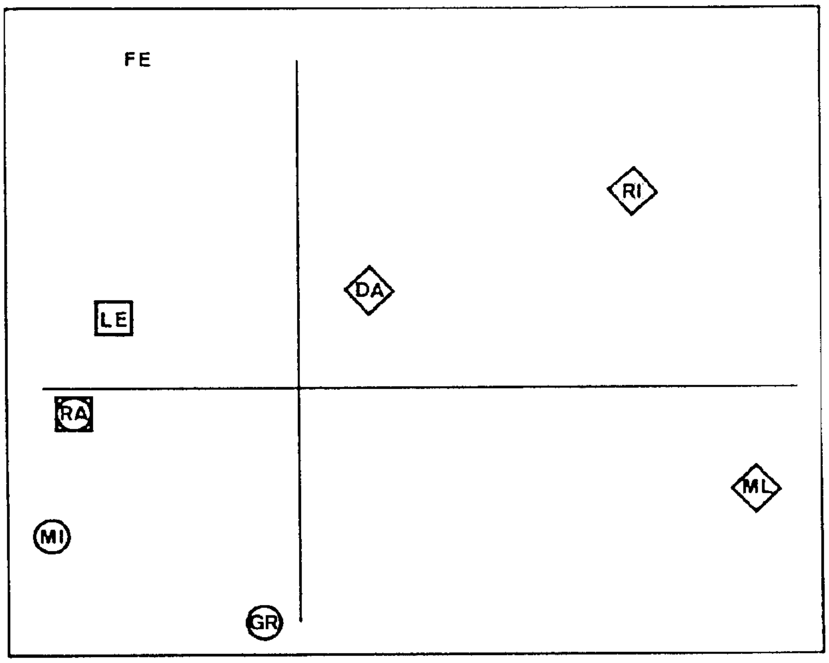


Figure 3.4: Une comparaison par analyse factorielle de diverses matrices de substitution réalisée par Risler *et al.* [Risler *et al.*, 1988] : DA (Dayhoff), FE (Feng), GR (Grantham), LE (Levin), MI (Miyata), ML (McLachlan), RI (Risler).

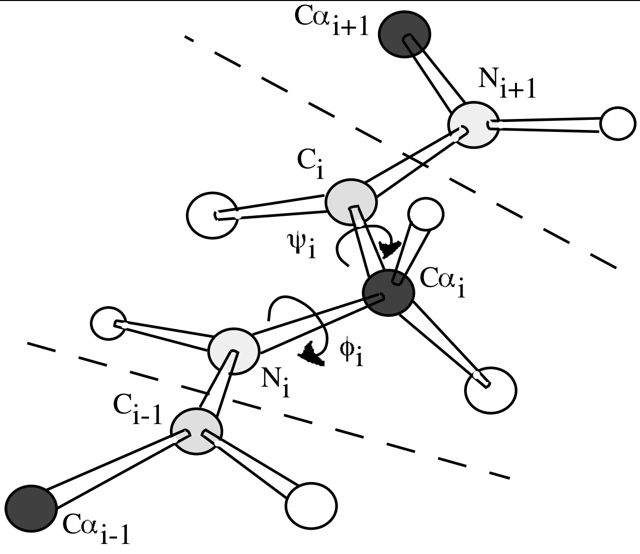
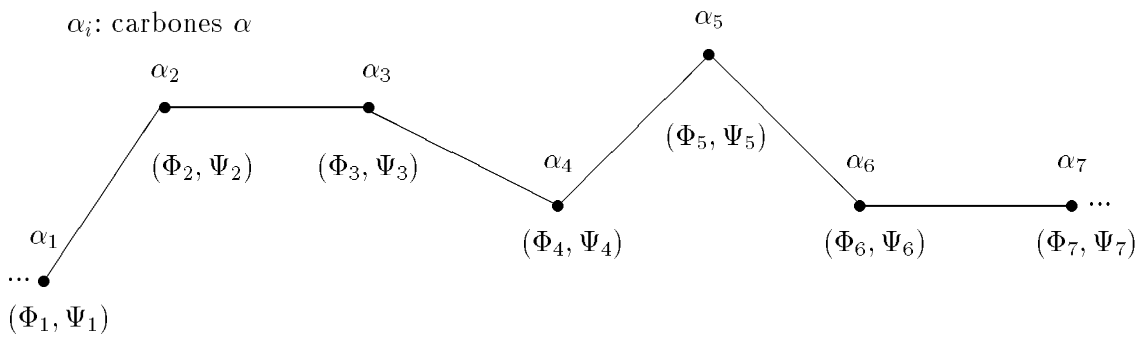


Figure 3.5: Angles Φ et Ψ.

squelette peut ainsi être définie de manière unique par la succession linéaire des paires d'angles (Φ, Ψ) le long de ce squelette. Il est habituel de ne représenter que les carbones α de ce dernier comme sur la figure 3.6. Observons qu'alors les traits qui sont dessinés entre deux de ces carbones ne sont que des pseudo-liens qui ne correspondent pas à des liaisons chimiques. Ces traits ont des longueurs approximativement égales car les distances entre carbones α d'une chaîne polypeptidique sont à peu près constantes (environ 3,8 Angström). Comme ces paires d'angles (Φ, Ψ) représentent des paires de valeurs réelles, elles doivent être recodées sur des valeurs discrètes de manière à pouvoir travailler avec un alphabet de symboles discrets. Pour réaliser cela, une grille de maille ϵ est construite sur la carte de Ramachandran [Pothier, 1993] [Sagot *et al.*, 1995b] (voir la figure 3.10 plus loin — notons aussi que ce qui apparaît comme étant plat sur la figure représente en fait la surface d'une sphère). Le centre de chaque carré de la grille — on appelle ce centre un noeud — reçoit un numéro allant de 1 à $(360/\epsilon)^2$. Ensuite, toute paire d'angles (Φ, Ψ) est mise en relation avec le centre du carré à l'intérieur duquel elle se place. Par conséquent, à chaque paire (Φ, Ψ) de la protéine correspond à la fois un noeud et le numéro qui a été attribué à celui-ci, mais il est clair qu'à un noeud donné, et donc à un numéro, peuvent correspondre plusieurs paires d'angles identiques ou très proches les unes des autres (toutes se référant à des points internes à une même maille). Une paire d'angles quelconque se trouve ainsi codée sur un seul symbole.

La ressemblance entre monomères comme des objets dans l'espace est, dans ce cas précis, immédiate à établir : elle va porter sur une ressemblance entre paires d'angles et donc, après codage, à une relation de voisinage sur les carrés de la grille ainsi que nous le montrons plus loin. Observons que, si l'ordre linéaire des monomères n'était pas pris en considération, c'est-à-dire, si nous voulions les considérer non plus comme des résidus le long d'une chaîne mais comme des objets libres dans l'espace, il ne serait plus possible d'attribuer à chacun un code qui permette de comparer deux monomères en tant qu'entités indépendantes. Tout codage dans ce cas dépendrait de l'existence d'un système de référence, qui pourrait être local ou global, mais qui ne pourrait être absolu.



Structure d'une protéine codée initialement comme :
$...(\Phi_1, \Psi_1)(\Phi_2, \Psi_2)(\Phi_3, \Psi_3)(\Phi_4, \Psi_4)(\Phi_5, \Psi_5)(\Phi_6, \Psi_6)(\Phi_7, \Psi_7)...$

Figure 3.6: Structure du squelette d'une protéine définie comme une succession linéaire de paires d'angles (Φ, Ψ).

3.2.3 Matrice, relation, couverture

3.2.3.1 Les matrices

Rappelons que notre but dans cette discussion sur la ressemblance des perles était celui d'établir le poids à donner à l'appariement de deux monomères dans un alignement. Une manière de procéder consiste à faire appel à une matrice comme celles vues précédemment. La façon la plus immédiate d'utiliser les mesures de ressemblance présentes dans une de ces matrices est d'attribuer directement leurs valeurs à cet appariement. Observons que dans le cas des nucléotides, cette matrice demeure le plus souvent l'identité.

Ces matrices en fait ne fournissent pas le même type d'information et ne possèdent pas toutes les mêmes propriétés mathématiques. Dans le cas des protéines, celles obtenues par comptage des mutations observées sont, à l'exception de celle de Risler, toutes des matrices de proximité. Celle de Risler ainsi que les matrices de Grantham, de Miyata et de Rao sont des matrices de distances. Finalement les matrices génétiques sont des matrices de similarité.

Commençons par établir certaines définitions [Cailliez and Pages, 1976] :

Définition 3.2.1 *Un* <u>*indice de proximité*</u> *sur un ensemble E est une application f de $E \times E$ dans l'ensemble $\mathcal{R}$ telle que :*

1. *$f(i,j) = f(j,i)$ pour tout $(i,j) \in E \times E$;*
2. *$f(i,j) =$ est d'autant plus élevé que i est proche de j, pour tout $(i,j) \in E \times E$.*

Définition 3.2.2 *Un* <u>*indice de similarité*</u> *sur un ensemble E est une application f de $E \times E$ dans l'ensemble $\mathcal{R}$ telle que :*

1. *$f(i,j) = f(j,i)$ pour tout $(i,j) \in E \times E$;*
2. *$f(i,i) =$ maximum parmi toutes les valeurs attribuées aux éléments de $E \times E$, cela pour tout $i \in E$.*

Définition 3.2.3 *Un* <u>*indice de dissimilarité*</u> *sur un ensemble* E *est une application* f *de* $E \times E$ *dans l'ensemble* $\mathcal{R}^+$ *des nombres positifs ou nuls telle que :*

1. $f(i,j) = f(j,i)$ *pour tout* $(i,j) \in E \times E$;
2. $f(i,i) = 0$ *pour tout* $i \in E$.

Définition 3.2.4 *Une* <u>*distance*</u> *sur un ensemble* E *est un indice de dissimilarité sur* E *qui vérifie les deux propriétés supplémentaires suivantes :*

1. $f(i,j) = 0 \Longrightarrow i = j$;
2. $f(i,k) \leq f(i,j) + f(j,k)$ *pour tout* $i, j, k \in E$ *(*<u>*inégalité triangulaire*</u>*).*

Utilisant ces divers indices, nous pouvons alors établir des définitions formelles pour les matrices construites dans la section précédente.

Définition 3.2.5 *Soit* Σ *un alphabet. Une* <u>*matrice de proximité/similarité*</u> *est un indice de proximité/similarité de* $\Sigma \times \Sigma$ *dans* $\mathcal{R}$.

Définition 3.2.6 *Une matrice de proximité/similarité* $\mathcal{M}$ <u>*normalisée*</u> *est une matrice telle que, pour tout* $a, b \in \Sigma$ *:*

1. $0 \leq \mathcal{M}(a,b) \leq 100$.

Définition 3.2.7 *Soit* Σ *un alphabet. Une* <u>*matrice de dissimilarité/distance*</u> *est un indice de dissimilarité/distance de* $\Sigma \times \Sigma$ *dans* $\mathcal{R}^+$.

Malgré leurs différences, des liens existent entre certaines de ces matrices. Ainsi, une matrice de distance peut facilement être transformée en une matrice de similarité. Il suffit de la normaliser, puis d'attribuer à chaque paire d'éléments de l'alphabet (nucléotides ou acides aminés) un score égal à 100, moins le score que la paire possédait dans la matrice d'origine. L'opération inverse, à savoir la même démarche effectuée sur une matrice de similarité, ne nous fournit pas nécessairement une matrice de distance, la raison étant que l'inégalité triangulaire n'y est pas toujours observée.

La non obéissance par les matrices de proximité, comme par exemple les matrices PAM et BLOSUM, de la règle 2 relative aux indices de similarité peut quant à elle paraître non intuitive. Elle s'explique facilement si l'on se souvient que les valeurs attribuées aux éléments de la diagonale principale de ces matrices indiquent la probabilité de retrouver le même acide aminé à une même position après n mutations. Il n'y a pas de raison pour que cette probabilité soit la même pour tous les résidus. Ceux en particulier possédant un ensemble de propriétés peu communes mutent rarement entre protéines homologues, un remplacement de l'un d'entre eux à un endroit quelconque de la molécule risquant de transformer complètement la structure de celle-ci. Cela est le cas de la cystéine par exemple qui est un résidu possédant plusieurs fonctions uniques comme celle de lier les atomes métalliques et d'établir des ponts disulfures.

Dans la plupart des cas cependant, lorsque nous utilisons ces matrices pour attribuer un score à l'appariement de deux monomères dans un alignement, nous travaillons avec des matrices de similarité normalisées entre 0 et 100. Une telle matrice, dérivée de la matrice PAM250 de la figure 3.3 et corrigée par Brutlag [Brutlag *et al.*, 1990] pour tenir compte de certaines des critiques de Wilbur, est indiquée dans la figure 3.7.

	S	G	A	I	V	L	T	D	E	R	K	M	F	Y	C	P	N	Q	H	W
S	17	17	17	6	8	6	15	10	10	8	12	4	4	3	8	13	11	8	6	3
G	17	48	20	5	9	5	12	13	12	5	9	4	3	2	4	11	12	8	5	1
A	17	20	23	9	13	9	17	12	12	7	11	6	4	3	5	16	11	9	7	1
I	6	5	9	18	20	18	8	3	3	3	6	7	8	3	3	4	3	3	2	1
V	8	9	13	20	30	20	11	5	5	4	7	10	6	3	5	7	5	5	4	1
L	6	5	9	18	20	60	8	4	4	5	7	20	16	7	1	6	5	6	5	4
T	15	12	17	8	11	8	18	8	8	6	12	5	3	2	4	10	9	6	5	2
D	10	13	12	3	5	4	8	20	18	5	11	2	1	1	1	6	12	10	7	1
E	10	12	12	3	5	4	8	18	21	5	11	2	1	1	1	7	10	13	7	1
R	8	5	7	3	4	5	6	5	5	29	23	3	2	1	2	7	6	8	8	6
K	12	9	11	6	7	7	12	11	11	23	42	9	2	2	2	8	13	12	9	3
M	4	4	6	7	10	20	5	2	2	3	9	11	4	1	1	3	2	3	1	1
F	4	3	4	8	6	16	3	1	1	2	2	4	58	31	1	2	2	1	3	4
Y	3	2	3	3	3	7	2	1	1	1	2	1	31	55	5	1	3	1	4	3
C	8	4	5	3	5	1	4	1	1	2	2	1	1	5	93	3	2	1	2	1
P	13	11	16	4	7	6	10	6	7	7	8	3	2	1	3	35	7	8	6	1
N	11	12	11	3	5	5	9	12	10	6	13	2	2	3	2	7	11	7	8	1
Q	8	8	9	3	5	6	6	10	13	8	12	3	1	1	1	8	7	16	11	1
H	6	5	7	2	4	5	5	7	7	8	9	1	3	4	2	6	8	11	26	1
W	3	1	1	1	1	4	2	1	1	6	3	1	4	3	1	1	1	1	1	100

Figure 3.7: Matrice PAM250 obtenue à partir de la matrice de la figure 3.3, corrigée par Brutlag et normalisée entre 0 et 100.

Nous allons voir plus loin que l'emploi direct de ces matrices conduit à l'affectation d'une valeur à un alignement qui est elle-même une mesure moyenne de la ressemblance des molécules comparées, reflétant ainsi la caractéristique principale de la construction de ces matrices. Mais celles-ci peuvent également servir de base de départ à d'autres façons de mesurer la ressemblance qui sont particulièrement appropriées à la réalisation de comparaisons multiples. C'est ce que nous allons voir maintenant.

3.2.3.2 Relation

3.2.3.2.1 Relation sur l'alphabet des monomères

La première approche correspond à une discrétisation de la notion de ressemblance. Cela paraît être un retour en arrière en quelque sorte puisque deux éléments sont à nouveau considérés comme étant soit similaires, soit non similaires, sans moyen terme entre les deux. Mais cette fois-ci, la similarité porte sur plus que la simple identité des monomères en jeu. Nous dirons ainsi que deux monomères se ressemblent s'il existe une relation entre eux établie sur la base de quelques-unes de leurs caractéristiques. La motivation principale de cette approche est qu'elle permet une première formalisation de la relation entre plus de deux objets à la fois qui sera mathématiquement bien établie. Mais commençons par voir comment définir une telle relation.

La manière la plus classique consiste à établir un seuil sur une des matrices de substitution vues précédemment [Brutlag *et al.*, 1990] [Cobbs, 1994] [Soldano *et al.*, 1995]. Nous obtenons alors :

Définition 3.2.8 *Soit Σ un alphabet et $\mathcal{M}$ une matrice de similarité (ou de proximité). Une relation R_{seuil} de similarité sur $\Sigma \times \Sigma$ est définie par :*

$$\textit{pour tout } a,\ b \in \Sigma,\ a\ R_{seuil}\ b$$
$$\Updownarrow$$

$$\mathcal{M}(a,b) \geq seuil$$
$$\underline{ou}$$
$$a = b \text{ (a et b désignent en fait le même monomère)}$$

où 'seuil' est une valeur prise entre les valeurs minimum et maximum de $\mathcal{M}$.

Une définition semblable peut être établie pour une relation obtenue à partir d'une matrice de distance. Bien entendu, si la matrice est l'identité (c'est le cas pour l'ADN et l'ARN), la relation obtenue est alors l'identité.

Notation 3.2.1 *On note I la relation d'identité.*

Il est clair que la relation R_{seuil} est symétrique (puisque la matrice l'est) et réflexive (par construction) mais qu'en général cependant elle n'est pas transitive. Elle n'est donc le plus souvent pas une relation d'équivalence et ne définit donc pas toujours une partition de l'ensemble des monomères. Une manière à la fois d'illustrer ce phénomène de non-transitivité et surtout de sentir le caractère subjectif qui peut intervenir quant au choix du seuil permettant d'établir une telle relation est donné dans la figure 3.8. Il est certain que la ressemblance du dessin a avec le dessin b est frappante. Il en est de même pour celle du dessin b avec c, et de c avec d. Mais où placer le seuil qui permet de distinguer ces objets? Ceci n'est qu'une image, mais il est évident que le même problème se pose ici aussi, et nous observons ainsi l'introduction d'un nouveau point de vue dans la notion de similarité entre monomères, outre celui déjà introduit lors de la construction de la matrice.

Bien entendu, la relation de similarité entre monomères peut être établie directement, sans utilisation d'une matrice. Nous employons alors la notation simplifiée R. Ce qui est intéressant pour nous, c'est que quelle que soit la façon de définir une telle relation R, celle-ci conduit de manière immédiate aux notions de clique et de clique maximale.

Définition 3.2.9 *Soit Σ un alphabet et R une relation binaire sur $\Sigma \times \Sigma$. Soit C un sous-ensemble de Σ. On appelle C une <u>clique</u> de R si, pour tout $a, b \in C$, on a $a \ R \ b$. En d'autres termes, C est une clique si tous ses éléments sont 2 à 2 en relation.*

Si la relation R est représentée par son graphe G, alors les cliques de R sont tous les sous-graphes complets de G.

Définition 3.2.10 *Soit Σ un alphabet, R une relation sur $\Sigma \times \Sigma$ et C une clique de R. C est dit être une <u>clique maximale</u> si, pour tout $a \in \Sigma \backslash C$, $C \cup \{a\}$ n'est plus une clique.*

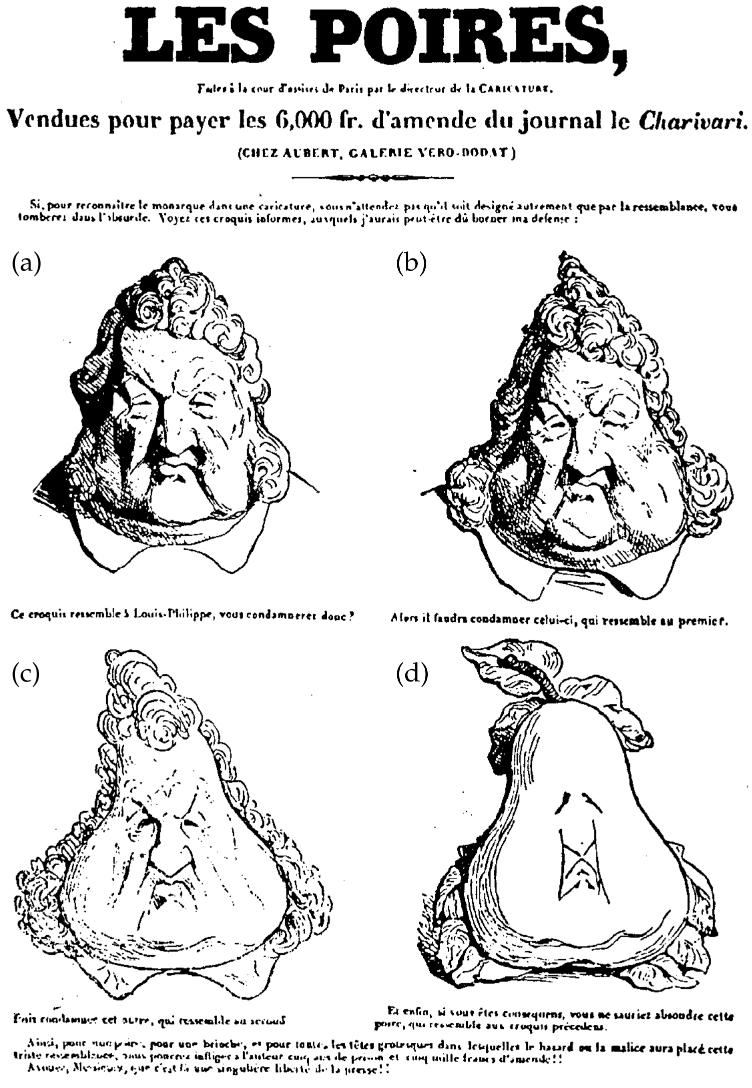
En termes du graphe G représentant R, les cliques maximales de R sont tous les sous-graphes complets maximaux de G.

Un exemple d'une telle relation et de ses cliques maximales est donnée dans la figure 3.9.

Cette représentation de la mesure de ressemblance entre deux monomères permet ainsi une extension au cas de plus de deux éléments dont nous allons voir plus loin qu'elle est à la fois très satisfaisante du point de vue conceptuel, et conduit à des algorithmes de comparaison de chaînes de caractères qui sont efficaces.

Avant de clore cette section, observons qu'à partir de la notion de clique maximale, il est possible de définir un concept qui sera utilisé par la suite et qui mesure l'indice de non-transitivité d'une relation R de similarité par rapport à l'identité :

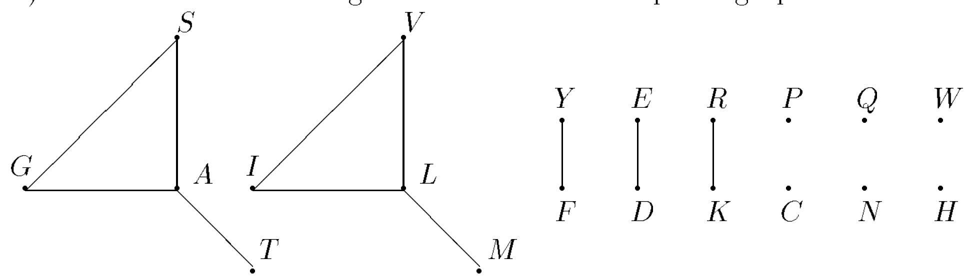
Définition 3.2.11 *Soit R une relation sur Σ et $a \in \Sigma$, on note g_a le nombre de cliques maximales auxquelles a appartient.*



282. C. Philipon. Les Poires. Reproduction d'après Le Charivari. *1834.*

Figure 3.8: Une illustration de la notion d'équivalence.

Soit $\Sigma = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ l'alphabet des acides aminés et R la relation de similarité entre ces acides aminés obtenue en seuillant (à la valeur 17) la matrice PAM de la figure 3.7 et schématisée par le graphe suivant :



Les cliques maximales de R sont les ensembles :
$\{A, S, G\}, \{A, T\}, \{I, L, V\}, \{L, M\}, \{F, Y\}, \{D, E\}, \{K, R\}, \{C\}, \{P\}, \{N\}, \{Q\}, \{W\}, \{H\}$.

Figure 3.9: Exemple d'une relation de similarité entre les acides aminés et de ses cliques maximales.

Définition 3.2.12 *Soit R une relation de Σ, on appelle g le plus grand nombre de cliques maximales de R à laquelle un symbole de Σ appartient, à savoir :*

$$g = Max\ \{g_a \mid a \in \Sigma\}.$$

On appelle $\overline{g}$ la valeur moyenne de g_a pour $a \in \Sigma$, c'est-à-dire :

$$\overline{g} = \frac{\sum_a g_a}{\mid \Sigma \mid}.$$

Dans l'exemple de R donné dans la figure 3.9, g est égal à 2 et $\overline{g}$ à 1.1. L'indice de non-transitivité d'une relation d'équivalence est la même que celle de l'identité, à savoir : $g = 1$.

3.2.3.2.2 Relation sur l'alphabet des objets 3D — Graphe d'intervalles

Dans le cas où les monomères sont considérés en tant qu'objets dans l'espace, plus spécialement en tant que résidus le long d'une chaîne qui se replie sur elle-même pour former une structure tridimensionnelle, nous avons vu (section 3.2.2.3) qu'il est possible de les associer de manière univoque à des paires d'angles, qui elles-mêmes sont codées sur des symboles qui sont les numéros des carrés d'une grille d'échantillonnage. Ces symboles (les numéros) représentent un nouvel alphabet sur lequel la structure de la protéine est définie et une relation R est alors établie entre ces symboles (noeuds de la carte) de la façon suivante :

$$\forall\ (\alpha, \beta) \in \Sigma^2,\ \alpha\ R\ \beta \text{ si et seulement si } \exists \text{ un carré de côté } 2K\epsilon \text{ qui contient } \alpha \text{ et } \beta$$

où α et β sont les noeuds de la grille et K est un paramètre (appelé marge) qui peut être ajusté pour augmenter ou diminuer la tolérance. Comme pour une relation sur l'alphabet des monomères, nous pouvons établir l'indice de non-transitivité de la relation R sur l'alphabet des

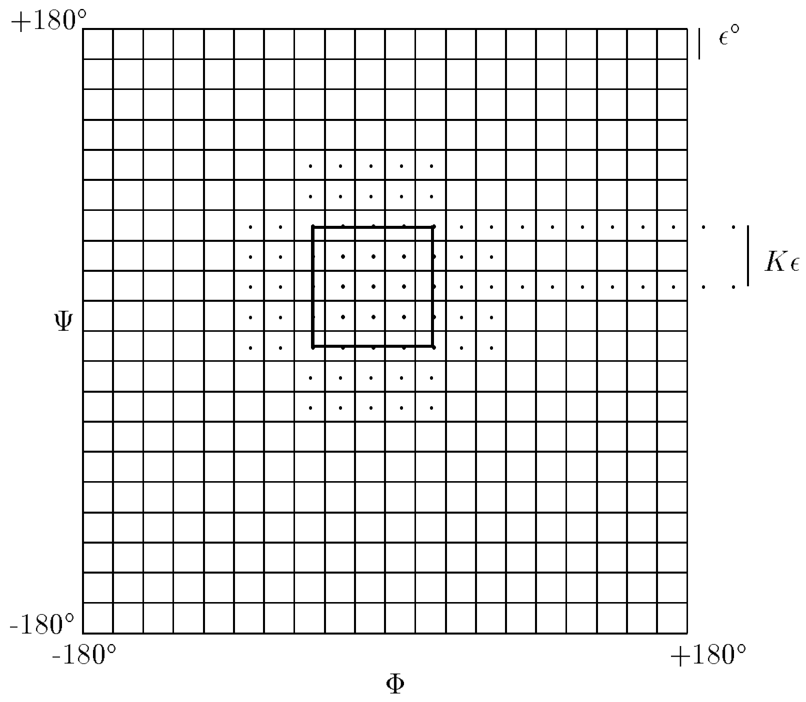


Figure 3.10: Carte de Ramachandran échantillonée : cliques maximales de la relation R = grands carrés de taille $K\epsilon$.

objets 3D. Diminuer les valeurs de ϵ et K conduit à des alphabets et des relations de tailles et indices de non-transitivité croissants respectivement (pour une maille de 5° par exemple l'alphabet possède 5182 symboles; et pour $K = 1$, g est égal à 9). En termes d'angles, la définition précédente signifie simplement que deux paires d'angles sont équivalentes si elles sont égales à plus ou moins ($\Delta\Phi = K\epsilon, \Delta\Psi = K\epsilon$). Bien que la relation R soit définie sur les ensembles de noeuds de la grille et non sur les paires d'angles elles-mêmes, elle possède la même caractéristique intrinsèquement non transitive de la relation de similarité entre angles et correspond à la relation associée à un graphe d'intervalles. Les cliques maximales de R sur Σ sont alors par définition tous les grands carrés de taille $K\epsilon$ centrés sur chaque noeud (voir la figure 3.10). Nous montrons plus loin comment cette relation R peut être étendue aux mots structuraux.

3.2.3.3 Couverture

3.2.3.3.1 Couverture sur l'alphabet des monomères

Initialement, le concept de couverture que nous utilisons ici a trouvé sa motivation dans la constatation qu'une seule relation n'est souvent pas suffisante pour capter toute la richesse des liens existants entre les acides aminés. La complexité de ces liens se trouve alors mieux illustrée

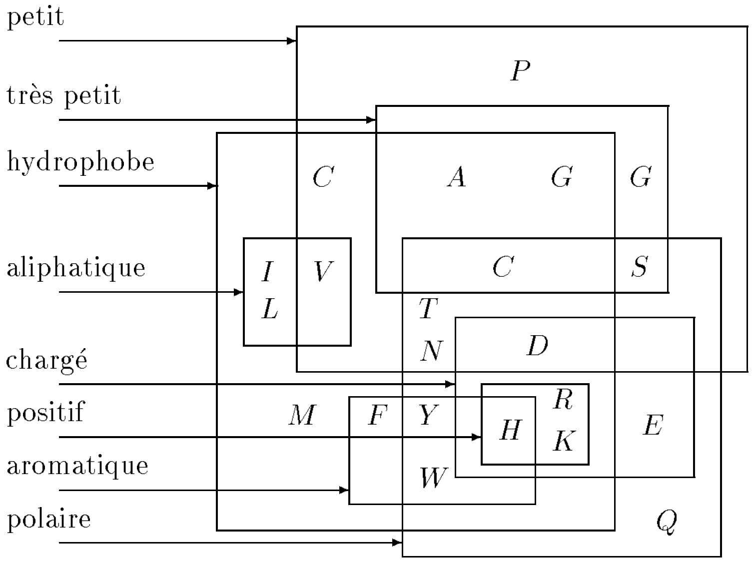


Figure 3.11: Diagramme Venn proposé par Taylor pour les acides aminés.

par un diagramme de Venn comme celui proposé par Taylor [Taylor, 1986a] et montré dans la figure 3.11. Ce diagramme a été établi à la fois à partir de données mutationnelles comme celles présentes dans une des matrices que nous avons décrites (dans son cas la matrice utilisée est celle de Dayhoff) et directement des propriétés physico-chimiques des acides aminés.

Une couverture est l'objet le mieux adapté pour modéliser ces liens [Sagot *et al.*, 1995a] [Sagot *et al.*, 1995d] [Sagot *et al.*, 1996].

Définition 3.2.13 *Soit Σ un alphabet, une couverture de Σ est un ensemble de sous-ensembles S_1, S_2, ... , S_p de Σ tels que :*

$$\bigcup_{i=1}^{p} S_i = \Sigma.$$

Sans recevoir un nom spécifique, cet objet a été utilisé par plusieurs personnes avant nous [Galas *et al.*, 1985] [Karlin and Ghandour, 1985b] [Karlin and Ghandour, 1985a] [Neuwald and Green, 1994] [Smith and Smith, 1990] [Waterman, 1984b] [Waterman *et al.*, 1984] [Waterman, 1989]. Dans tous les cas à l'exception de Neuwald, la couverture en question est en fait une partition.

Un exemple d'une couverture qui n'est pas une partition et qui s'inspire du diagramme de Taylor est donné dans la figure 3.12.

Observons qu'une couverture C peut être représentée par un hypergraphe sur l'alphabet Σ. Par ailleurs, nous parlons de l'indice de non-transitivité g d'une couverture de la même façon que nous avons parlé de l'indice de non-transitivité d'une relation R. Formellement :

Définition 3.2.14 *Soit C une couverture de Σ et $a \in \Sigma$, on note g_a le nombre d'ensembles de C auxquels a appartient.*

La couverture donnée ci-dessous est fortement inspirée du regroupement proposé par Taylor [Taylor, 1986a] et indiqué dans la figure 3.11.

$S_1 = \{A, C, G, S, T\}$ — très petit
$S_2 = \{A, C, D, G, N, S, T, V\}$ — petit
$S_3 = \{F, H, W, Y\}$ — aromatique
$S_4 = \{V, L, I, M\}$
$S_5 = \{K, R, H\}$ — basique
$S_6 = \{D, E\}$ — acide
$S_7 = \{Q, N\}$ — amide
$S_8 = \{P\}$ — proline

Observez que cette couverture ne représente ni une partition de Σ ni les cliques maximales d'une relation.

Figure 3.12: Exemple d'une couverture de l'alphabet des acides aminés.

Définition 3.2.15 *Soit C une couverture de Σ, on appelle g le nombre maximum d'ensembles de C auquel un symbole de Σ appartient, à savoir :*

$$g = Max \{g_a \mid a \in \Sigma\}.$$

On appelle par ailleurs $\overline{g}$ la valeur moyenne de g_a pour $a \in \Sigma$, c'est-à-dire :

$$\overline{g} = \frac{\sum_a g_a}{\mid \Sigma \mid}.$$

Dans l'exemple de la figure 3.12, g est égal à 2 et $\overline{g}$ à 1.45. g mesure ainsi la non-transitivité de la couverture par rapport à ce que nous pouvons appeler la couverture identité définie par :

Notation 3.2.2 *On note CI la couverture identité égale à $\{\{a\} \mid a \in \Sigma\}$.*

La couverture identité est celle habituellement employée pour les nucléotides mais ce n'est pas la seule ainsi que nous l'indiquons à la fin de cette section.

Lorsque C représente en fait une partition, sa non-transitivité est la même que celle de CI, c'est-à-dire 1.

Observons en outre qu'à nouveau ici la mesure de ressemblance est discrétisée car deux monomères sont considérés comme étant soit similaires s'il existe un ensemble de la couverture qui les contient tous deux, soit non similaires si cela n'est pas le cas. Comme pour les cliques, une couverture permet également une formalisation très satisfaisante de la ressemblance entre plus de deux objets à la fois. Le rapport existant entre clique et clique maximale d'un côté et couverture d'un autre va plus loin puisque l'ensemble des cliques maximales d'une relation forme une couverture.

La réciproque cependant n'est pas vraie. Le diagramme de Taylor en particulier ne peut être obtenu par simple seuillage d'une matrice comme c'est le cas pour la relation de la section précédente puisqu'en fait il exprime plusieurs relations à la fois. Il permet ainsi par exemple de caractériser un ensemble d'acides aminés à deux niveaux différents de généralité, les plaçant d'abord dans une même unité qui ne contient qu'eux, puis les groupant avec un ou plusieurs autres acides aminés qui possèdent les mêmes attributs que les éléments de l'ensemble plus petit, plus quelques autres qui leur sont propres.

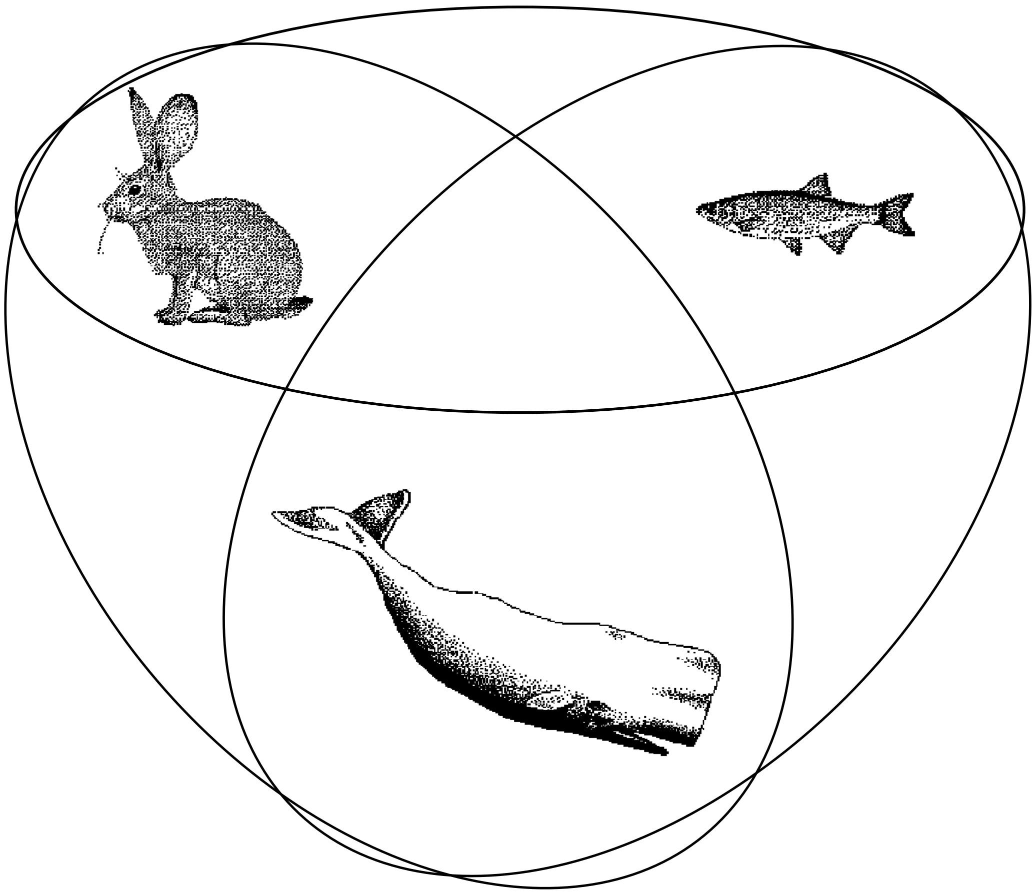


Figure 3.13: Points de vue différents.

L'information captée par une couverture peut être encore plus subtile. Illustrons cela par un exemple. Imaginons que l'on veuille comparer non plus les monomères d'une macromolécule biologique, mais certains animaux, par exemple des lapins, des baleines et des sardines. Pour un zoologiste, un lapin est plus proche d'une baleine que les deux d'une sardine car tous deux sont des mammifères et l'attribut 'être un mammifère' est important en zoologie. Pour le patron d'une entreprise alimentaire par contre, il est sans doute plus cohérent de classer la baleine avec la sardine car bien que les trois animaux soient (en principe) comestibles, la baleine et la sardine proviennent du même milieu, la mer, et c'est sans doute un des facteurs les plus importants pour lui en termes de l'infrastructure dont il a besoin pour installer son industrie (traiter des lapins exige d'autres moyens que ceux nécessaires pour traiter les produits de la pêche). Finalement, un enfant souhaitant adopter un animal n'hésitera pas à classer le lapin avec la sardine car il sait que tous deux peuvent entrer dans son appartement, une baleine non. Ces points de vue divers donnent ainsi lieu à trois classifications (voir la figure 3.13) qu'une couverture reproduit fidèlement mais que le concept de clique maximale gommerait.

Cette image sert donc à illustrer encore une fois le fait que les critères permettant de mesurer la ressemblance de deux objets dépendent de ce que nous recherchons. Ces critères sont très probablement différents suivant que nous désirons identifier par exemple l'emplacement des sites actifs d'une protéine, ou des zones d'interaction sur l'ADN et l'ARN, ou encore établir les facteurs à partir desquels une chaîne polymérique se replie dans l'espace pour produire une structure. Ils sont différents aussi suivant le temps depuis lequel les molécules que nous voulons comparer ont divergé à partir d'un ancêtre commun. Le problème est que ces critères ne sont pas toujours connus, et ceux qui le sont n'ont probablement pas un caractère définitif.

Parfois alors il faudrait pouvoir aller plus loin et jouer sans *a priori* avec plusieurs points de vue différents à la fois, aussi bien dans le cas de l'ADN et de l'ARN que dans celui des protéines. Un premier outil pour réaliser cela va constituer ce que nous désignons sous le nom de couverture combinatoire pondérée. Dans sa définition la plus complète, une couverture combinatoire pondérée CCP est simplement l'ensemble des parties de l'alphabet avec une pondération affectée à chacun de ces sous-ensembles. La signification précise de ces contraintes deviendra claire lorsque nous passerons de la comparaison des perles à celles des colliers. Pour l'instant, il suffit de dire qu'elles sont nécessaires afin d'éviter que tout ou presque tout finisse par se ressembler. Nous laissons donc pour plus tard leur définition et une discussion des limites conceptuelles de cette approche.

Observons cependant dès maintenant que, même lorsque nous aurons établi des contraintes sur ces couvertures combinatoires, la souplesse qu'elles autorisent vont exiger aussi des raffinements afin de nous permettre de les utiliser d'une manière qui soit algorithmiquement efficace. Ces raffinements sont formellement intéressants en eux-mêmes et peut-être extensibles à des problèmes d'une autre nature. Ils sont présentés au chapitre 5.

3.2.3.3.2 Couverture sur l'alphabet des objets 3D

Notons simplement ici que le codage de la structure d'une macromolécule en tant que chaîne sur des symboles qui représentent les numéros des carrés d'une grille discrétisants des paires d'angles peut *a priori* conduire à une mesure de la ressemblance utilisant le concept d'une couverture. Les éléments de cette couverture seraient comme auparavant des ensembles de carrés (numéros) proches les uns des autres mais comme il n'y a plus ici de notion de maximalité, ces ensembles pourraient être représentés non seulement par des grands carrés de taille $K\epsilon$ mais également par n'importe quel autre découpage de la grille.

3.2.4 La ressemblance des colliers de perles revue

3.2.4.1 Valeur d'un alignement

3.2.4.1.1 Alignement sans trous

Nous allons vouloir à présent passer d'une mesure de la ressemblance entre perles à une mesure de la ressemblance entre colliers alignés.

Commençons par supposer que nous avons réalisé un alignement de deux molécules et, pour simplifier, considérons pour l'instant le cas où cet alignement ne comporte pas de trous. Quelle valeur lui affecter?

Il y a deux façons de procéder qui représentent en fait deux philosophies différentes de la mesure de ressemblance mais qui peuvent être formalisées de manière unique puisque dans les deux cas, la valeur attribuée à l'alignement, appelée un score, est égale à la somme des scores affectés aux paires alignées. Ces derniers cependant peuvent représenter soit une valeur numérique, entière ou réelle, trouvée dans une matrice de substitution, soit une valeur booléenne dans le cas d'une relation ou d'une couverture.

Notation 3.2.3 *Soient $a_1a_2...a_n$ et $b_1b_2...b_n$ deux chaînes sur un alphabet Σ. La suite définie sur $\Sigma \times \Sigma$ égale à $(a_1, b_1)(a_2, b_2)...(a_n, b_n)$ et notée $\begin{pmatrix} a_1 \\ b_1 \end{pmatrix} \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} ... \begin{pmatrix} a_n \\ b_n \end{pmatrix}$ représente l'alignement sans trous des deux chaînes (dans ce cas un seul alignement est possible).*

Formellement nous avons :

Définition 3.2.16 *Soit $\begin{pmatrix} a \\ b \end{pmatrix}$ un alignement de deux symboles de Σ. Alors la valeur de score $\begin{pmatrix} a \\ b \end{pmatrix}$ est donnée par :*

- *$\mathcal{M}(a, b)$ pour une matrice de substitution $\mathcal{M}$;*
- $\begin{cases} 1 \; si \; aR\,b \\ 0 \; sinon \end{cases}$ *pour une relation de similarité R;*
- $\begin{cases} 1 \; si \; \exists S \in C \; tel \; que \; a, b \in S \\ 0 \; sinon \end{cases}$ *pour une couverture C de l'alphabet.*

Définition 3.2.17 *Soit $\mathcal{A} = \begin{pmatrix} a_1 \\ b_1 \end{pmatrix} \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} ... \begin{pmatrix} a_n \\ b_n \end{pmatrix}$ un alignement des deux chaînes (définies sur un alphabet Σ) $a_1a_2...a_n$ et $b_1b_2...b_n$. Alors :*

$$score\mathcal{A} = \sum_{i=1}^{n} score \begin{pmatrix} a_i \\ b_i \end{pmatrix}.$$

Exemple 3.2.1 *Soient les deux chaînes $s1$ et $s2$ définies sur l'alphabet $\Sigma = \{a, b, c, d\}$ et données par :*
$s1 = abcbdcb$
$s2 = acbbdca$.

Supposons que $\mathcal{M}_S = \begin{pmatrix} 3 & -1 & 0 & -2 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ -2 & 0 & 0 & 4 \end{pmatrix}$, *$R = I$ (voir la notation 3.2.1), $C = CI$ (voir la notation 3.2.2).*

Alors le score de l'alignement $\mathcal{A} = \begin{pmatrix} a \\ a \end{pmatrix} \begin{pmatrix} b \\ c \end{pmatrix} \begin{pmatrix} c \\ b \end{pmatrix} \begin{pmatrix} b \\ b \end{pmatrix} \begin{pmatrix} d \\ d \end{pmatrix} \begin{pmatrix} c \\ c \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix}$ *sans trous de $s1$ et $s2$ est égal à 7 si nous utilisons une matrice, et à 4 si nous travaillons avec une relation ou une couverture.*

Si la matrice utilisée est une matrice de dissimilarité (resp. de distance), alors le score de $\mathcal{A}$ représente en fait une dissimilarité (resp. une distance) entre les deux chaînes. Dans le cas d'une mesure entre monomères ayant pour base une relation ou une couverture et où les valeurs données dans la définition 3.2.16 sont inversées (i.e. 0 si $a \; R \; b$ ou si $\exists \; S \in C$ tel que $a, \, b \in S$, 1 sinon), le score de $\mathcal{A}$ indique une distance. Si de plus $R = I$ ou $C = CI$, cette distance représente la distance de Hamming entre les deux chaînes.

Exemple 3.2.2 *La distance de Hamming entre les chaînes $s1$ et $s2$ de l'exemple 3.2.1 est égale à 3.*

La définition précédente peut être réécrite de manière récursive. Cette façon de faire va nous être particulièrement utile par la suite. En effet, nous allons voir que la récursion est à la base de presque toutes les méthodes d'exploration de la ressemblance que nous présentons ici.

Définition 3.2.18 *Soit* $\mathcal{A} = \begin{pmatrix} a_1 \\ b_1 \end{pmatrix} \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} \ldots \begin{pmatrix} a_n \\ b_n \end{pmatrix}$ *un alignement des deux chaînes (définies sur un alphabet* Σ*)* $a_1a_2...a_n$ *et* $b_1b_2...b_n$ *avec* $n \geq 1$*. Alors :*

$$score\mathcal{A} = score\mathcal{A}' + score \begin{pmatrix} a_n \\ b_n \end{pmatrix}$$

où $\mathcal{A}' = \begin{pmatrix} a_1 \\ b_1 \end{pmatrix} \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} \ldots \begin{pmatrix} a_{n-1} \\ b_{n-1} \end{pmatrix}$.

Pourquoi avons-nous dit qu'il y avait en fait ici deux philosophies différentes de la mesure de ressemblance entre deux objets suivant que nous travaillons avec une matrice ou avec une relation ou couverture? Prenons l'exemple des examens que doivent passer les étudiants. La réussite à un cours peut être déterminée d'au moins deux manières : soit une moyenne (éventuellement pondérée) est établie des notes obtenues sur l'ensemble des disciplines que l'étudiant doit suivre et ce dernier est reçu si cette moyenne est au-dessus de 10, soit pour être reçu celui-ci doit obtenir une note au-dessus de 10 dans au moins un certain nombre x de disciplines. Tout le monde sait pourquoi la seconde façon de faire peut être plus sévère : il n'y a pas de rattrapage possible des notes. Si vous obtenez 19 dans une discipline, vous l'avez bien sûr réussie, mais vous ne l'avez pas plus réussie que si vous aviez obtenu 10. Et si vous obtenez 9, vous l'avez tout aussi bien ratée que si vous aviez obtenu 0.

Le résultat est le même dans le cas qui nous concerne et la première manière de procéder, utilisant une matrice, est ainsi plus flexible que la seconde. Il est important d'observer que deux facteurs sont en jeu ici : l'effet de compensation et l'effet de pondération. Le second permet d'attribuer des poids différents à une même propriété suivant les objets qui la vérifient. Par exemple, l'appariement de deux monomères identiques dans un alignement peut recevoir un score variable suivant la nature du monomère en question. Un acide aminé plus rare a ainsi un score d'alignement avec lui-même plus élevé indiquant son plus faible taux de mutabilité. L'effet de compensation quant à lui est une façon détournée de prendre en considération le contexte d'un appariement, du moins en partie. Il permet ainsi, selon la définition de ressemblance établie, de moduler le score d'un mauvais appariement par les scores des appariements avoisinants dans un alignement. Cette modulation n'est plus possible lorsque l'on travaille avec une relation R ou une couverture C. Une autre façon de voir cela est de considérer que travailler avec ces deux concepts revient à travailler avec un système de score qui attribue la valeur 1 à un appariement de deux monomères en relation par R (ou qui appartiennent à un même sous-ensemble de C) et $-\infty$ sinon. Les deux effets, de pondération et de compensation, sont illustrés dans la figure 3.14 dans le cas d'une chaîne de caractère. Cette plus grande flexibilité permise par l'utilisation d'un système de score peut cependant être aussi plus délicate à manier et à interpréter. Les deux approches sont en fait valables, et nous laissons pour le chapitre consacré aux illustrations une discussion de leurs mérites respectifs ainsi que des problèmes que chacune présente.

Observons toutefois avant de clore cette section qu'il n'est en fait pas obligatoire d'attribuer la valeur 1 à une paire de symboles en relation ou appartenant à un même ensemble d'une couverture. Nous pourrions en effet utiliser une valeur w_i qui dépendrait de la clique maximale C_i de R ou de l'ensemble S_i de la couverture C. Notons que, dans le cas d'une relation, ces scores w_i seraient définis après l'établissement de la relation, qui elle-même aurait pu être obtenue à

Soit $\Sigma = \{a, b, c\}$ et soient :

$$\mathcal{M}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ et } \mathcal{M}_2 = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}.$$

Si nous établissons que deux chaînes sont similaires si :

- le nombre d'identités est ≥ 5;

ou

- le score d'un alignement sans trou des deux est ≥ 5;

alors par le premier critère les chaînes *aacbcba* and *aaabaaa* ne sont pas similaires tandis que par le second elles le sont que la matrice utilisée soit $\mathcal{M}_1$ ou $\mathcal{M}_2$. La première illustre l'effet de pondération appliqué sur une même propriété (ici l'identité de deux monomères) et la seconde l'effet de compensation, les 3 mauvais appariements étant 'rattrapés' localement par l'alignement des paires identiques.

Figure 3.14: Exemple de la plus grande flexibilité que permet l'utilisation de scores sur un simple comptage du nombre d'erreurs ou d'identités.

partir du seuillage d'une matrice de scores. Les scores w_i et les scores de la matrice servant éventuellement à définir R ou C n'ont ainsi pas besoin d'être identiques. L'établissement de poids w_i dépendants de C_i ou de S_i peuvent alors introduire un effet de pondération et de compensation comme dans le cas des matrices, mais la nature de ces effets est différente. Celui de pondération interviendrait au niveau de groupes d'objets (des cliques maximales ou non) plutôt qu'au niveau des objets individuels. Celui de compensation quant à lui ne serait observé que si un score négatif mais non infini était attribué aux paires de symboles non en relation par R ou appartenant à des éléments de C différents. Ces scores négatifs pourraient eux-mêmes dépendre de quelles cliques de R ou éléments de C sont impliqués par la paire appariée. Il n'est pas du tout clair cependant comment les poids w_i ainsi que les poids de mésappariement pourraient être déterminés. Aucune théorie n'existe dans ce sens, même sujette à précaution comme celle associée à la construction des matrices. Relation et couverture sont donc utilisés de manière discrétisée et le score que nous employons en pratique est celui indiqué dans la définition 3.2.16.

Nous allons faire l'hypothèse maintenant que nous disposons d'un alignement des deux molécules qui présente des trous. Voyons comment lui attribuer une valeur.

3.2.4.1.2 Alignement avec trous

Un trou dans un alignement de deux molécules ayant même origine signifie qu'une mutation correspondant soit à une délétion, soit à une insertion a eu lieu. De tels événements sont considérés plus rares que les substitutions non pas parce que les délétions et insertions surviennent moins souvent, mais parce qu'elles sont observées moins souvent. Nous avons vu dans le chapitre précédent que ces mutations en effet conduisent plus fréquemment à une perte d'activité de la macromolécule. Toute la difficulté d'attribuer une valeur aux trous présents dans un alignement vient alors de cette constatation. En effet, les méthodes qui essayent de déterminer quelle pondération il faudrait affecter à un trou sont soit *ad hoc*, soit emploient les mêmes approches que celles utilisées pour calculer certaines matrices de substitution : c'est-à-dire, par une observation des délétions ou insertions survenues entre macromolécules homologues ([Benner *et al.*, 1993] [Pascarella and Argos, 1992] pour les protéines). C'est comme si pour estimer la qualité de l'alimentation des jeunes français des années trente, on décidait de mesurer l'état de santé des personnes âgées entre soixante-quinze et quatre-vingt-dix ans aujourd'hui. Même si nous laissons de côté la question de comment effectivement quantifier une telle qualité qui fasse la part des multiples autres facteurs pouvant avoir une influence sur la santé des individus, une telle mesure ne peut être qu'un pauvre reflet de la qualité réelle de cette alimentation puisqu'elle néglige toutes les personnes décédées entre temps pour diverses raisons (dont quelqu'unes peuvent être justement liées à leur alimentation).

Outre la difficulté que nous venons d'illustrer, et qui est insurmontable, l'attribution d'une valeur à un trou dans un alignement pose deux autres types de problèmes. Le premier est semblable à celui rencontré pour l'affectation d'une mesure à la ressemblance de deux monomères. Il n'est pas possible d'attribuer une seule valeur à un trou quel que soit son contexte. Dans le cas d'une protéine par exemple, une délétion ou insertion observée à l'intérieur d'une structure secondaire ou dans une région correspondant à un site ne peut pas avoir le même poids qu'une délétion ou insertion observée ailleurs, soit à une des extrémités de la molécule, soit dans les parties moins structurées ou qui ne sont pas impliquées dans son activité. Le poids doit varier également suivant les caractéristiques physico-chimiques de l'acide aminé qui a été délété ou inséré. L'insertion d'un 'gros' acide aminé par exemple peut en effet rompre localement ou globalement la structure d'une chaîne, alors que celle d'un 'petit' résidu comme la glycine ne modifie éventuellement que peu cette structure et peut être tolérée. Toute la difficulté réside à nouveau ici dans le fait que ce contexte est rarement connu.

L'attribution d'une valeur à un trou pose un second problème qui concerne sa longueur cette fois. Si une substitution est un événement ponctuel, une délétion ou une insertion implique souvent plusieurs résidus en même temps. Il n'est pas nécessairement question ici de corrélations entre mutations (bien que cela puisse également être le cas) mais d'un unique événement mutationnel affectant une région plus ou moins étendue. Ceci correspond donc à une suite de trous dans un alignement. Il n'est pas admissible d'attribuer une valeur constante à chacun d'eux puisqu'ils se réfèrent à un même phénomène survenu une seule fois. Ici aussi, plusieurs manières de procéder ont été proposées, la plus usuelle consiste à pénaliser plus fortement le premier trou. Le score d'une suite de trous est ainsi une fonction affine de sa longueur, c'est-à-dire est égale à $ak + b$ où k est la longueur de la suite, a et b deux constantes [Altschul and Erickson, 1986] [Gotoh, 1982]. Un cas particulier de cette fonction est obtenue pour $a = 0$, le score d'un trou est alors lui-même une constante [Sellers, 1974] [Ukkonen, 1985a]. D'autres fonctions ont été imaginées, par exemple, des fonctions affines par morceaux [Gotoh, 1990] [Miller and Myers, 1988] où la valeur de a change par intervalles, ou encore des fonctions concaves [Fitch and Smith, 1983] [Galil and Giancarlo, 1989] [Miller and Myers, 1988] [Miller and Myers, 1988] [Waterman, 1984a]. Gonnet et Benner ont toutefois suggéré que les fonctions affines ne seraient pas un modèle valable pour les alignements protéiques. D'après les résultats de leur analyse de la distribution des trous dans les alignements de protéines homologues [Benner *et al.*, 1993], le poids à accorder à l'ouverture d'un trou varie avec la distance évolutive des protéines tandis que celui à affecter à l'allongement d'un trou, bien qu'indépendant de cette distance, augmente non pas linéairement avec la longueur mais en fonction de cette longueur élevée à la puissance 3/2. Ces résultats ne correspondent cependant pas tout à fait à ceux obtenus par Pascarella [Pascarella and Argos, 1992].

Nous voyons ainsi que les controverses autour de la valeur à attribuer à l'introduction d'un trou ou à son allongement sont encore plus grandes que celles concernant les substitutions. Ces controverses sont importantes, il a été montré que la qualité d'un alignement dépend en effet plus fortement du choix des pénalités affectées aux trous que de celui des scores de substitutions [Gonnet *et al.*, 1992]. Différents travaux ont porté aussi sur les rapports existant entre ce choix et celui du meilleur alignement qu'il est possible d'obtenir entre deux chaînes [Gusfield *et al.*, 1992] [Vingron and Waterman, 1994] [Waterman *et al.*, 1992].

Cependant notre propos n'est pas de les discuter ici. La raison principale en est que nous nous intéressons plus dans ce travail aux mots qu'aux chaînes entières, et que ces mots représentent comme nous l'avons vu les régions les mieux conservées des macromolécules. Il n'est pas abusif de considérer que, dans ce cas, le choix d'un score à attribuer aux trous n'est pas aussi crucial. En effet, la présence de trous y est observée moins fréquemment que sur le reste de la molécule, et ces trous correspondent le plus souvent à des événements mutationnels ponctuels, les délétions ou insertions de blocs entiers se produisant plutôt entre ces régions. Compter simplement chacun d'eux ou leur attribuer un poids constant apparaît en grande partie justifié dans ce cadre [Altschul, 1989]. Nous allons donc admettre ce choix pour l'instant et voir quel est alors le score d'un alignement avec trous.

Notation 3.2.4 *Soient $a_1a_2...a_n$ et $b_1b_2...b_m$ deux chaînes sur un alphabet Σ avec $n \leq m$ et soit $\overline{\Sigma} = \Sigma \cup \{\lambda\}$. La suite définie sur $\overline{\Sigma} \times \overline{\Sigma}$ égale à $\begin{pmatrix} \overline{a}_1 \\ \overline{b}_1 \end{pmatrix} \begin{pmatrix} \overline{a}_2 \\ \overline{b}_2 \end{pmatrix} ... \begin{pmatrix} \overline{a}_p \\ \overline{b}_p \end{pmatrix}$ avec $m \leq p \leq n+m$ et telle que :*

- $\overline{a}_i = a_j$ *ou* $\overline{a}_i = \lambda$ *pour* $1 \leq i \leq p$ *et* $1 \leq j \leq n$
- $\overline{b}_i = b_j$ *ou* $\overline{b}_i = \lambda$ *pour* $1 \leq i \leq p$ *et* $1 \leq j \leq m$
- $\mid \overline{a}_1...\overline{a}_p \mid = n$ *et* $\overline{a}_1...\overline{a}_p = a_1...a_n$
- $\mid \overline{b}_1...\overline{b}_p \mid = m$ *et* $\overline{b}_1...\overline{b}_p = b_1...b_m$
- *pour tout* $i \in \{1,...,p\}$, *il n'est jamais vrai que* $\overline{a}_i = \lambda = \overline{b}_i$ *(pas de trou mis face à un trou)*

représente un alignement de $a_1a_2...a_n$ avec $b_1b_2...b_m$.

Exemple 3.2.3 *Soient les deux chaînes $s1$ et $s2$ définies sur l'alphabet $\{a, b, c, d\}$ et données par :*
$s1 = abcbadcb$
$s2 = acbbdca$.
Alors $\begin{pmatrix} a \\ a \end{pmatrix} \begin{pmatrix} b \\ \lambda \end{pmatrix} \begin{pmatrix} c \\ c \end{pmatrix} \begin{pmatrix} b \\ b \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \begin{pmatrix} d \\ d \end{pmatrix} \begin{pmatrix} c \\ c \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix}$ *est un alignement de $s1$ et $s2$.*

Si l'on considère la dernière paire d'un alignement $\begin{pmatrix} \overline{a}_1 \\ \overline{b}_1 \end{pmatrix} \begin{pmatrix} \overline{a}_2 \\ \overline{b}_2 \end{pmatrix} ... \begin{pmatrix} \overline{a}_p \\ \overline{b}_p \end{pmatrix}$ de deux chaînes $a_1a_2...a_n$ et $b_1b_2...b_m$, trois cas de figures sont alors possibles :

- $\begin{pmatrix} \overline{a}_p \\ \overline{b}_p \end{pmatrix} = \begin{pmatrix} a_n \\ b_m \end{pmatrix}$
- $\begin{pmatrix} \overline{a}_p \\ \overline{b}_p \end{pmatrix} = \begin{pmatrix} a_n \\ \lambda \end{pmatrix}$

- $\begin{pmatrix} \overline{a}_p \\ \overline{b}_p \end{pmatrix} = \begin{pmatrix} \lambda \\ b_m \end{pmatrix}$.

Nous utilisons ce fait pour donner une définition récursive du score d'un alignement $\mathcal{A}$.

Définition 3.2.19 *Soit* $\begin{pmatrix} \overline{a} \\ \overline{b} \end{pmatrix}$ *un alignement de deux symboles de* $\Sigma \cup \{\lambda\}$ *dont l'un des deux au moins est différent de* λ. *Alors, score* $\begin{pmatrix} \overline{a} \\ \overline{b} \end{pmatrix}$ *est égal à :*

- $\begin{cases} \mathcal{M}(a,b) \text{ si } \overline{a} = a \neq \lambda \text{ et } \overline{b} = b \neq \lambda \\ w_\lambda \text{ sinon} \end{cases}$ *pour une matrice* $\mathcal{M}$ *et une constante* w_λ

  (w_λ représente la pénalité à attribuer à un trou);
- $\begin{cases} 1 \text{ si } \overline{a} = a \neq \lambda \text{ et } \overline{b} = b \neq \lambda \text{ et } aR\,b \\ 0 \text{ sinon} \end{cases}$ *pour une relation de similarité R;*
- $\begin{cases} 1 \text{ si } \overline{a} = a \neq \lambda \text{ et } \overline{b} = b \neq \lambda \text{ et } \exists S \in C \text{ tel que } a,b \in S \\ 0 \text{ sinon} \end{cases}$ *pour une couverture C.*

Définition 3.2.20 *Soit* $\mathcal{A} = \begin{pmatrix} \overline{a}_1 \\ \overline{b}_1 \end{pmatrix} \begin{pmatrix} \overline{a}_2 \\ \overline{b}_2 \end{pmatrix} ... \begin{pmatrix} \overline{a}_p \\ \overline{b}_p \end{pmatrix}$ *un alignement des chaînes* $a_1a_2...a_n$ *et* $b_1b_2...b_m$. *Alors :*

$$score\mathcal{A} = score\mathcal{A}' + score \begin{pmatrix} \overline{a}_p \\ \overline{b}_p \end{pmatrix}$$

où $\mathcal{A}' = \begin{pmatrix} \overline{a}_1 \\ \overline{b}_1 \end{pmatrix} \begin{pmatrix} \overline{a}_2 \\ \overline{b}_2 \end{pmatrix} ... \begin{pmatrix} \overline{a}_{p-1} \\ \overline{b}_{p-1} \end{pmatrix}$.

Exemple 3.2.4 *Soient les deux chaînes s1 et s2 définies sur l'alphabet* $\Sigma = \{a,b,c,d\}$ *et données par :*

s1 = abcbadcb

s2 = acbbdca.

Supposons que $\mathcal{M}_S = \begin{pmatrix} 3 & -1 & 0 & -2 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ -2 & 0 & 0 & 4 \end{pmatrix}$, $R = I$ *(voir la notation 3.2.1),* $C = CI$ *(voir la notation 3.2.2) et* w_λ = -1, *Alors le score de* $\mathcal{A} = \begin{pmatrix} a \\ a \end{pmatrix} \begin{pmatrix} b \\ \lambda \end{pmatrix} \begin{pmatrix} c \\ c \end{pmatrix} \begin{pmatrix} b \\ b \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \begin{pmatrix} d \\ d \end{pmatrix} \begin{pmatrix} c \\ c \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix}$ *est égal à 8 si nous utilisons une matrice et à 5 si nous travaillons avec une relation ou une couverture.*

Dans le cas d'une relation ou d'une couverture, il est habituel de travailler avec une mesure de dissimilarité plutôt qu'avec une mesure de similarité, et ainsi la valeur d'un alignement est donné en inversant les rôles de 0 et 1 dans la définition 3.2.19. Lorsque $R = I$ ou $C = CI$ et que score $\mathcal{A} \leq$ score $\mathcal{B}$ pour tout alignement $\mathcal{B}$ de $a_1a_2...a_n$ avec $b_1b_2...b_m$, alors le score de $\mathcal{A}$ mesure en fait la distance d'édition classique que nous avons mentionnée au début de ce chapitre. Elle correspond bien au plus petit nombre d'opérations de substitutions, délétions et insertions nécessaires pour passer d'une chaîne à l'autre. Nous appelons aussi cette distance la distance de Levenshtein [Levenshtein, 1966] séparant deux chaînes de caractères et nous utilisons indifféremment les deux termes dans le reste de ce travail.

Exemple 3.2.5 *La distance de Levenshtein des chaînes $s1$ et $s2$ de l'exemple 3.2.4 est égale à* 3.

3.2.4.2 Mesure d'une ressemblance : la valeur du meilleur alignement

Nous venons de montrer comment attribuer un score à un alignement. Une mesure de la ressemblance entre objets décomposables en unités élémentaires sur lequel un ordre total a pu être établi est alors classiquement donnée par le plus grand (ou le plus petit) score obtenu parmi tous les alignements possibles de ces objets (suivant que l'on considère une similarité ou une dissimilarité entre eux).

Définition 3.2.21 *Soient $s1$ et $s2$ deux chaînes. Soit $VR(s1, s2)$ la valeur de similarité (resp. dissimilarité) entre $s1$ et $s2$, alors :*

$$VR(s1, s2) = max\ (resp.min)\ \{score\mathcal{A} \mid \mathcal{A}\ un\ alignement\ de\ s1\ et\ s2\}.$$

Exemple 3.2.6 *En reprennant les données de l'exemple 3.2.4, alors $VR(s1, s2)$ est égal à* 8 *ou* 5 *suivant que nous travaillons avec une matrice ou avec une relation/couverture puisque l'alignement* $\mathcal{A} = \begin{pmatrix} a \\ a \end{pmatrix} \begin{pmatrix} b \\ \lambda \end{pmatrix} \begin{pmatrix} c \\ c \end{pmatrix} \begin{pmatrix} b \\ b \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \begin{pmatrix} d \\ d \end{pmatrix} \begin{pmatrix} c \\ c \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix}$ *sans trou représente un meilleur alignement de $s1$ et $s2$.*

Cette mesure correspond ainsi au score du meilleur alignement de $s1$ et $s2$ et est appelé score optimal des deux chaînes. Dans le cas de l'exemple ci-dessus, $\mathcal{A}$ représente le seul alignement optimal de $s1$ et $s2$. Mais il est clair que deux chaînes $s1$ et $s2$ peuvent posséder plusieurs alignements optimaux. Un exemple est donné ci-dessous :

Exemple 3.2.7 *Soient les deux chaînes $s1$ et $s2$ définies sur l'alphabet $\Sigma = \{a, b, c\}$ et données par :*

$s1 = abca$

$s2 = acba.$

Supposons que $\mathcal{M}$ soit la matrice identité et w_λ = -$\frac{1}{2}$. Alors : $\mathcal{A} = \begin{pmatrix} a \\ a \end{pmatrix} \begin{pmatrix} b \\ c \end{pmatrix} \begin{pmatrix} c \\ b \end{pmatrix} \begin{pmatrix} a \\ a \end{pmatrix}$ *et* $\mathcal{B} = \begin{pmatrix} a \\ a \end{pmatrix} \begin{pmatrix} \lambda \\ c \end{pmatrix} \begin{pmatrix} b \\ b \end{pmatrix} \begin{pmatrix} c \\ \lambda \end{pmatrix} \begin{pmatrix} a \\ a \end{pmatrix}$ *sont deux alignements optimaux (de score* 2*) de $s1$ et $s2$.*

Dans sa forme présente, la mesure de la ressemblance a porté sur deux objets uniquement qui sont comparés directement entre eux. La comparaison d'objets peut se faire d'autres façons. C'est ce que nous allons montrer en commençant par parler des méthodes locales qui vont nous permettre de repérer les objets qui nous intéressent : les mots communs à un ensemble de chaînes.

3.3 Méthodes de comparaison

3.3.1 Comparaison globale ou locale

Dans les mesures de ressemblance que nous avons présentées jusqu'à maintenant, c'est dans leur totalité que nous comparions les objets. Or nous avons montré dans le chapitre 2 que c'est localement surtout que ces objets vont nous intéresser.

La manière classique de procéder pour effectuer des comparaisons locales de chaînes demeure cependant essentiellement la même : par un alignement implicite ou explicite des objets à comparer.

Introduisons donc d'abord une notation concernant l'alignement de deux mots dans une chaîne.

Définition 3.3.1 *Soient $s1 = a_1...a_n$ et $s2 = b_1...b_m$ deux chaînes et soient (i, j) et (k, l) deux paires d'entiers tels que $1 \leq i < j \leq n$, $1 \leq k < l \leq m$ et $j - i \leq l - k$. Ces paires définissent ainsi deux mots $u1 = a_i...a_j$ et $u2 = b_k...b_l$ de $s1$ et $s2$ respectivement. Alors l'alignement $\begin{pmatrix} \overline{a}_1 \\ \overline{b}_1 \end{pmatrix} \begin{pmatrix} \overline{a}_2 \\ \overline{b}_2 \end{pmatrix} ... \begin{pmatrix} \overline{a}_p \\ \overline{b}_p \end{pmatrix}$ avec $(j - i + 1) \leq p \leq (j - i + 1) + (l - k + 1)$ tel que :*

- *$\overline{a}_r = a_s$ ou $\overline{a}_r = \lambda$ pour $1 \leq r \leq p$ et $i \leq s \leq j$*
- *$\overline{b}_r = b_s$ ou $\overline{b}_r = \lambda$ pour $1 \leq r \leq p$ et $k \leq s \leq l$*
- *$\mid \overline{a}_i...\overline{a}_{i+p} \mid = j - i + 1$ et $\overline{a}_1...\overline{a}_p = a_i...a_j$*
- *$\mid \overline{b}_k...\overline{b}_{k+p} \mid = l - k + 1$ et $\overline{b}_1...\overline{b}_p = b_k...b_l$*
- *pour tout $r \in \{1, ..., p\}$, il n'est jamais vrai que $\overline{a}_r = \lambda = \overline{b}_r$*

est un alignement des mots $u1$ et $u2$.

Exemple 3.3.1 *Soient $\Sigma = \{a, b, c, d\}$ et les deux chaînes définies sur cet alphabet données ci-dessous:*
$s1 = abcbdcb$
$s2 = acbbdca$.
Alors $\begin{pmatrix} b \\ b \end{pmatrix} \begin{pmatrix} d \\ d \end{pmatrix} \begin{pmatrix} \lambda \\ c \end{pmatrix}$ est un alignement du mot $u1 = bd$ de $s1$ avec le mot $u2 = bdc$ de $s2$.

Le score de l'alignement des mots $u1$ et $u2$ est défini exactement de la même façon que le score des deux chaînes $s1$ et $s2$ dont ils font partie, et à nouveau ici la mesure de ressemblance entre $u1$ et $u2$ correspond traditionnellement à celle du meilleur alignement possible des deux mots, c'est-à-dire, celui ayant un score optimal. Le score de l'alignement de deux mots peut représenter aussi en fait, comme dans le cas des chaînes, une distance, par exemple celle de Levenshtein.

Exemple 3.3.2 *La distance de Levenshtein séparant les deux mots bd et bdc de l'exemple 3.3.1 est égale à 1.*

Nous allons alors nous intéresser surtout à identifier les mots qui se ressemblent suffisamment dans ces chaînes car ce sont ceux-là que nous allons vouloir définir comme étant similaires. Avec tout ce que nous venons d'établir, cette identification est immédiate :

Définition 3.3.2 *Étant donné deux chaînes* $s1 = a_1...a_n$ *et* $s2 = b_1...b_m$ *et une constante seuil, deux mots* $u1$ *et* $u2$ *de* $s1$ *et* $s2$ *respectivement sont considérés comme étant similaires si* $score\mathcal{A} \geq seuil$ *avec* $\mathcal{A}$ *un meilleur alignement de* $u1$ *et* $u2$ *(ou* $score\mathcal{A} \leq seuil$ *si le score d'un alignement mesure une dissimilarité ou une distance).*

Observons enfin que lorsque nous travaillons avec une mesure de similarité, il est possible d'attribuer une valeur à la ressemblance locale de deux chaînes. Celle-ci est en effet donnée par le plus grand score obtenu parmi tous les alignements (optimaux) possibles de mots dans ces chaînes [Smith and Waterman, 1981].

Définition 3.3.3 *Soient* $s1$ *et* $s2$ *deux chaînes et un système de score ayant pour base sur une mesure de similarité. Soit* $VRL(s1, s2)$ *la valeur de la ressemblance locale de* $s1$ *et* $s2$*. Alors :*

$$VRL(s1, s2) = max\ \{score\mathcal{A} \mid \mathcal{A} \text{ un alignement optimal possible de } u1 \text{ et } u2 \text{ pour tous mots } u1 \text{ et } u2 \text{ dans } s1 \text{ et } s2 \text{ respectivement}\}.$$

Bien sûr, une telle valeur n'a vraiment de sens que si certains des scores attribués aux paires de monomères (et *a fortiori* aux paires symbole-trou) sont négatifs.

3.3.2 Comparaison deux à deux ou multiple

Il y a essentiellement deux raisons pour lesquelles comparer plusieurs objets à la fois peut nous permettre de réaliser une analyse plus fine de leur degré de ressemblance. La principale est liée à l'idée de consistance. Considérons un ensemble S de $n > 2$ objets. Même si nous réalisons toutes les comparaisons deux à deux possibles de ces objets, comme chaque comparaison est indépendante, aucune ne peut faire profiter les autres automatiquement des résultats qu'elle obtient. Bien qu'une comparaison multiple, simultanée, de tous les éléments de S ne représente en fait rien d'autre qu'un très grand nombre de comparaisons deux à deux, ces dernières partagent cependant alors ce que l'on peut appeler une 'mémoire associative' en quelque sorte.

Illustrons ce point sur un exemple simple. Supposons que nous sommes en train de comparer trois chaînes $s1$, $s2$, and $s3$, et que, étant donné une certaine définition de ressemblance entre les mots de ces chaînes (par exemple celle basée sur la distance de Hamming), les mots u dans $s1$, v dans $s2$ et w dans $s3$ sont tous similaires entre eux (c'est-à-dire, leurs distances deux à deux sont toutes inférieures à une certaine valeur). Un exemple est donné dans la figure 3.15 : au maximum 2 substitutions sont nécessaires pour passer de l'un quelconque de ces mots à chacun des deux autres. Nous comparons maintenant ces chaînes soit de manière binaire en effectuant toutes les combinaisons 2 à 2 possibles de $s1$, $s2$ et $s3$, soit de manière multiple impliquant les 3 chaînes en même temps.

Dans le premier cas, nous observons que les mots u et v sont similaires, de même que v et w ainsi que u et w. Sans réaliser cependant un travail supplémentaire sur ces résultats préliminaires, il n'est pas possible de noter qu'en fait $\{u, v, w\}$ forme un seul groupe de mots similaires. Une comparaison multiple est capable d'établir ces liens de manière automatique. Elle aussi procède par comparaison de deux objets seulement à chaque fois, mais elle est capable de conserver la trace de l'information que chacune de ces comparaisons deux à deux fournit de manière à pouvoir les récupérer facilement ensuite.

La seconde raison pour laquelle une comparaison multiple est plus intéressante qu'un grand nombre de comparaisons deux à deux peut être appelée l''amplification de la signification' d'une ressemblance. Un autre exemple illustre ce point dans la figure 3.16 (a). Les similarités entre u et v, v et w, et u et w, sont maintenant plus lointaines (3 substitutions sont nécessaires pour passer de l'un de ces mots à chacun des deux autres au lieu de 2 comme dans l'exemple

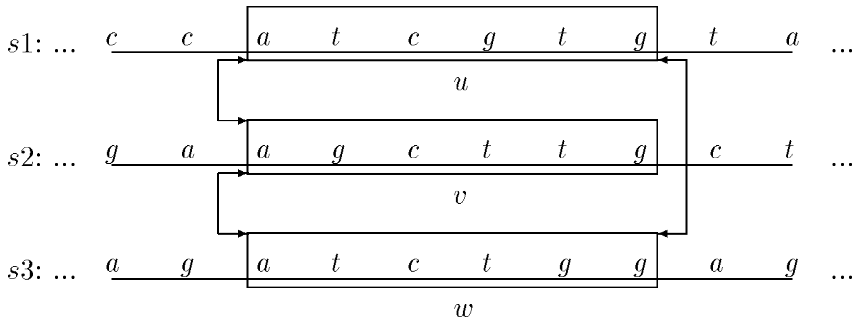
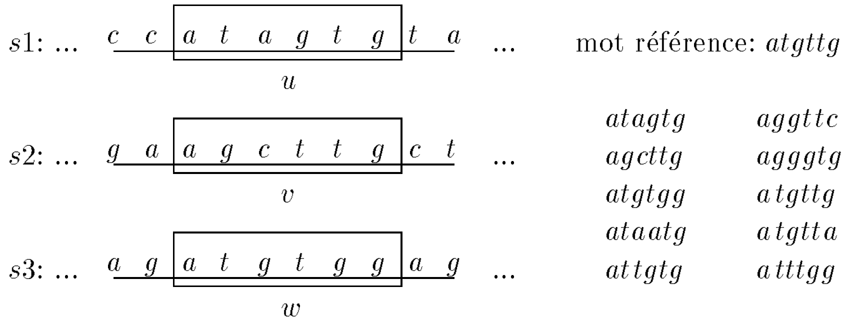


Figure 3.15: Illustration du concept de consistance dans une comparaison multiple.



(a) (b)

Figure 3.16: Illustration du concept de l'amplification de la signification dans une comparaison multiple.

précédent (figure 3.15)). Ces similarités sont devenues suffisamment faibles en fait pour que nous puissions imaginer qu'une comparaison binaire locale de deux quelconques des chaînes $s1$, $s2$ et $s3$ produise un résultat qui serait considéré comme non-significatif sauf à relâcher tellement les critères de similarité que presque tout finirait par se ressembler. Dans une comparaison multiple, nous obtenons bien sûr que ces trois mots se ressemblent tous entre eux d'une certaine façon. Et dans ce cas, ceci pourrait être vu comme étant significatif car, malgré le fait que la ressemblance de chacune des paires de mots possibles n'est jamais assez forte en elle-même, leur co-existence en est un renforcement. Cet effet devient d'autant plus marqué qu'au lieu d'avoir trois mots faiblement similaires entre eux, nous en avons dix (comme dans la figure 3.16), ou même une centaine. Un exemple concret et très illustratif d'un tel phénomène est donné dans le chapitre consacré aux illustrations.

C'est donc à des comparaisons multiples que nous allons nous attacher dans tout le reste de ce travail. Nous ne parlons des comparaisons deux à deux que dans la mesure où elles servent de base de départ pour une comparaison multiple. En particulier, le score d'un alignement multiple par exemple, qu'il s'agisse de chaînes ou de mots, est classiquement obtenu à partir de celui des alignements deux à deux le composant.

Bien sûr, réaliser une comparaison multiple est un problème intrinsèquement combinatoire que le concept de relation (et ses cliques maximales) et de couverture permettent toutefois de rendre efficace. Dans ce cas aussi, il est souvent important d'utiliser une méthode indirecte de comparaison et c'est de la distinction existant entre les méthodes directe et indirecte dont nous parlons maintenant.

3.3.3 Comparaison directe ou indirecte

Il y a deux manières de comparer des objets lorsque nous cherchons à en identifier des parties communes. Cette observation est valable quel que soit le nombre d'objets à analyser, mais la distinction devient particulièrement intéressante en termes pratiques lorsqu'il s'agit de comparer plus de deux objets à la fois.

Prenons le cas concret qui nous intéresse ici où ces objets sont des chaînes de caractères et ce que nous cherchons à identifier sont tous les groupes de mots similaires présents dans ces chaînes. La première manière de faire procède par une comparaison directe de ces mots entre eux. S'il y a un nombre total p de mots, il faut alors, si nous voulons être exhaustifs, procéder à au moins 2^p comparaisons. La seconde façon d'agir utilise une approche indirecte. Elle identifie ainsi les groupes de mots similaires en comparant les mots non plus directement entre eux, mais plutôt à des objets qui sont externes aux chaînes. Ces objets n'ont pas besoin d'être des mots, ils peuvent par exemple représenter un produit cartésien d'éléments d'une couverture. Même lorsqu'il s'agit de mots, c'est-à-dire lorsqu'ils sont définis sur le même alphabet que les chaînes, ces objets peuvent ne pas être eux-mêmes présents dans ces chaînes. Dans tous les cas, ils font ainsi fonction de points de références contre lesquels tous les objets internes (les mots dans les chaînes) sont comparés. Un ensemble de mots similaires est alors un ensemble de mots tels qu'il existe un de ces points de référence avec lequel tous les mots de l'ensemble se trouvent dans une certaine relation de similarité (voir la figure 3.17). Le nombre de comparaisons à effectuer pour le même total p de mots varie suivant la relation adoptée, nous reviendrons là-dessus plus tard, mais notons seulement pour l'instant qu'il sera en général considérablement plus petit que 2^p. Dans certains cas, il peut être proportionnel à p. Cette idée d'utiliser un objet externe pour réaliser une comparaison multiple de mots d'une chaîne a été, à notre connaissance, initialement introduite par Waterman [Galas *et al.*, 1985] [Waterman, 1984b] [Waterman, 1986] [Waterman, 1989] [Waterman, 1990] [Waterman *et al.*, 1984].

Chacun de ces objets externes représente également d'une certaine façon un 'résumé' d'un groupe de mots similaires. De tels résumés sont obtenus même lorsque les comparaisons sont effectuées de manière directe mais il est important de distinguer les deux. Dans ce second cas, le résumé des mots similaires est toujours obtenu <u>après</u> que les mots aient été identifiés. Dans le premier cas où l'on procède à des comparaisons indirectes, le résumé est le plus souvent établi <u>en même temps</u> que les groupes de mots sont trouvés. Lorsque cela est ainsi, ces objets externes peuvent être vus comme des constructeurs de ces groupes et c'est ce type d'objet qui est utilisé en ce qui concerne notre propre travail.

Dans les deux sections qui suivent, nous allons affiner les notions générales que nous venons d'introduire et montrer aussi que la différence entre comparaisons directe et indirecte est en fait souvent mince. Ceci va nous conduire à présenter jusqu'à la fin de ce chapitre une série de définitions de similarité entre mots dont nous nous servirons par la suite pour identifier les groupes de mots qui se ressemblent suffisamment entre eux dans l'espoir que ces groupes reflètent certaines

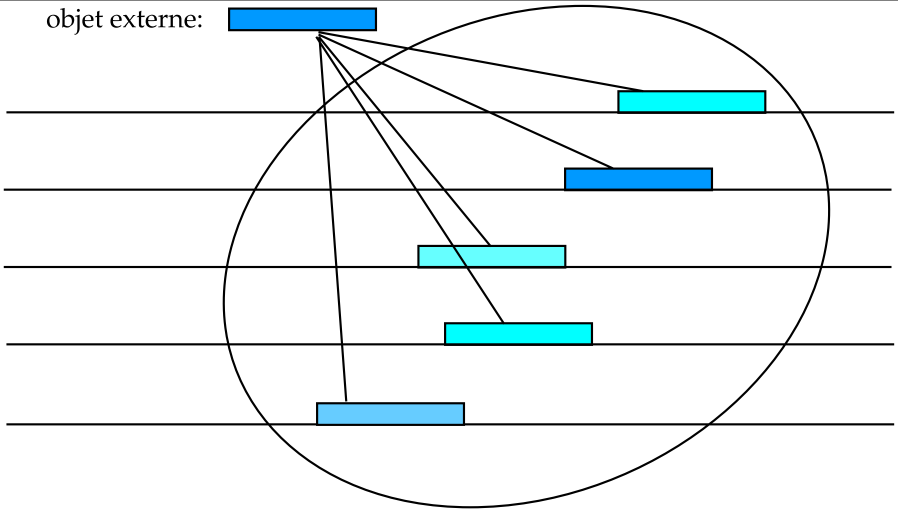


Figure 3.17: Illustration de la méthode indirecte de comparaison : utilisation d'objets externes.

caractéristiques fonctionnelles, structurales ou évolutives des molécules comparées.

Chaque groupe de mots similaires entre eux d'une certaine façon constitue ce qui est appelé un bloc [Henikoff, 1994] [Henikoff and Henikoff, 1991] [Posfai *et al.*, 1989] (voir la figure 3.18).

Notation 3.3.1 *Étant donné un ensemble de chaînes, nous appelons* <u>*bloc*</u> *un groupe de mots dans ces chaînes similaires suivant une certaine mesure de ressemblance.*

Comme nous l'avons mentionné dans le chapitre 2, un bloc peut comporter des mots présents dans chacune des chaînes que l'on analyse, ou seulement dans un sous-ensemble de celles-ci. Nous allons donc introduire là où cela est possible une contrainte supplémentaire, que nous appelons contrainte de quorum. Cette contrainte va indiquer le nombre minimum de chaînes de l'ensemble où au moins un mot d'un bloc de similarité doit apparaître pour que ce bloc soit considéré comme intéressant. Cette contrainte est essentielle car elle permet de traiter le cas où les caractéristiques communes que nous voulons déterminer dans l'ensemble des chaînes ne sont pas en fait observées sur <u>toutes</u> celles-ci. Ce cas, qui correspond à un ensemble de données comportant du bruit, est très fréquemment rencontré en pratique. Observons que cette contrainte intervient également lorsque nous travaillons avec une seule chaîne : elle indique alors le nombre minimum de fois où un mot doit apparaître répété dans la chaîne pour mériter notre attention. Nous en donnons ici une définition formelle.

Définition 3.3.4 *Étant donné un ensemble $\mathcal{S}$ de chaînes définies sur Σ^* et un bloc B de mots similaires dans $\mathcal{S}$, on appelle :*

- *le* <u>*quorum en mots*</u> q_m *de B la cardinalité de B :*

$$q_m(B) = Card(B);$$

- *le* <u>*quorum en chaîne*</u> q_c *de* B *le nombre de chaînes (distinctes) de* $\mathcal{S}$ *dans lesquelles au moins un mot de* B *est présent :*

$$q_c(B) = \sum_{s \in \mathcal{S}} \delta_0(B, s)$$

où $\delta_0(B,s) = \begin{cases} 1 \text{ si } \exists u \in B \text{ tel que } u \text{ est un mot dans } s \\ 0 \text{ sinon} \end{cases}$

Comme nous allons nous soucier surtout du quorum en chaîne, dorénavant lorsque nous parlerons de quorum, c'est donc au quorum q_c que nous ferons référence. Lorsque cela ne sera pas le cas, nous l'indiquerons de manière explicite.

Notation 3.3.2 *Nous appelons* <u>*quorum*</u> *d'un bloc* B *de mots similaires d'un ensemble de chaînes, le quorum en chaîne* q_c *de* B*.*

Nous disons alors que :

Définition 3.3.5 *Un bloc* B *de mots similaires d'un ensemble* N *de chaînes définies sur* Σ^* *satisfait la* <u>*contrainte de quorum*</u> q *si* $q_c(B) \geq q$*.*

Finalement, observons qu'il peut être important dans certaines situations de repérer des mots qui sont non plus communs à un nombre suffisamment grand de chaînes d'un ensemble, mais observés sur un nombre 'anormalement' petit de celles-ci. En d'autres termes, il peut être important de repérer les mots qui sont <u>rares</u>. Cette idée a été énoncée par [Blaisdell *et al.*, 1993] et [Waterman, 1989] en particulier et est résumée de manière très suggestive par Watanabe [Watanabe, 1985] : "*The holes in Emmenthal cheese (or lotus root) certainly can be considered as an important constituent part of the object. This reinterpretation of absence as presence may be compared with the interpretation of a 'bubble' in the sea of negative electrons in negative energy states as a positive electron*".

Nous pouvons ainsi vouloir définir le concept d'une <u>contrainte maximale de quorum</u> q en disant qu'un bloc B de mots similaires d'un ensemble N de chaînes définies sur Σ^* satisfait cette contrainte si $q_c(B) \leq q$. La contrainte de quorum de la définition 3.3.5 est dans ce cas une contrainte que nous appelons <u>minimale</u>. Les deux notions ne sont toutefois pas symétriques. En effet, nous avons :

- $q_c(B) \geq q \Longrightarrow q_m(B) \geq q$
  <u>mais</u>
- $q_c(B) \leq q \not\Longrightarrow q_m(B) \leq q$.

Un bloc peut ainsi satisfaire une contrainte maximale de quorum et cependant être présent un nombre 'anormalement' grand de fois dans une des chaînes de l'ensemble. Bien sûr, un bloc peut satisfaire une contrainte minimale de quorum et être totalement absent d'une des chaînes, mais biologiquement les deux phénomènes ne seront sans doute pas considérés comme étant également significatifs. Traiter les mots rares est aussi techniquement plus difficile et bien que représentant un problème très intéressant, nous considèrons dans ce travail uniquement les blocs satisfaisant une contrainte minimale de quorum. Le terme plus général de contrainte de quorum fera ainsi référence à cette dernière.

Nous allons commencer par décrire les mesures directes de ressemblance, et d'abord par une approche qui demeure globale. Celle-ci est importante car c'est avec elle que nous montrons comment passer de la valeur d'un alignement deux à deux donnée précédemment à celle d'un alignement multiple, puis à celle d'un bloc. La première définition de similarité entre mots que nous donnons a ainsi toujours pour base l'optimisation d'une fonction de score.

bloc de similarité:

AAAGTTGTTGACT
AAAGTTGTTGACA
GAAGTTGTTGACT
ATAGTTGTTGACT
AAAGTTATTGACT
AAAGTTATTGACT
AAAGGTGTTGACT
AAAGTTGTTGACA

Figure 3.18: Exemple d'un bloc de similarité (2 substitutions au plus autorisées entre chaque paire de mots du bloc).

3.4 Mesures directes de la ressemblance

3.4.1 Approche globale : optimisation d'une fonction de score

3.4.1.1 Deux façons d'aligner plusieurs chaînes

L'approche globale et directe pour mesurer la ressemblance entre deux objets correspond traditionnellement à celle présentée dans la section 3.2.4.2 [Needleman and Wunsch, 1970] [Sellers, 1974] [Wagner and Fischer, 1974]. La mesure de ressemblance de deux chaînes $s1$ et $s2$ est ainsi donnée par $VR(s1,s2)$.

Le problème est que, comme nous l'avons vu, nous allons vouloir comparer toujours plus de deux objets à la fois. En général, il s'agit de comparer plusieurs mots simultanément, mais nous pouvons aussi vouloir traiter des chaînes entières.

La mesure de la ressemblance globale de plus de deux chaînes que l'on compare directement entre elles est une extension de la mesure de ressemblance de deux de ces objets seulement. À la base de cette mesure multiple se trouve donc toujours un alignement comme dans le cas du deux à deux. C'est dans la façon d'obtenir cet alignement que des différences surgissent et il existe ainsi deux manières principales de procéder. Nous n'allons pas nous soucier pour l'instant de méthodes, celles-ci sont montrées rapidement dans le chapitre suivant, mais seulement de l'idée qui se cache derrière chacune des deux approches. Pour cela, revenons un peu à la biologie et rappelons que ces chaînes représentent des objets ayant une histoire évolutive éventuellement commune. Les deux méthodes d'alignement multiple varient ainsi suivant qu'il est ou non tenu compte de cette histoire. Dans le second cas, comme aucune supposition *a priori* n'est faite des liens existant possiblement entre les objets, ceux-ci sont tous comparés entre eux. Dans le premier, on essaye de reconstituer l'histoire évolutive des objets, au préalable ou en même temps que l'alignement, construisant ainsi ce que l'on appelle un arbre phylogénétique, et chaque objet n'est comparé qu'avec certains des autres objets de l'ensemble, ceux auxquels il est lié par une branche de l'arbre. En fait, cette approche s'assimile d'une certaine façon aux méthodes indirectes dans la mesure où les objets initiaux sont en réalité souvent comparés à des objets ancestraux plutôt que directement ou, du moins, que ces objets ancestraux sont établis en même temps que l'alignement se réalise.

Dans les deux cas cependant, l'alignement obtenu à la fin présente la forme d'un tableau, comme pour deux chaînes, seulement avec plus de dimensions. Établissons donc d'abord une notation pour un tel alignement.

Définition 3.4.1 *Soient* $s1 = s1_1...s1_{n_1}$, $s2 = s2_1...s2_{n_2}$, ..., $sN = sN_1...sN_{n_N}$ *N chaînes définies sur* Σ^*. *Alors* $\begin{pmatrix} \overline{s1}_1 \\ \overline{s2}_1 \\ ... \\ \overline{sN}_1 \end{pmatrix} \begin{pmatrix} \overline{s1}_2 \\ \overline{s2}_2 \\ ... \\ \overline{sN}_2 \end{pmatrix} ... \begin{pmatrix} \overline{s1}_p \\ \overline{s2}_p \\ ... \\ \overline{sN}_p \end{pmatrix}$ *avec* $min\{n_1, n_2, ..., n_N\} \leq p \leq n_1 + n_2 + ... + n_N$ *tel que :*

- $\overline{si}_j = si_k$ *ou* $\overline{si}_j = \lambda$ *pour* $1 \leq k \leq n_i$ *et* $1 \leq j \leq p$
- $\mid \overline{si}_1...\overline{si}_p \mid = n_i$ *et* $\overline{si}_1...\overline{si}_p = si_1...si_{n_i}$
- *pour tout* $j \in \{0, ..., p\}$, *il n'est jamais vrai que pour tout* $i \in \{1, ..., N\}$, $\overline{si}_j = \lambda$ *(aucune colonne de l'alignement n'est composée que de trous)*

représente un alignement des chaînes $s1$, $s2$, ..., sN.

Exemple 3.4.1 *Soient* $\Sigma = \{a, b, c, d, e\}$ *et les chaînes* $s1$, $s2$, $s3$ *et* $s4$ *données par :*
$s1 = acdeceb$
$s2 = acaebeb$
$s3 = cdbeb$
$s4 = bcdbc$.
Alors $\mathcal{A} = \begin{pmatrix} a \\ a \\ \lambda \\ b \end{pmatrix} \begin{pmatrix} c \\ c \\ c \\ c \end{pmatrix} \begin{pmatrix} d \\ a \\ d \\ d \end{pmatrix} \begin{pmatrix} e \\ e \\ \lambda \\ \lambda \end{pmatrix} \begin{pmatrix} c \\ b \\ b \\ b \end{pmatrix} \begin{pmatrix} e \\ e \\ e \\ e \end{pmatrix} \begin{pmatrix} b \\ b \\ b \\ \lambda \end{pmatrix}$ *est un alignement multiple de* $s1$, $s2$, $s3$ *et* $s4$.

La question est maintenant : quel score attribuer à cet alignement? C'est dans cette attribution des scores que va justement apparaître la différence entre un alignement obtenu par comparaison de toutes les chaînes entre elles et un alignement obtenu à partir d'un arbre phylogénétique. Voyons d'abord le premier.

3.4.1.2 Les scores SP

Dans le cas où toutes les chaînes d'un ensemble sont comparées entre elles, il est naturel de considérer que le score de l'alignement multiple est la somme des scores de tous les alignements deux à deux qu'il induit, en d'autres termes, ceux qui sont la projection de cet alignement sur chacune des paires de chaînes de l'ensemble (voir la figure 3.19).

Dans ce cas, le score d'une colonne de l'alignement est la somme des scores de toutes les paires de symboles ou paires symbole-trou de cette colonne. Ces scores sont appelés des scores SP, de l'anglais 'Sum of the Pairs' [Altschul and Lipman, 1989].

Plus formellement, nous pouvons établir une définition de ce score qui ressemble à celle du cas deux à deux, en particulier elle aussi est récursive.

Définition 3.4.2 *Soit* $\mathcal{A} = \begin{pmatrix} \overline{s1}_1 \\ \overline{s2}_1 \\ ... \\ \overline{sN}_1 \end{pmatrix} \begin{pmatrix} \overline{s1}_2 \\ \overline{s2}_2 \\ ... \\ \overline{sN}_2 \end{pmatrix} ... \begin{pmatrix} \overline{s1}_p \\ \overline{s2}_p \\ ... \\ \overline{sN}_p \end{pmatrix}$ *avec* $min\{n_1, n_2, ..., n_N\} \leq p \leq n_1 + n_2 + ... + n_N$ *un alignement des chaînes* $s1 = s1_1...s1_{n_1}$, $s2 = s2_1...s2_{n_2}$, ..., $sN = sN_1...sN_{n_N}$. *Alors :*

```
Alignement multiple :              Alignement induit :

1. A  A  G  A  A  -  A
                                   1. A  A  G  A  A  A
2. A  T  -  A  A  T  G
                                   3. C  T  G  -  G  G
3. C  T  G  -  G  -  G
```

Figure 3.19: Alignement multiple utilisant des scores SP et alignement deux à deux induit de cet alignement multiple.

$$score\mathcal{A} = score\mathcal{A}' + score\begin{pmatrix}\overline{s1}_p\\ \overline{s2}_p\\ \dots\\ \overline{sN}_p\end{pmatrix}$$

où :

$$\mathcal{A}' = \begin{pmatrix}\overline{s1}_1\\ \overline{s2}_1\\ \dots\\ \overline{sN}_1\end{pmatrix}\begin{pmatrix}\overline{s1}_2\\ \overline{s2}_2\\ \dots\\ \overline{sN}_2\end{pmatrix}\dots\begin{pmatrix}\overline{s1}_{p-1}\\ \overline{s2}_{p-1}\\ \dots\\ \overline{sN}_{p-1}\end{pmatrix}$$

et

$$score\begin{pmatrix}\overline{s1}_p\\ \overline{s2}_p\\ \dots\\ \overline{sN}_p\end{pmatrix} = \sum_{1\leq i<j\leq N} score\begin{pmatrix}\overline{si}_p\\ \overline{sj}_p\end{pmatrix}$$

avec score $\begin{pmatrix}\overline{si}_p\\ \overline{sj}_p\end{pmatrix}$ *donné par la définition 3.2.19.*

Exemple 3.4.2 *Si nous travaillons avec une mesure de similarité et une relation d'identité entre les symboles de l'alphabet, alors le score de l'alignement des trois chaînes de la figure 3.19 est égal à* 6.

La pondération affectée aux trous dans un alignement multiple est celle qu'utilise Gotoh [Gotoh, 1982]. Elle correspond à celle que nous avons décidée d'adopter : chaque trou dans un alignement reçoit un poids constant. Comme pour l'attribution d'un score à un trou dans le cas de deux chaînes, d'autres fonctions ont été élaborées, notamment par Altschul [Altschul and Lipman, 1989], Fredman [Fredman, 1984] et Murata [Murata *et al.*, 1985]. Altschul en particulier n'affecte un score qu'à l'ouverture d'un trou dans l'alignement induit sur chaque paire, pas à son prolongement. La différence entre le score de Gotoh et celui d'Altschul est montré dans la figure 3.20 à l'aide d'exemples. Il est clair que les coûts d'Altschul sont plus complexes à calculer, ils impliquent un retour en arrière constant dans l'alignement pour vérifier où a commencé un trou.

3.4.1.3 Scores obtenus en utilisant un arbre

Une des motivations pour l'étude d'objets qui ont évolué au cours du temps à partir d'un certain nombre d'ancêtres communs est d'essayer justement d'établir leur histoire et les liens

	Altschul	Gotoh
× − × × × × × × ×	2	2
× − × × − × × × ×	2	2
× − − × × × − × × × × ×	3	4
× − × × × − − × × × − ×	4	4
× − − × × × × − − × × × × × ×	4	6
× − × × × × × − × × × × − − ×	5	6
× − − − × × × − × × × × × × ×	3	6
× − − − − − × × × − × − × × × × × × × × ×	4	10

Figure 3.20: Deux façons d'attribuer un coût à un trou dans un alignement multiple : celui d'Altschul (appelé 'natural gap cost') et celui de Gotoh.

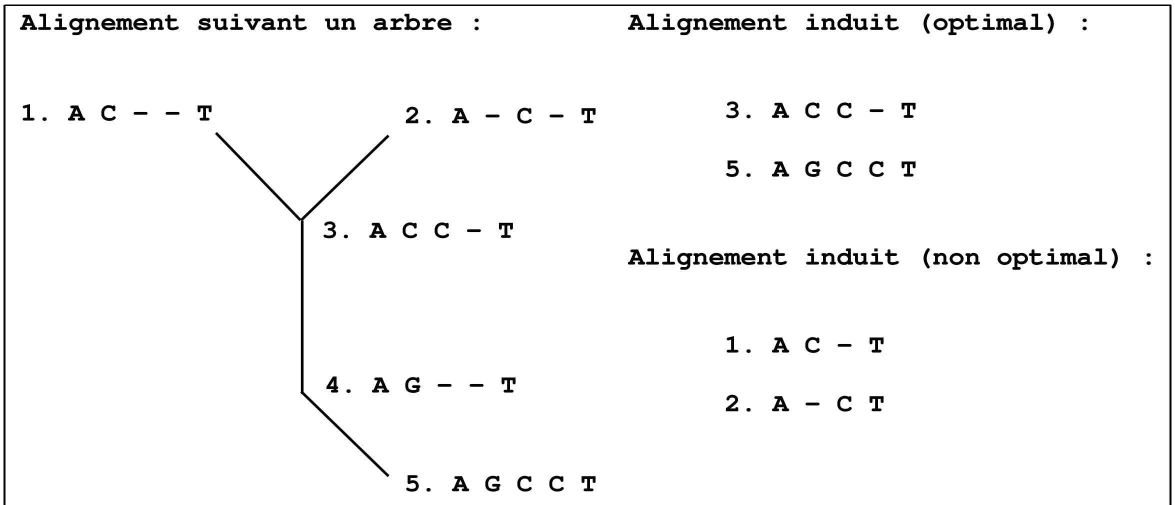


Figure 3.21: Alignements deux à deux induits d'un alignement phylogénétique.

existant entre ces objets. Ce processus peut fonctionner dans le sens inverse cependant. En effet, si l'histoire évolutive d'un ensemble taxonomique a pu être bien établie à partir de données fossiles ou morphologiques, cette histoire peut être utilisée pour éclairer l'évolution au niveau moléculaire des chaînes biologiques et donc pour comparer et aligner des macromolécules. Une première formalisation de cette approche, appelée alignement multiple à partir d'un arbre, est due principalement à Sankoff [Sankoff *et al.*, 1973] [Sankoff, 1975] [Sankoff *et al.*, 1976]. On appelle cela aussi réaliser un alignement multiple phylogénétique.

Habituellement, les chaînes de l'ensemble à aligner constituent les feuilles d'un tel arbre, et des chaînes additionnelles, correspondant à ce que l'on pourrait appeler des chaînes ancestrales, sont placées aux nœuds internes de l'arbre. Ces chaînes ancestrales sont construites pendant l'alignement et sont ajoutées à celui-ci. L'alignement multiple final présente ainsi la forme de deux tableaux placés l'un en dessous de l'autre, dont seul celui du dessus nous intéresse. Celui-ci en effet concerne l'alignement des chaînes de départ, et celui du dessous l'alignement des chaînes ancestrales. Le score de cet alignement des chaînes de départ plus les chaînes ancestrales est alors la somme des scores de tous les alignements deux à deux associés à chaque branche de l'arbre et qui sont induits par l'alignement multiple (voir la figure 3.21) [Sankoff, 1975]. Ces alignements deux à deux portent donc toujours sur une chaîne de départ et une chaîne ancestrale.

Formellement, nous avons :

Définition 3.4.3 *Soit* $\mathcal{A} = \begin{pmatrix} \overline{s1}_1 \\ \overline{s2}_1 \\ ... \\ \overline{sN}_1 \end{pmatrix} \begin{pmatrix} \overline{s1}_2 \\ \overline{s2}_2 \\ ... \\ \overline{sN}_2 \end{pmatrix} ... \begin{pmatrix} \overline{s1}_p \\ \overline{s2}_p \\ ... \\ \overline{sN}_p \end{pmatrix}$ *avec* $min\{n_1, n_2, ..., n_N\} \leq p \leq n_1 + n_2 + ... + n_N$ *un alignement des chaînes* $s1 = s1_1 ... s1_{n_1}$, $s2 = s2_1 ... s2_{n_2}$, ..., $sN = sN_1 ... sN_{n_N}$ *obtenu à partir d'un arbre phylogénétique ayant* M *nœuds internes, et donc* M *chaînes ancestrales* $a1, a2, ..., aM$ *qui ont été reconstruites en même temps que l'alignement et qui sont de longueurs* $m_1, m_2, ..., m_M$ *respectivement. Alors :*

$$score\mathcal{A} = score\mathcal{A}' + score \begin{pmatrix} \overline{s1}_p \\ \overline{s2}_p \\ ... \\ \overline{sN}_p \end{pmatrix}$$

où :

$$\mathcal{A}' = \begin{pmatrix} \overline{s1}_1 \\ \overline{s2}_1 \\ ... \\ \overline{sN}_1 \end{pmatrix} \begin{pmatrix} \overline{s1}_2 \\ \overline{s2}_2 \\ ... \\ \overline{sN}_2 \end{pmatrix} ... \begin{pmatrix} \overline{s1}_{p-1} \\ \overline{s2}_{p-1} \\ ... \\ \overline{sN}_{p-1} \end{pmatrix}$$

et

$$score \begin{pmatrix} \overline{s1}_p \\ \overline{s2}_p \\ ... \\ \overline{sN}_p \end{pmatrix} = \sum_{\substack{1 \leq i \leq N \\ 1 \leq j \leq M \\ aj \text{ chaîne ancestrale} \\ \text{auquelle } si \text{ est liée}}} score \begin{pmatrix} \overline{si}_p \\ aj_{m_j} \end{pmatrix}$$

avec score $\begin{pmatrix} \overline{si}_p \\ aj_{m_j} \end{pmatrix}$ *donné par la définition 3.2.19 sachant que* $aj_{m_j} \in \Sigma$.

Exemple 3.4.3 *Si nous travaillons avec une mesure de distance et une relation d'identité entre les symboles de l'alphabet, alors le score de l'alignement des trois chaînes de la figure 3.21 est égal à* 6.

Comme nous venons de le dire, l'arbre phylogénétique est donné au départ, il est obtenu en fait à partir d'autres mesures. Mais en réalité, il est possible de procéder dans les deux directions à la fois, c'est-à-dire de faire usage d'un arbre pour obtenir un meilleur alignement des molécules comparées, et utiliser l'alignement progressivement obtenu pour construire l'arbre. Dans ce cas au départ il n'y a rien : ni alignement, ni arbre. Lorsque l'arbre est connu au départ, nous connaissons le nombre d'ancêtres, donc le nombre de nœuds internes de l'arbre ainsi que sa forme, et ce qu'il faut reconstituer ce sont alors uniquement les chaînes ancestrales. Lorsque l'arbre n'est pas fourni initialement, et que sa forme et le nombre d'ancêtres ne sont donc pas connus, c'est tout cela plus les chaînes ancestrales qu'il faut découvrir. Nous indiquons dans la section suivante quel critère est habituellement adopté pour opérer ces reconstructions.

3.4.1.4 Mesure de la ressemblance : optimisation d'une fonction de score

Quel que soit le score utilisé, qu'il s'agisse d'un score SP ou de celui obtenu par construction d'un arbre phylogénétique, la mesure de ressemblance d'un ensemble de chaînes est habituellement donnée, comme dans le cas du deux à deux, par le score du meilleur alignement multiple qu'il est possible de réaliser avec ces chaînes.

Définition 3.4.4 *Soient N chaînes s1,s2,...,sN et soit* $VRM(s1, s2, ...sN)$ *une valeur de la similarité multiple (resp. dissimilarité) de s1,s2,...,sN. Alors :*

$$VRM(s1, s2, ..., sN) = max \text{ (resp. min) } \{score\mathcal{A} \mid \mathcal{A} \text{ un alignement de } s1,s2,...,sN\}.$$

Dans le cas où l'alignement est déterminé à partir d'un arbre, il est usuel de travailler avec des distances et c'est donc le score minimal qui va nous intéresser. Il faut observer alors que la minimisation de la fonction de récurrence présentée dans la section 3.4.1.3 implique en fait une double minimisation du fait que les chaînes ancestrales ne sont pas connues au départ. Ainsi nous avons :

$$\min(score\mathcal{A}) = \min(score\mathcal{A}' + \min(score \begin{pmatrix} \overline{s1}_p \\ \overline{s2}_p \\ ... \\ \overline{sN}_p \end{pmatrix})).$$

C'est cette double minimisation, l'externe et l'interne, qui constitue le critère utilisé pour reconstituer les chaînes ancestrales d'un arbre lorsque celui-ci est connu à l'avance. Ce critère est appelé critère de parsimonie. Lorsque l'arbre n'est pas connu, pour évaluer la mesure de ressemblance d'un ensemble d'objets en tenant compte de leur histoire évolutive il faut alors encore déterminer, parmi tous les arbres possibles, celui qui obtient un score minimal sachant que les chaînes de départ sont toujours placées aux feuilles. La combinatoire explosive d'une telle approche fait que des algorithmes exacts pour mesurer la ressemblance de plusieurs chaînes à partir d'un arbre n'existent que lorsque celui-ci est connu à l'avance.

3.4.1.5 Deux cas particuliers de score multiple : les étoiles centrales et les arbres étoiles

Les scores SP présentés dans la section précédente sont souvent coûteux à calculer mais leur principal inconvénient est qu'il n'est pas vraiment intuitif de comprendre quel sens ils peuvent avoir en biologie. Que signifie en effet attribuer comme score à un alignement (c'est-à-dire en fait établir comme mesure de sa qualité), la somme de tous les scores de paires de chaînes de l'ensemble puis à considérer que le meilleur alignement est celui dont le score ainsi obtenu est le plus élevé (ou le plus petit si l'on travaille avec des distances)? En quoi cette mesure peut-elle refléter la vraie histoire évolutive des chaînes puisqu'elle agit comme si chacune de celles-ci pouvait être un ancêtre de toutes les autres alors que la plupart n'ont pas de lien évolutif direct entre elles et ne sont en relation que parce qu'elles sont issues d'un ancêtre commun (qui peut ne même pas être présent dans l'ensemble). Sans doute une meilleure mesure d'une telle relation est celle qui s'appuie sur le concept d'étoile centrale (en anglais 'Center Star') introduite par Gusfield [Gusfield, 1993]. Gusfield lui-même considère cette mesure comme une approximation des scores SP et l'a donc établie pour des raisons pragmatiques. Nous la voyons plutôt comme une mesure intéressante en elle-même. Elle consiste à définir le score d'un alignement multiple de la façon suivante :

Définition 3.4.5 *Soient s1,s2,...,sN des chaînes définies sur Σ^*. Le score de similarité (resp. dissimilarité) d'un alignement EC ('Étoile Centrale') de s1,s2,...,sN est égal à :*

$$score(EC) = max\ (resp.\ min)\ \{\sum_{j=1}^{N} score(si, sj) \mid i \in \{1, ..., N\}\}$$

où $score(si, sj)$ est le score d'un alignement optimal des chaînes si et sj. C'est donc le maximum (resp. minimum) de la somme de tous les scores optimaux deux à deux par rapport à si, pour toute chaîne si dans l'ensemble de départ. La chaîne si pour laquelle ce maximum (ou minimum) est atteint est appelé l'étoile centrale. C'est la chaîne (étoile) qui se trouve au centre de toutes les autres chaînes (étoiles).

Exemple 3.4.4 *Si nous travaillons avec une mesure de distance et une relation d'identité entre les symboles de l'alphabet, alors le score de l'alignement des cinq chaînes de la figure 3.22 est égal à 4.*

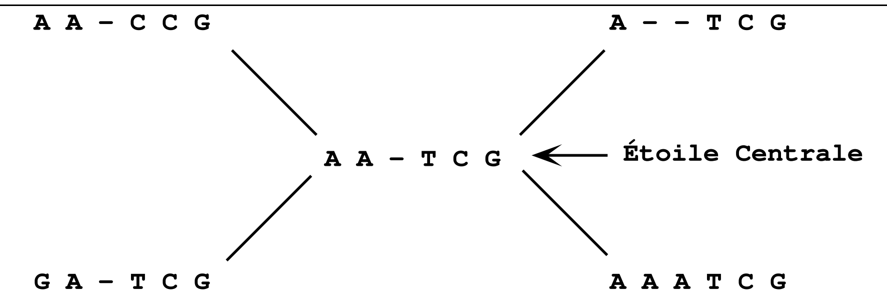


Figure 3.22: Exemple d'un alignement Étoile Centrale.

Une approximation du même genre peut être considérée également lorsque le système de score utilisé s'appuie sur un arbre phylogénétique. Dans ce cas, l'arbre est défini comme étant une étoile : il ne comporte donc qu'un seul nœud interne [Altschul and Lipman, 1989] et les chaînes à comparer sont alors supposées être les descendants directs d'un unique ancêtre commun. Si en termes de phylogénie, cette hypothèse peut paraître absurde en pratique, l'idée d'arbre étoile est au fond tout à fait semblable à celle d'étoile centrale à la seule différence que dans le cas de l'arbre, l'étoile centrale n'a plus besoin d'être une des chaînes de l'ensemble de départ. C'est un concept qu'il est tout à fait intéressant d'explorer dans le cadre des méthodes indirectes de comparaison et que nous reverrons alors sous le nom de centre de gravité ou de mot de Steiner.

3.4.2 Résumé des mesures directes et globales de la ressemblance

Un résumé des mesures directes et globales de la ressemblance est donné dans le tableau ci-dessous.

Mesures Directes et Globales de la Ressemblance	
Caractéristiques principales	Référence aux définitions
optimisation d'une fonction de score scores SP	définition 3.4.4 et définition 3.4.2
optimisation d'une fonction de score scores ayant pour base un arbre phylogénétique	définition 3.4.4 et définition 3.4.3
optimisation d'une fonction de score scores Étoile Centrale	définition 3.4.5

3.4.3 Approche locale

3.4.3.1 Optimisation d'une fonction de score

L'approche locale et directe pour mesurer la ressemblance entre plusieurs objets simultanément commence par une application des méthodes traditionnelles globales que nous avons vues un peu plus haut et fait donc appel encore une fois à l'optimisation d'une fonction de score d'un alignement. Dans ce cas cependant, le score attribué à un alignement de mots dans les chaînes que l'on compare est toujours un score SP et jamais un score obtenu à partir d'un arbre phylogénétique. *A priori*, il n'y a pas de raison pour que les connaissances apportées par un tel arbre ne puissent être utilisées pour comparer des mots, mais la formalisation d'un tel problème devrait sans doute en être différente que celle introduite dans le cas de chaînes entières.

Il suit donc une première définition de similarité entre mots donnée par :

Définition 3.4.6 *Étant donné N chaînes* $s1,s2,...,sN$ *et une constante seuil, les mots* $u1,u2,...,uN$ *de* $s1,s2,...,sN$ *respectivement sont considérés comme étant similaires si* $score\mathcal{A} \geq seuil$ *avec* $\mathcal{A}$ *un meilleur alignement multiple de* $u1,u2,...,uN$ *(ou* $score\mathcal{A} \leq seuil$ *si le score d'un alignement mesure une dissimilarité ou une distance).*

À nouveau ici, si nous travaillons avec une mesure de similarité, il est possible d'attribuer une valeur à la ressemblance locale d'un ensemble de chaînes de la façon suivante :

Définition 3.4.7 *Soient N chaînes* $s1,s2,...,sN$ *et soit* $VRML(s1,s2,...sN)$ *la valeur de la ressemblance multiple locale de* $s1,s2,...,sN$*. Alors :*

$$VRML(s1,s2,...,sN) = max\ \{score\mathcal{A} \mid \mathcal{A} \text{ un alignement de } u1,u2,...,uN \text{ pour tous mots } u1,u2,...,uN \text{ dans } s1,s2,...,sN \text{ respectivement}\}.$$

Encore une fois, cette valeur n'a de sens que si les scores de certaines des paires de monomères sont négatifs. Un groupe de mots $u1,u2,...,uN$ dont l'alignement produit ce score maximal est alors un bloc des mots les plus similaires qu'il soit possible de trouver dans chacune des chaînes.

Sans entrer pour l'instant dans des considérations de complexité algorithmique sur lesquelles nous reviendrons dans le chapitre suivant, nous pouvons déjà observer que cette définition de similarité entre mots est en pratique la plus coûteuse de toutes avec laquelle travailler, et cela même lorsque les trous ne sont pas permis. Observons aussi que cette définition suppose implicitement que nous fixons la contrainte de quorum q (voir la définition 3.3.5) à une valeur qui correspond en pourcentage à 100% : un bloc de similarité doit avoir au moins une occurrence, c'est-à-dire au moins un mot présent dans chacune des chaînes de l'ensemble. Or nous avons vu qu'il est important de pouvoir travailler en présence de 'bruit', à savoir avec des valeurs variables de quorum. Cela est bien sûr possible mais la définition donnée ici est particulièrement mal adaptée pour traiter de manière algorithmiquement efficace ce genre de situation. Nous en donnerons la raison précise dans le chapitre suivant, mais intuitivement nous pouvons déjà voir que l'introduction d'une telle contrainte augmente la combinatoire du problème. C'est ce qui a motivé l'élaboration d'autres définitions de similarité entre mots. Celle que nous présentons maintenant ne traite que le cas des blocs de mots similaires dont l'alignement ne comporte pas de trous et seulement des substitutions 'conservatives'. Il s'agit en effet d'une extension directe de la relation entre monomères présentée dans la section 3.2.3.2.1 et les substitutions dites conservatives sont celles impliquant des éléments en relation par R.

3.4.3.2 Recherches combinatoires exactes

Une relation R de similarité entre monomères et le concept de cliques maximales tels que ceux-ci ont été vus dans la section 3.2.3.2 peuvent être étendus de façon très naturelle aux mots, que ceux-ci soient lexicaux 3.2.3.2.1 ou structuraux 3.2.3.2.2. Comme ceux qui nous intéressent sont présents dans un ensemble de chaînes desquelles nous voulons extraire les traits communs, nous avons indiqué (définition 3.1.2) que chacun d'eux peut être identifié de manière non ambiguë par sa position dans ces chaînes. Il est alors traditionnel d'étendre la relation entre monomères en tant que symboles ou en tant qu'objets 3D à une relation entre positions dans les chaînes

Soit $\Sigma = \{I, L, M, P, C, W\}$ et soient $\{\{I, L\}, \{L, M\}, \{P\}, \{C\}, \{W\}\}$ les cliques maximales d'une relation R sur Σ.
Soient :

$s1 = LLP$
$s2 = LIC$
$s3 = MLIW$

et $s = s1s2s3 = LLPLICMLIW$.
Alors {1,4,8} et {1,4,7,8} sont des cliques de R_2 mais comme {1,4,8} est inclus dans {1,4,7,8}, ce dernier ensemble est l'unique clique maximale de R_2.

Figure 3.23: Exemple de clique maximale C_k d'une relation R_k établie sur l'alphabet des acides aminés.

sachant que celles-ci font référence aux mots. Pour faciliter la notation, nous considérons aussi que les chaînes sont en fait concaténées en une seule, notée s (nous avons donc $s = s1s2...sN$), et la relation de similarité entre mots porte finalement sur leurs indices dans s.

Définition 3.4.8 *Soit une chaîne $s \in \Sigma^n$ et i, j deux positions différentes dans s (c'est-à-dire, deux indices) tels que $1 \leq i, j \leq n$, alors :*

$$i\ R_1\ j \Leftrightarrow s_i\ R\ s_j.$$

Soient maintenant $k \geq 1$ un entier et i, j deux positions différentes dans s telles que $i, j \leq n - k + 1$, alors :

$$i\ R_k\ j \Leftrightarrow (i+p)\ R_1\ (j+p) \text{ pour tout } p \in \{0, ..., k-1\}.$$

Nous disons de i et j vérifiant la condition ci-dessus qu'ils sont en k-relation.

Définition 3.4.9 *Étant donné une chaîne $s \in \Sigma^n$, un ensemble C_k de positions dans s est une clique de la relation R_k si pour tout $i, j \in C_k$, $i\ R_k\ j$. C_k est dit maximal si pour tout $l \in [1..n] \setminus C_k$, $C_k \bigcup \{l\}$ n'est plus une clique.*

Exemple 3.4.5 *Voir la figure 3.23.*

Les cliques maximales d'une relation R_k fournissent alors une façon différente de définir un tel bloc de mots similaires d'une ou plusieurs chaînes dont nous verrons par la suite qu'elle conduit à des algorithmes exacts bien plus efficaces que ceux utilisant la définition 3.4.6 établie à partir d'une fonction de score. L'introduction d'une contrainte de quorum q (voir la définition 3.3.5) est en outre bien plus aisée à traiter. La définition obtenue est alors la suivante [Soldano *et al.*, 1995] :

Définition 3.4.10 *Étant donné une chaîne $s \in \Sigma^n$ (qui est la concaténation de N chaînes), une relation R sur Σ, un entier q entre 2 et N représentant la contrainte de quorum et un autre entier k, les groupes de mots similaires de longueur k dans s sont les cliques maximales C_k de R_k qui satisfont la contrainte q.*

R peut bien sûr être égal à l'identité, ou peut être une relation d'équivalence. Dans le premier cas, les groupes de mots similaires sont en fait des ensembles de mots identiques. Cette définition de similarité entre mots est celle utilisée dans les algorithmes de Karp, Miller et Rosenberg [Karp *et al.*, 1972], Crochemore [Crochemore, 1981] [Crochemore and Rytter, 1991], Landraud [Landraud *et al.*, 1989] et Martinez [Martinez, 1983]. La relation d'identité a été étendue à une relation de similarité R quelconque, c'est-à-dire non nécessairement transitive, par Brutlag [Brutlag *et al.*, 1990] et Cobbs [Cobbs, 1994] pour le cas du deux à deux et par Soldano [Soldano *et al.*, 1995] pour le cas du multiple. Lorsque les chaînes représentent un codage linéaire de structures protéiques, et R une relation entre paires d'angles codées sur les nœuds d'une grille comme indiqué dans la section 3.2.3.2.2, alors les cliques maximales de R_k représentent tous les motifs structuraux contigus de longueur k d'un ensemble de protéines [Pothier, 1993] [Sagot *et al.*, 1995b].

Cette approche comporte une inflexibilité d'une autre sorte que celle vue précédemment puisqu'elle ne permet dans sa définition de similarité ni les substitutions non conservatives (i.e. celles impliquant des éléments non en relation par R) ni les trous. Avant de présenter notre travail qui a porté sur cet aspect-là justement, nous voulons discuter une dernière définition de similarité entre mots. Celle-ci ne résoud qu'en partie les problèmes soulevés par les deux premières mais elle représente une approche qu'il est essentiel de connaître car elle fournit de très bons résultats en pratique. C'est souvent contre ces résultats-là que nous avons mesuré les notres, nous reviendrons là-dessus dans le chapitre consacré aux illustrations. Cette approche fait fortement appel à divers critères de signification statistique, et c'est cela qui en fait à la fois son intérêt et résume toute la difficulté que nous pouvons avoir à bien saisir ce que la mesure de ressemblance utilisée capte vraiment.

3.4.3.3 Optimisation stochastique

Cette approche utilise pour mesurer la ressemblance d'un ensemble de mots une fonction d'optimisation stochastique qui n'est pas tout à fait immédiate à comprendre. Nous n'en fournissons ici qu'une idée très générale. La mesure de ressemblance que nous schématisons ci-dessous est donc présentée sous son aspect le plus simple, celui que donne Lawrence dans deux articles principaux [Lawrence *et al.*, 1993] [Lawrence and Reilly, 1990] et qui s'inspire de la formule utilisée par Hertz et Stormo [Cardon and Stormo, 1992] [Hertz *et al.*, 1990].

Définition 3.4.11 *Étant donné un ensemble de chaînes $s1,s2,...,sN$, le groupe de mots $u1,u2,...,uN$ dans $s1,s2,...,sN$ respectivement ayant même longueur k et que l'on considère comme étant le plus similaire entre eux est celui qui maximize la valeur de la formule :*

$$\log L = \sum_{i=1}^{k} \sum_{j \in \Sigma} n_{ij} \log \frac{q_{ij}}{p_j}$$

où Σ est l'alphabet des monomères (nucléotides ou acides aminés), n_{ij} est le nombre de fois où le monomère j apparaît dans la colonne i de l'alignement sans trous des mots, q_{ij} est la fréquence d'apparition du monomère j à cette position et p_j est la fréquence d'apparition du monomère j sur les chaînes ailleurs que dans les mots du groupe.

Intuitivement, cette définition a pour base la quantité d'information que peut représenter un groupe de mots similaires. Cette information est résumée en une matrice qu'on espère spécifique du signal [Cardon and Stormo, 1992]. La ressemblance entre mots présents dans des chaînes est ainsi fonction de l'intensité du signal biologique qu'ils portent.

Observons toutefois que la mesure de la ressemblance que cette définition donne s'appuie sur des fréquences d'apparition observées. En particulier, il faut estimer la valeur de q_{ij} qui devrait indiquer la fréquence d'apparition du monomère j à la position i de tous les mots représentant un certain signal biologique, d'après la seule valeur connue qui est celle des n_{ij} indiquant le nombre de fois où le monomère j apparaît à la position i dans les exemples dont on dispose pour caractériser ce signal. Nous retrouvons donc ici la même difficulté rencontrée lorsque nous avons discuté du mode de construction des matrices de substitution pour les acides aminés (voir section 3.2.2.2) et de l'attribution d'une pondération aux trous dans un alignement (voir section 3.2.4.1.2). Elle devient plus critique dans ce cas puisqu'en général le nombre d'exemples (de chaînes) desquels nous devons extraire une information peut être relativement petit. Si par exemple q_{ij} est choisi proportionnel à n_{ij} (ce qui correspond à la façon la plus naturelle de procéder), nous pourrions avoir $q_{ij} = 0$ si l'acide aminé j n'est jamais observé dans l'ensemble des exemples dont nous disposons. La manière la plus classique de contourner cette difficulté consiste à employer des 'régularisateurs' qui peuvent être vus comme des rectificateurs des valeurs observées pour n_{ij} sur les exemples. Ces régularisateurs produisent ainsi des estimateurs pour q_{ij} (et d'une manière similaire non montrée ici pour p_j) appelées valeurs postérieures (en anglais 'posterior counts') et notées x_{ij}. Il existe plusieurs types de régularisateurs. Les trois principaux sont les pseudo-comptages (en anglais 'pseudo-counts'), les matrices de substitutions et les mélanges de Dirichlet. Nous ne parlerons que des deux premières, une référence à la troisième peut être trouvée dans l'article de Karplus [Karplus, 1995] dont nous avons adopté également les définitions concernant les pseudo-comptages et les matrices de substitution.

Les pseudo-comptages sont des régularisateurs non nuls qui dépendent du monomère en question et qui sont additionnés aux n_{ij}. On obtient ainsi :

$$x_{ij} = n_{ij} + z_j.$$

Le plus souvent, z_{ij} est choisi égal à $a.f_j$ où f_j est la fréquence d'apparition du monomère j partout sur les chaînes (donc $f_j \neq p_j$) et a une constante. C'est le régularisateur utilisé par Lawrence [Lawrence *et al.*, 1993]. Ce type de pseudo-comptage est appelé pseudo-comptage de bruit de fonds (en anglais 'background pseudo-counts'). Il fait référence aussi à la méthode de prédiction de Bayes.

Les régularisateurs qui emploient une matrice de substitution $\mathcal{M}$, et qui ne sont donc applicables qu'aux protéines, produisent des valeurs postérieures qui sont une combinaison linéaire des valeurs observées pour n_{ij} :

$$x_{ij} = \sum_k \mathcal{M}(k,j).n_{ik}.$$

Enfin les deux approches sont le plus souvent combinées pour produire des valeurs postérieures définies par :

$$x_{ij} = \sum_k \mathcal{M}(k,j).n_{ik} + z_j.$$

Karplus montre que les matrices de substitution seules sont les meilleurs régularisateurs pour estimer les valeurs de n_{ij} en présence d'un seul exemple mais qu'elles sont plus mauvaises que les pseudo-comptes au-delà de trois exemples. Il montre également que les mélanges de Dirichlet sont les meilleurs régularisateurs de tous mais ce sont les plus coûteux à calculer.

Il faut savoir également que des mesures plus complexes que celle donnée dans la définition sont possibles lorsque l'on veut identifier plusieurs blocs de similarité, modifier la longueur des mots similaires, permettre la présence de trous ou introduire une contrainte de quorum [Cardon and Stormo, 1992].

Toutes ces mesures font appel à des modèles que l'on peut voir comme des objets externes dont on se sert pour comparer les objets concrets que sont les mots. Dans le cas de Lawrence, les modèles sont des matrices de fréquences d'apparition des monomères aux positions dans les chaînes où se situe un signal. Dans les méthodes utilisant des HMM ('Hidden Markov Model') [Krogh *et al.*, 1994], les modèles sont des automates stochastiques. Nous pouvons donc considérer que les approches regroupées dans cette section se trouvent entre les mesures directes de comparaison et les indirectes. Les objets externes avec lesquels nous travaillons ne sont cependant pas les mêmes et, surtout, la manière de les obtenir est très différente. Les notres sont en effet construits par récurrence tandis que la taille des matrices de Lawrence et celle des automates stochastiques est le plus souvent fixée à l'avance, ou du moins estimée. En conséquence, les algorithmes utilisant la définition de similarité donnée ici, ou les formules que l'on peut trouver dans les articles sur les HMM, emploient tous des heuristiques pour établir ces matrices ou automates. Il est important cependant de noter que la raison n'est pas seulement qu'ils ne seraient pas efficaces en pratique s'il en était autrement, mais est propre à la philosophie même de cette approche. Nous reviendrons là-dessus dans le chapitre suivant. Avant d'examiner les objets externes que nous utilisons et qui conduisent à des algorithmes exacts de recherche de blocs de similarité, nous voulons cependant conclure cette section en parlant des diverses représentations qu'il est possible de donner des groupes de mots similaires obtenus par comparaison directe de ces mots entre eux.

3.4.4 Représentation des blocs obtenus par comparaison directe des mots

En effet, l'exploration de la ressemblance d'un ensemble d'objets dont nous pensons qu'ils exercent une même activité a en fin de compte pour but de caractériser ce que les objets ont en commun. L'idée est que ces caractérisations, si elles sont suffisamment précises et complètes, devraient nous permettre alors d'identifier d'autres macromolécules que celles de l'ensemble initial possédant la même activité. Une fois les traits communs repérés, dans le cadre de ce travail il s'agit de mots, il faut donc encore trouver un moyen de représenter l'information qu'ils contiennent afin de pouvoir s'en servir pour localiser ces traits ailleurs. Ainsi que nous l'avons mentionné à la fin du chapitre 2, l'idéal serait que le trait commun avec sa représentation soit non seulement caractéristique des molécules jouant un même rôle, mais qu'il leur soit aussi spécifique. Repérer un trait spécifique est bien sûr beaucoup plus difficile à réaliser, et il est quasiment impossible de vérifier que nous y sommes bien parvenus.

Les représentations qui ont été faites de ces groupes de mots similaires dans le cas où ceux-ci ont été comparés de façon directe sont aussi diverses que les définitions et les moyens élaborés pour les obtenir. La terminologie employée est aussi très confuse et ambiguë, certains termes ayant souvent été employés avec des significations différentes.

Nous pouvons cependant classer ces représentations en cinq grandes catégories que sont les alignements, les consensus, les expressions régulières de portée variable, les tableaux de propriétés et de relations et enfin les automates stochastiques auxquels nous avons fait référence dans la section précédente.

Les alignements sont la manière la plus simple et immédiate de représenter ces mots communs. C'est aussi à partir des alignements que la plupart des autres représentations sont établies. Les consensus et les expressions régulières en représentent ainsi des formes plus compactes qui résument et parfois simplifient considérablement l'information contenue dans un alignement, et facilitent la recherche de cette information ailleurs.

La catégorie des consensus est elle-même divisée en deux sous-catégories, celle des matrices ou des profils, et celle des motifs consensuels. Les matrices de consensus [Goldstein and Waterman, 1994] [Schneider *et al.*, 1986] [Stormo, 1990] [Tatusov and Koonin, 1994] [Tatusov *et al.*, 1994] ne sont rien d'autre en fait que des tableaux où chaque entrée présente, pour chaque symbole de l'alphabet, la probabilité de trouver ledit symbole à une position donnée dans un groupe de mots alignés. Les profils [Gribskov *et al.*, 1987] [Gribskov *et al.*, 1988] [Gribskov *et al.*, 1990] contiennent en outre la probabilité d'observer un trou à une position quelconque dans l'alignement. Les motifs consensuels quant à eux sont des mots sur le même alphabet Σ que celui servant à représenter les macromolécules. Ils ont reçu des noms divers dans la littérature : motif [Bairoch, 1992] [Barton and Sternberg, 1990] [Posfai *et al.*, 1989] [Smith *et al.*, 1990] [Smith and Smith, 1990], motif consensuel [Galas *et al.*, 1985] [Hertz *et al.*, 1990] [Waterman, 1984b] [Waterman *et al.*, 1984] [Waterman, 1986] [Waterman, 1989] [Waterman, 1990], signature [Sheridan and Venkataraghavan, 1992]. Nous pouvons les classer en fait aussi parmi les expressions régulières dont ils représentent la forme la plus simple. Les formes plus complexes englobent les motifs dégénérés définis sur une partition de $\mathcal{P}(\Sigma) \setminus \{\Sigma\}$ [Galas *et al.*, 1985] [Smith and Smith, 1990] [Waterman, 1984b] [Waterman *et al.*, 1984] [Waterman, 1989], ceux définis sur un sous-ensemble quelconque de $\mathcal{P}(\Sigma) \setminus \{\Sigma\}$ [Taylor, 1986c] [Taylor, 1986b] [Taylor, 1989] ou de $\mathcal{P}(\Sigma)$ [Neuwald and Green, 1994] (dans ce second cas on dit que le joker $\{\Sigma\}$ — en anglais 'wild card' ou 'don't care symbol' — est autorisé) et enfin les expressions régulières complètes. Des exemples de consensus (matrice et motif) et d'expressions régulières sont données dans la figure 3.24.

Les représentations peuvent être plus sophistiquées et, dans le cas des protéines, essayer d'inclure une information structurale indiquant les propriétés physico-chimiques des mots et les relations proches ou distantes reliant chaque position du groupe aux autres. Des illustrations de ce type de représentation peuvent être trouvées dans l'article de Sali [Sali and Blundell, 1990] et dans celui d'Overington [Overington *et al.*, 1990]. Ces représentations ne sont possibles que si pour au moins une des chaînes de l'ensemble que l'on a comparé la structure tertiaire de la protéine est connue par ailleurs (par cristallographie par exemple). Enfin, les groupes de mots similaires identifiés par optimisation stochastique sont représentés dans leur forme la plus simple par des matrices [Lawrence *et al.*, 1993] [Lawrence and Reilly, 1990] ou dans leur forme plus complexe par un automate non déterministe dont les transitions sont données par des probabilités [Krogh *et al.*, 1994]. Les automates comme les matrices sont établis non pas à la fin de la recherche comme cela est le cas de toutes les autres représentations, mais au cours de celle-ci et sont ainsi modifiés en permanence. Ils servent donc à la fois à résumer des ensembles de mots et à construire ces ensembles. C'est dans ce sens également que nous avons dit que cette approche par optimisation stochastique est à mi-chemin entre les mesures directes de comparaison et les indirectes que nous avons élaborées et dont nous allons parler maintenant.

3.4.5 Résumé des mesures directes et locales de la ressemblance

Un résumé des mesures directes et locales de la ressemblance est donné dans le tableau ci-dessous.

Mesures Directes et Locales de la Ressemblance	
Caractéristiques principales	Référence aux définitions
optimisation d'une fonction de score	définition 3.4.6
recherches combinatoires exactes utilisation d'une relation R de similarité	définition 3.4.10
optimisation stochastique	définition 3.4.11

Soit $\Sigma = \{a, b, c\}$. Nous présentons ci-dessous quelques-unes des manières les plus couramment utilisées pour représenter d'une manière condensée un groupe de mots similaires :

- matrice consensus :

position	1	2	3	4	5
a	0.70	0.10	0.50	0.20	0.85
b	0.20	0.10	0.40	0.00	0.05
c	0.10	0.80	0.10	0.80	0.10

- motif simple : *abaca*
- motif dégénéré, joker $\{\Sigma\}$ non autorisé :
  - les ensembles ne peuvent être que des classes: $[ab]b[ab]ca$
  - les ensembles peuvent être quelconques : $[ab][bc][c][a]$
- motif dégénéré, joker autorisé : $[ab]n[b]nn[c]$ où n est le joker $[abc]$
- expression régulière : $[ab]n\star[ab]$ où $n = [abc]$ et $\star$ représente un nombre quelconque de symboles.

Figure 3.24: Exemples de représentations de groupes de mots similaires.

Tout ce que nous avons présenté jusqu'ici a été des définitions de similarité (dissimilarité) proposées par d'autres que nous. Les définitions que nous établissons à partir de maintenant concernent toutes celles que nous avons élaborées dans le cadre de ce travail.

3.5 Mesures indirectes de la ressemblance

3.5.1 Introduction aux modèles : à la fois des constructeurs et des résumés de la ressemblance

Les définitions de similarité entre mots que nous introduisons maintenant ont pour caractéristique principale de faire appel à des objets externes contre lesquels les mots dans les chaînes sont comparés. Ces objets externes sont à la fois la représentation résumée des blocs de similarité et leurs constructeurs. Ils sont donc obtenus en même temps que les blocs sont recherchés et toute définition de similarité entre mots est ainsi une définition de similarité entre un ensemble de mots et un certain objet externe (voir la figure 3.17).

Ces définitions essayent toutes de répondre, de diverses façons possibles, à certains problèmes que peuvent présenter les trois définitions précédentes. Elles permettent donc des erreurs (substitutions non conservatives et/ou trous) et l'utilisation d'une contrainte de quorum tout en conduisant à des algorithmes qui sont exhaustifs, exacts et relativement efficaces. Tous les blocs de mots similaires suivant une de ces définitions et qui satisfont une contrainte de quorum q sont ainsi localisées en un temps qui demeure dans la plupart des cas raisonnable.

Les raisons pour lesquelles plusieurs définitions ont été établies et non pas une seule sont à la fois historiques et pragmatiques. La raison historique est celle à laquelle nous avons déjà fait allusion dans l'introduction. Tout au long de ce travail, nous avons essayé de modéliser la notion de ressemblance entre macromolécules biologiques. Au fur et à mesure que nous avancions, et que des premiers résultats étaient obtenus, nos idées ont changé, et des nécessités nouvelles sont apparues. Ces idées évoluent encore et il est vraisemblable qu'elles conduisent à de nouvelles définitions.

La raison pragmatique est liée au fait qu'il n'est sans doute pas réaliste d'espérer qu'une seule définition permette de traiter tous les cas auxquels nous pouvons être confrontés. Chaque définition peut ainsi être plus adaptée à la reconnaissance de traits communs particuliers. Pour l'instant, nous allons nous contenter de donner ces définitions, en essayant toujours cependant à chaque fois de les motiver biologiquement.

Ces définitions peuvent être divisées en deux grandes catégories : celles qui comptent les erreurs et celles qui affectent un score à ceux-ci. Nous disons des premières qu'elles travaillent avec une notion de distance basée sur des symboles, et de la dernière qu'elle utilise une notion de distance basée sur des mots.

Les objets externes employés sont également de deux types : il s'agit soit de mots définis sur le même alphabet que les chaînes que l'on compare, soit des produits cartésiens d'ensembles d'une couverture établie à partir de cet alphabet. Dans tous les cas, ces objets sont appelés des modèles. La représentation qu'ils fournissent d'un ensemble de mots similaires correspond ainsi soit aux motifs simples soit aux motifs dégénérés des mesures directes. Il est sans doute possible d'imaginer d'autres modèles.

Observons finalement que, dans tout le reste de ce chapitre, nous allons adopter la notation de la section 3.4.3.2. Chaque mot présent dans une des chaînes d'un ensemble $s1,s2,...,sN$ va ainsi être identifié par sa position (son indice) dans la chaîne s qui est la concaténation de $s1,s2,...,sN$. Par ailleurs, la contrainte de quorum qui va jouer sur un modèle est bien sûr la contrainte sur les blocs de la définition 3.3.5 puisqu'un modèle identifie toujours un bloc de mots similaires.

3.5.2 Les différentes définitions de similarité travaillant avec des modèles

3.5.2.1 Définitions utilisant une notion de distance basée sur des symboles

3.5.2.1.1 Des modèles qui sont des mots — une application à l'ADN et à l'ARN

Nous avons indiqué au début de ce chapitre (section 3.2.2.1) que lorsqu'il s'agit d'analyser des macromolécules d'ADN ou d'ARN, il est traditionnel de considérer que deux nucléotides sont soit identiques, soit différents sans moyen terme entre les deux. Des relations d'équivalences ou de similarités sont rarement établies dans ce cas. De telles relations sont pourtant possibles et envisageables dans certaines situations, mais ici nous décidons que nous allons simplement autoriser un certain nombre d'erreurs entre les mots que nous définissons comme étant similaires. En fait, comme nous venons de le voir, les erreurs entre mots vont être introduites de manière indirecte : c'est ainsi entre chaque mot d'un ensemble et un même modèle qu'un certain nombre d'erreurs va être permis. Comme aucune relation de similarité n'est établie entre les nucléotides, la définition d'un modèle que nous employons dans ce cas est celle d'un mot sur le même alphabet Σ que celui des monomères [Sagot *et al.*, 1995a] :

Définition 3.5.1 *Un modèle m de longueur k est un élément de Σ^k.*

La relation entre un modèle et des mots présents dans s est alors donnée par :

Définition 3.5.2 *Étant donné un modèle m, m est dit être L-présent dans une chaîne s à la position i s'il existe au moins un mot u qui se trouve situé à cette position dans s et qui est tel que $distance_L(u,m) \leq e$ où $distance_L$ est la distance de Levenshtein entre u et m. Nous*

Soient $\Sigma = \{A, B, C\}$ et $e = 1$. Étant donné le modèle :

$$m = ABCABC$$

le mot $u = ACABC$ est une L-image de m avec une délétion (premier B) et le mot $v = ABCABBC$ est une L-image de m avec une insertion (second B). Les deux occurrences sont notées (1, 0, 1, 0) et (1, 0, 0, 1) respectivement.

Figure 3.25: Convention pour noter le nombre de délétions et insertions.

appelons u une L-image de m.

Soit $d_s + d_d + d_i = distance_L(u,m)$ où d_s, d_d et d_i sont le nombre de substitutions, délétions et insertions respectivement (par convention, nous notons d_d et d_i respectivement le nombre de délétions et insertions réalisées sur u — voir la figure 3.25).

Nous disons que (i,d_s,d_d,d_i) est une L-occurrence de m dans s. Nous avons $| u | = | m | - d_d + d_i$.

Finalement, nous notons $L(m)$ l'ensemble de toutes les L-occurrences de m dans s, c'est-à-dire :

$$L(m) = \{(i,d_s,d_d,d_i) \mid d = d_s+d_d+d_i \text{ est une } distance_L(u,m) \text{ et } d \leq e\}.$$

Exemple 3.5.1 *Voir la figure 3.26.*

Observons que la décomposition en d_s, d_d et d_i peut ne pas être unique. Un modèle peut donc être L-présent dans s à la position i plus d'une fois, avec différentes L-images. Il peut même être présent pour la même image u avec différentes L-occurrences.

Par ailleurs, la relation qui est établie entre un modèle et son image induit une relation entre les images elles-mêmes, c'est-à-dire entre les mots de s, puisque la distance de Levenshtein entre deux quelconques d'entre eux est au plus $2e$.

Il convient d'observer toutefois que la réciproque n'est pas vraie, à savoir, étant donné un ensemble de mots tels que la distance entre deux quelconques d'entre eux est au plus $2e$, il n'existe pas toujours un modèle avec lequel tous ces mots soient en relation. Un exemple de ce fait est donné ci-dessous :

Exemple 3.5.2 *Soient $\Sigma = \{A, C, G, T\}$ et $e = 1$. Soient les mots $AGCG$, $ACCC$ et $ATCT$. La distance de Levenshtein entre deux quelconques de ceux-ci est toujours égale à $2e = 2$, mais il est clair qu'il n'existe aucun modèle m tel que la distance de chacun des mots à m soit au plus 1.*

Des modèles qui sont des mots et la distance de Levenshtein nous fournissent une façon de définir une relation de similarité entre des mots présents dans des chaînes.

Définition 3.5.3 *Étant donné une chaîne $s \in \Sigma^n$ (qui est la concaténation de N chaînes), un entier q entre 2 et N représentant une contrainte de quorum et un autre entier k, les groupes de mots similaires dans s sont les ensembles de L-occurrences de tous les modèles m de longueur k satisfaisant cette contrainte.*

Soient $\Sigma = \{A, C, G, T\}$ et $e = 1$. Soient les trois chaînes :

$s1 = CAAATTCATA\underline{ACAAATGGAGGA}GGATGTATATGTATCCGGA$
$s2 = TATTTAACCGTGACA\underline{ACAAGGAGGA}AACGTAATGATGAATC$
$s3 = TGGCATAAGCTGAT\underline{ACATAGGAGGA}CGAATATGACTTGGAG$

Le plus long modèle trouvé dans les trois chaînes est le modèle :
$ACAAAGGAGGA$
situé à la position 11 de $s1$ (avec 1 insertion), 15 de $s2$ (avec 1 délétion) et 15 de $s3$ (avec 1 substitution).

Figure 3.26: Exemple d'un modèle et de ses L-occurrences.

Notons que, bien qu'établie de façon différente, cette définition est très similaire à celle utilisée par Waterman et introduite avant nous dans la littérature [Galas *et al.*, 1985] [Waterman, 1984b] [Waterman *et al.*, 1984] [Waterman, 1986] [Waterman, 1989] [Waterman, 1990]. Sa façon d'obtenir les modèles, et donc les ensembles de mots similaires, est toutefois très différente et d'une application pratique souvent plus limitée, ainsi que nous le montrons dans le chapitre 5.

3.5.2.1.2 Des modèles comme produits d'ensembles d'une couverture — une application aux protéines

Dans le cas des protéines, nous avons vu qu'il existe des relations de similarité entre les acides aminés (section 3.2.2.2) et que la complexité de telles relations est captée au mieux à l'aide d'une couverture. La définition d'un modèle que nous employons dans ce cas utilise donc le concept d'une couverture C de l'alphabet et nous avons [Sagot *et al.*, 1995d] :

Définition 3.5.4 *Un modèle M de longueur k est un élément de C^k. Nous notons $\mid M \mid$ la longueur de M.*

Avant d'indiquer quelle est la relation existant entre un modèle et des mots dans s, nous introduisons deux notations qui vont nous être utiles :

Notation 3.5.1 *Étant donné $u = u_1...u_k \in \Sigma^k$ et un modèle $M = M_1...M_k \in S^k$, nous disons que u est une <u>instance</u> de M si, pour tout $i \in \{1,...,k\}$, $u_i \in M_i$.*
En d'autres termes, u est une instance de M si $u \in M$.

Notation 3.5.2 *Étant donné $u \in \Sigma^*$ et un modèle $M \in S^k$, nous appelons <u>$distance_{EL}$</u> entre u et M, notée $distance_{EL}(u, M)$, le nombre minimum de substitutions, délétions et insertions nécessaires pour passer de u à une instance de M ($distance_{EL}$ signifie 'distance Ensembliste de Levenshtein').*

La relation entre un modèle et un mot est alors définie par :

Soient $\Sigma = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ et C la couverture de Σ donnée par :

$S_1 = \{I, L, V, M\}$
$S_2 = \{A, G, S, C, T\}$
$S_3 = \{K, R, H\}$
$S_4 = \{K, R\}$
$S_5 = \{Y, W, F, H\}$
$S_6 = \{Y, W, F\}$
$S_7 = \{D, E\}$
$S_8 = \{Q, N\}$
$S_9 = \{P\}$
$S_{10} = \{C\}$
$S_{11} = \{W\}$

Soient les quatre chaînes suivantes :

$s1 = VMGPIKDM\underline{VHITHGPIG}CSFYTWGGRRFKSKPENGTGLNFNEYV$
$s2 = CAYAGSKGVVWGPIKDM\underline{IHISHGPVG}CGQYSRAGRRNYYIGTTGV$
$s3 = EFCGGHTHAISRYGLEDMLPAN\underline{VRMIHGPGC}PVCVLPAGRIDMAI$
$s4 = MKLAHWMYAGPAHIGTLRVASSFKN\underline{VHAIMHAPLG}DDYFNVMR$

Les plus longs modèles trouvés dans ces chaînes sont les modèles suivants :

$S_1S_3S_1S_1S_5S_2S_9S_1S_2$

situé à la position 9 de $s1$ (avec 1 substitution), 18 de $s2$ (avec 1 substitution), 23 de $s3$ (deux fois, avec 1 substitution et 1 délétion) et 26 de $s4$ (avec 1 insertion), et :

$S_1S_3S_1S_1S_3S_2S_9S_1S_2$

situé à la position 9 de $s1$ (avec 1 substitution), 18 de $s2$ (avec 1 substitution), 23 de $s3$ (deux fois, avec 1 substitution et 1 délétion) et 26 de $s4$ (avec 1 insertion).

Figure 3.27: Exemple de modèles et de leurs EL-occurrences

Soient Σ l'alphabet des acides aminés et C la couverture donnée dans la figure 3.27. Supposons par ailleurs qu'aucune erreur ne soit autorisée entre un modèle et ses occurrences et que la contrainte de quorum q soit égale à 2.
Un exemple de situation où :

- $M_1 \subseteq M_2$ et
  - $EL(M_1)$ égal à $EL(M_2)$ est donnée par les chaînes :

    $s1 = CKD$
    $s2 = GRE$
    $s3 = SRD$

    et les modèles valides $S_2S_4S_7$ et $S_2S_3S_7$;
  - $EL(M_1)$ strictement inclus dans $EL(M_2)$ est donnée par les chaînes :

    $s1 = CKD$
    $s2 = GRE$
    $s3 = SHD$

    et les modèles valides $S_2S_4S_7$ et $S_2S_3S_7$;
- $M_1 \cap M_2 \neq \emptyset$ avec $M_1 \neq M_2$ et
  - $EL(M_1)$ égal à $EL(M_2)$ est donnée par les chaînes :

    $s1 = CHD$
    $s2 = GHE$
    $s3 = SHD$

    et les modèles valides $S_2S_3S_7$ et $S_2S_5S_7$;
  - $EL(M_1)$ strictement inclus dans $EL(M_2)$ est donnée par les chaînes :

    $s1 = CHD$
    $s2 = GHE$
    $s3 = SYD$

    et les modèles valides $S_2S_3S_7$ et $S_2S_5S_7$.

Figure 3.28: Exemple des rapports pouvant exister entre les ensembles d'occurrences des modèles présents dans un ensemble de chaînes.

Définition 3.5.5 *Étant donné un entier e et un modèle M, nous disons que M est* <u>*EL-présent*</u> *dans s à la position i s'il existe au moins un mot u situé à cette position dans s tel que* $distance_{EL}(u, M) \leq e$. *Nous appelons u une* <u>*EL-image*</u> *de M.*

Soit $d_s + d_d + d_i = distance_{EL}(u, M)$ *où* d_s, d_d *et* d_i *sont le nombre de substitutions, délétions et insertions respectivement entre u et l'instance la plus proche de M (à nouveau ici, par convention, nous notons* d_d *et* d_i *respectivement le nombre de délétions et insertions réalisées sur u). Observons que cette décomposition peut ne pas être unique. Nous disons que* (i,d_s,d_d,d_i) *est une* <u>*EL-occurrence*</u> *de M dans s et nous avons* $\mid u \mid = \mid M \mid - d_d + d_i$.

Finalement, nous notons EL(M) l'ensemble des EL-occurrences de M dans s. Nous avons alors :

$$EL(M) = \{(i,d_s,d_d,d_i) \mid d = d_s+d_d+d_i \text{ est une } distance_{EL} \text{ et } d \leq e\}.$$

Exemple 3.5.3 *Voir la figure 3.27.*

À nouveau ici, un modèle peut être EL-présent dans s à la position i plus d'une fois avec différentes EL-images (voir la figure 3.27), ou, pour la même image u, avec différentes EL-occurrences.

De même, la relation qui est établie entre un modèle et ses images induit une relation entre les images elles-mêmes, c'est-à-dire entre les mots de s, puisque la distance de Levenshtein, 'simple' cette fois, entre deux quelconques d'entre eux est au plus $2e$. La réciproque n'est pas vraie ainsi que nous l'avons montré dans la section précédente. L'exemple 3.5.2 donné alors s'applique également ici, il suffit de prendre $C = CI = \{\{a\} \mid a \in \Sigma\}$.

Quant à la couverture que nous utilisons, elle est *a priori* quelconque. En particulier, une couverture peut contenir un sous-ensemble qui est totalement inclus dans un autre comme cela est le cas des exemples donnés dans les figures 3.12 et 3.27. Ce qui n'est par contre absolument pas permis par la définition de similarité entre mots donnée dans cette section sont les couvertures qui contiennent comme un de leurs sous-ensembles tout l'alphabet — ce que nous avons appelé le joker (rappelons qu'en anglais il s'agit du 'wild card' ou du 'don't care symbol'). Il est aisé de voir pourquoi autoriser le joker représente un problème algorithmique complètement différent exigeant qu'une contrainte additionnelle soit imposée sur le nombre de fois où le joker est permis dans un modèle. Supposons en effet que le joker $\{\Sigma\}$ soit un des sous-ensembles de la couverture. Alors, étant donné un ensemble de chaînes et en faisant l'hypothèse qu'aucune erreur n'est autorisée, le plus long modèle qui est présent dans ces chaînes est trivialement le modèle $\{\Sigma\}^p$ où p est la longueur de la plus courte chaîne de l'ensemble. Des couvertures où le joker est permis mais avec une contrainte, et des couvertures plus combinatoires que celles que nous employons ici, y compris la couverture $\mathcal{P}(\Sigma)$, sont introduites dans la définition de la section suivante. Elles correspondent aux couvertures combinatoires pondérées mentionnées dans la section 3.2.3.3.1 et s'appliquent aussi bien à l'analyse des protéines qu'à celle des acides nucléiques.

Observons qu'il peut apparaître une certaine redondance de l'information captée par des modèles définis sur une couverture de l'alphabet. En effet, étant donné deux modèles M_1 et M_2, nous pouvons avoir les quatre situations suivantes :

- $M_1 \subseteq M_2$ et
  - $EL(M_1)$ égal à $EL(M_2)$;
    <u>ou</u>
  - $EL(M_1)$ strictement inclus dans $EL(M_2)$;

- $M_1 \cap M_2 \neq \emptyset$ avec $M_1 \neq M_2$ et
  - $EL(M_1)$ égal à $EL(M_2)$;
    ou
  - $EL(M_1)$ strictement inclus dans $EL(M_2)$.

Un exemple de chacun de ces cas est donné dans la figure 3.28. Nous pouvons ainsi définir deux types de non-maximalité de l'ensemble d'occurrences d'un modèle.

Définition 3.5.6 *Soient s1,...,sN des chaînes définies sur Σ^*, C une couverture de Σ et $M \in C^k$ un modèle de longueur k. Nous disons que :*

- *$EL(M)$ est intrinsèquement non-maximal s'il existe au moins un modèle $M' \in C^k$ tel que $M \subset M'$ et $EL(M) \subset EL(M')$;*
- *$EL(M)$ est non-maximal par rapport aux chaînes s1,...,sN s'il existe au moins un modèle $M' \in C^k$ tel que $M \cap M' \neq \emptyset$ avec $M \neq M'$ et $EL(M) \subset EL(M')$.*

Bien sûr, le premier point ne peut être vérifié que s'il existe S_i, $S_j \in C$ tel que $S_i \subset S_j$. Dans ce cas seulement, on peut définir un concept de non-minimalité d'un modèle par rapport à un autre.

Définition 3.5.7 *Soient M et M' deux modèles définis sur une couverture C de l'alphabet Σ. Si $M \subset M'$ et $EL(M) = EL(M')$, nous disons que le modèle M' est non-minimal par rapport au modèle M.*

Une telle redondance en termes d'ensembles d'occurrences n'est pas forcément mauvaise, au contraire, puisque nous sommes le plus souvent intéressés non seulement par les blocs de mots similaires que nous réussissons à identifier, mais aussi par leur représentation - les modèles. Exiger soit la maximalité d'un ensemble d'occurrences (intrinsèque ou liée aux chaînes que l'on compare), soit la minimalité d'un modèle peut être important cependant. Cela introduit toutefois une contrainte supplémentaire au niveau à la fois de la définition de similarité entre mots et de l'algorithme qui permet de localiser ces mots. Cette contrainte étant intimement liée à l'algorithme lui-même, elle est présentée dans le cas de la minimalité des modèles au chapitre 5. Nous verrons alors que la minimalité que nous obtenons en pratique est en fait plus restreinte que celle de la définition 3.5.7. Nous l'appelons 'minimalité gauche-droite'. Quant à la contrainte de maximalité sur les ensembles d'occurrences, nous en discutons plus loin lorsque nous montrons le rapport entre les modèles et les cliques maximales d'une relation R_k telles que définies dans la section 3.4.3.2. Ce rapport ne peut être établi bien sûr que si aucun élément de la couverture C n'est entièrement inclus dans un autre, ce qui en pratique sera le plus souvent le cas.

Des modèles qui sont des produits d'ensembles d'une couverture et la distance Ensembliste de Levenshtein nous fournissent donc une autre définition d'un groupe de mots similaires.

Définition 3.5.8 *Étant donné une couverture C de Σ, une chaîne $s \in \Sigma^n$ (qui est la concaténation de N chaînes), un entier q entre 2 et N représentant une contrainte de quorum et un autre entier k, les groupes de mots similaires dans s sont les ensembles de EL-occurrences de tous les modèles M de longueur k qui satisfont la contrainte q.*

Observons finalement qu'un exemple particulier d'une couverture est celle composée uniquement de singletons. Il est clair qu'il n'existe pas vraiment de différence entre des modèles comme produits cartésiens d'ensembles d'une telle couverture et des modèles définis comme des mots comme dans la section précédente. En termes des méthodes utilisées pour trouver les groupes de mots similaires, les idées de base sont exactement les mêmes pour les deux définitions et seule la version la plus générale de l'algorithme pour des modèles comme produits cartésiens d'ensembles d'une couverture (non combinatoire et non pondérée) est présentée dans le chapitre 5.

3.5.2.1.3 Des modèles comme produits d'ensembles d'une couverture combinatoire pondérée — une application aux protéines et aux acides nucléiques

Nous allons introduire dans cette section une définition de ressemblance utilisant des modèles définis cette fois sur une couverture quelconque, qui peut aussi bien être $\mathcal{P}(\Sigma)$. La motivation est double. La première, et sans doute la plus importante, est de nature biologique. Elle représente le fait que même lorsqu'une activité macromoléculaire en général est identifiable à un ensemble de mots, et donc caractérisable par un modèle, toutes les positions dans cet ensemble de mots peuvent ne pas être impliquées dans l'activité en question. Cela a été maintes fois observé, entre autres par Posfai [Posfai *et al.*, 1989], Bashford [Bashford *et al.*, 1987], Rooman [Rooman *et al.*, 1990] [Rooman and Wodak, 1988], Smith [Smith *et al.*, 1990] et Neuwald [Neuwald and Green, 1994] dans le cas des protéines. Le modèle peut ainsi contenir des positions qu'on appelle non-spécifiques, c'est-à-dire des positions où la conservation d'un monomère, ou d'un type particulier de monomères, n'est pas essentielle. Autoriser la présence du joker dans un modèle permet de capter cette non-spécificité de certaines positions. La seconde motivation est liée à l'idée mentionnée dans la section 3.2.3.3.1 d'essayer de se libérer de tout point de vue dans la recherche de mots communs à un ensemble de chaînes. Cela signifie en particulier travailler avec plusieurs alphabets en même temps. Nous introduisons ici les outils nécessaires dans ce but. Le principal est ce que nous avons appelé une couverture combinatoire pondérée (voir la fin de la section 3.2.3.3.1). Les modèles que nous employons dans cette nouvelle définition de similarité entre mots sont donc des produits cartésiens d'ensembles d'une telle couverture [Sagot and Viari, 1996].

Définition 3.5.9 *Soit Σ l'alphabet des monomères. Une couverture combinatoire pondérée CCP de Σ est définie par :*

$$CCP = \{(S, w_S) \mid S \in \mathcal{P}^+(\Sigma) \text{ et } w_S \text{ est un entier non négatif}\} \text{ où } \mathcal{P}^+(\Sigma) = \mathcal{P}(\Sigma)\backslash\emptyset.$$

Définition 3.5.10 *Soit CCP une couverture combinatoire pondérée de Σ, un modèle $M = S_1...S_k$ est un élément de CCP^k si, pour tout $1 \leq i \leq k$, $S_i \in \mathcal{P}^+(\Sigma)$ et le nombre de fois que S_i apparaît dans M n'est pas supérieur à w_{S_i}.*

Exemple 3.5.4 *Soit $\Sigma = \{A, C, G, T\}$ et :*

$$CCP = \left\{ \begin{array}{l} (\{X\}, \infty)\ \forall X \in \Sigma, \\ (\{X, Y\}, 1)\ \forall X, Y \in \Sigma, \\ (\{X, Y, Z\}, 1)\ \forall X, Y, Z \in \Sigma, \\ (\{A, C, G, T\}, 1) \end{array} \right\}$$

Alors nous avons :

$\{A, C\}\{A, G\}\{A, T\}\{C, G\}\{C, T\}\{G, T\} \in CCP^6$

mais par exemple les modèles :

Soient $\Sigma = \{A, C, G, T\}$ et :

$$CCP = \left\{ \begin{array}{l} (\{X\}, \infty)\ \forall X \in \Sigma, \\ (\{X, Y\}, 1)\ \forall X, Y \in \Sigma, \\ (\{X, Y, Z\}, 1)\ \forall X, Y, Z \in \Sigma, \\ (\{A, C, G, T\}, 1) \end{array} \right\}.$$

Soient les trois chaînes :

$s1 = AAC\underline{TGTGAGCATGGTCA}TATTT$
$s2 = TAA\underline{TGTGAGTTAGCTCA}CTCAT$
$s3 = AAT\underline{TGTGACACAGTGCA}AATTC$

Un plus long modèle trouvé dans les trois chaînes est le modèle :

$\{T\}\{G\}\{T\}\{G\}\{A\}\{C, G\}\{A, C, T\}\{A, C, G, T\}\{A, T\}\{G\}\{C, G, T\}\{G, T\}\{C\}\{A\}$
situé à la position 4 de $s1$, $s2$ et $s3$.

Figure 3.29: Exemple d'un modèle et de ses ECL-occurrences.

$\{A, C\}\{A, C\}\{A, T\}\{C, G\}\{C, T\}\{G, T\}$
et
$\{A, C\}\{A, G\}\{A, T\}\{C, T\}\{A, C, G, T\}\{A, C, G, T\}$
ne sont pas des éléments de CCP^6 *(parce que les ensembles* $\{A, C\}$ *et* $\{A, C, G, T\}$ *respectivement sont présents deux fois dans le modèle).*

Définition 3.5.11 *Étant donné une couverture combinatoire contrainte* CCP*, une chaîne* $s \in \Sigma^*$ *et un modèle* $M \in CCP^k$*,* M *est dit être* <u>*ECL-présent*</u> *à la position* i *si le mot* u *de longueur* $|M|$ *situé à cette position dans* s *est tel que* $u \in M$*.* u *est appelé une* <u>*ECL-image*</u> *de* M *et nous disons que* (i, u) *est une* <u>*ECL-occurrence*</u> *de* M *dans* s*.*

Nous notons ECL(M) l'ensemble des occurrences de M *dans* s*, à savoir :*

$$ECL(M) = \{(i, u = s_i ... s_{i+|M|-1}) \mid u \in M\}$$

Observons que la notation ECL vient de 'Ensembliste Combinatoire de Levenshtein', le L étant justifié par le fait que, bien qu'il n'y ait pas de notion d'erreur (et donc de distance) dans cette définition, nous pouvons en introduire une, ce que nous ferons dans le chapitre 5 (section 5.4.2.3).

Exemple 3.5.5 *Voir la figure 3.29.*

Des modèles comme produits particuliers d'ensembles d'une couverture combinatoire pondérée nous donnent alors une nouvelle définition de la similarité entre mots.

Définition 3.5.12 *Étant donné une couverture combinatoire pondérée* CCP *de* Σ*, une chaîne* $s \in \Sigma^n$ *(qui est la concaténation de* N *chaînes), un entier* q *entre 2 et* N *représentant une* *contrainte de quorum et un autre entier k, les groupes de mots similaires dans s sont les ensembles de ECL-occurrences de tous les modèles M de longueur k qui satisfont la contrainte q.*

Observons que cette définition ressemble beaucoup à la précédente mais qu'algorithmiquement les façons de construire les modèles et obtenir ainsi les groupes de mots similaires sont différentes et sont donc présentées séparément dans le chapitre 5.

Il est très important de noter cependant également que les deux définitions peuvent être regroupées, en particulier que des erreurs peuvent être autorisées entre un modèle M défini à partir d'une couverture combinatoire pondérée et ses ECL-occurrences. Pour des raisons de clarté à nouveau, nous avons maintenu les deux définitions séparées.

3.5.2.2 Définition utilisant une notion de similarité basée sur des sous-mots

Avec la dernière définition de similarité entre mots que nous présentons maintenant, nous revenons à des modèles plus simples puisque ceux-ci vont à nouveau être des mots sur le même alphabet que celui des chaînes que nous comparons [Sagot *et al.*, 1995c]. La définition d'un modèle utilisée dans cette section est donc identique à celle de la section 3.5.2.1.1.

La raison de ce retour en arrière est que nous allons travailler cette fois avec une notion de similarité entre mots (qui peut aussi être une distance) qui est elle-même établie sur des sous-mots d'une certaine longueur w - et non plus sur un simple comptage de symboles identiques (ou différents). La motivation a été donnée dans la section 3.2.4.1.1 où nous avons montré les deux philosophies différentes de la mesure de ressemblance. Les définitions antérieures suivent une de ces philosophies, celle d'une ressemblance basée sur un comptage d'erreurs ou d'identités, la définition de cette section s'appuie sur l'autre qui utilise un système de score et est en général plus flexible ainsi que le suggérait la figure 3.14.

Dans l'exemple de cette figure, les trous ne sont pas permis entre les mots. Nous commençons donc par présenter une définition où les trous ne sont pas pris en considération. Cela est fait dans le but de simplifier la présentation. La définition permettant des trous est plus complexe et est donnée à la fin de cette section.

Observons que l'idée de cette définition dans son cas le plus simple ressemble beaucoup à celle employée pour effectuer des comparaisons deux à deux par le plus populaire des programmes de recherche rapide sur banque, appelé BLAST [Altschul *et al.*, 1990]. Nous appelons de tels programmes des cribleurs (en anglais 'scans'). Ici nous l'appliquons au cas multiple dans le cadre d'un ensemble de chaînes. Nous présentons BLAST dans la section 4.2.2.1.2.1.

Définition 3.5.13 *Soient $u = u_1u_2...u_k \in \Sigma^k$, $m = m_1m_2...m_k$ un modèle de longueur k et $\mathcal{M}$ une matrice de similarité, nous notons :*

$$score_{\mathcal{M}}(u, m) = \sum_{i=1}^{k} \mathcal{M}(u_i, m_i).$$

Définition 3.5.14 *Étant donné une valeur numérique t, un modèle m et un entier w entre* 1 *et $\mid m \mid$, nous disons que m est <u>w-présent</u> dans s à la position i si le mot u de longueur $\mid m \mid$ situé à cette position dans s est tel que :*

$$score_{\mathcal{M}}(u_j...u_{j+w-1}, m_j...m_{j+w-1}) \geq t$$

pour tout $j \in \{1,...,\mid m \mid -w+1\}$, c'est-à-dire si tout sous-mot de longueur w de u possède un score supérieur ou égal à t avec le sous-mot correspondant de m.

Soient Σ l'alphabet des acides aminés, $\mathcal{M}$ la matrice PAM250 de la figure 3.7, $w = 3$ et $t = 60$. Soient les trois chaînes :

$s1 = MTDKAPQKT\underline{HLNIVIICG}G$

$s2 = MGKEK\underline{HINLVVIWW}WGHVN$

$s3 = MKMPKDK\underline{HVNIVFICV}GEE$

alors le plus long modèle qui est w-présent dans les trois chaînes est le modèle $HINIVVICW$ de longueur 9 (à la position 10 de $s1$, 6 de $s2$ et 8 de $s3$).

Figure 3.30: Un exemple de modèle et de ses w-occurrences.

Nous appelons u une <u>*w-image*</u> *de m.*

Nous disons que (i, u) *est une* <u>*w-occurrence*</u> *de m dans s et nous notons* $K(m)$ *l'ensemble de toutes les w-occurrences de m dans s, à savoir,*

$$K(m) = \{(i, u) \mid \forall j \in \{1, ..., \mid m \mid -w+1\}, score_{\mathcal{M}}(u_j...u_{j+w-1}, m_j...m_{j+w-1}) \geq t\}.$$

Finalement, nous notons $l_w(u, m) = score_{\mathcal{M}}(u_{|m|-w+1}...u_{|m|}, m_{|m|-w+1}...m_{|m|})$. *Il s'agit simplement du score du dernier sous-mot de longueur w de u et m.*

Exemple 3.5.6 *Voir la figure 3.30.*

Observons que lorsque la matrice utilisée est une matrice de distance (dans ce cas, l'inégalité dans la définition 3.5.14 est renversée), la relation qui est établie entre un modèle et ses w-images induit une relation entre les images elles-mêmes à cause de l'inégalité triangulaire.

Les considérations précédentes nous fournissent alors une dernière définition de la similarité entre un groupe de mots.

Définition 3.5.15 *Étant donné une chaîne* $s \in \Sigma^n$ *(qui est la concaténation de N chaînes), un entier q entre 2 et N représentant une contrainte de quorum, un autre entier k et une constante w entre 1 et k, les groupes de mots similaires de longueur k dans s sont les ensembles de w-occurrences de tous les modèles m de longueur k qui satisfont cette contrainte.*

Que se passe-t-il lorsque des trous sont permis? Rappelons (voir section 3.2.4.1.2) que nous travaillons avec des scores constants pour les trous.

Notation 3.5.3 *Soit* $\mathcal{A}$ *un alignement avec trous d'un modèle m avec un mot u. On appelle longueur d'une paire de sous-mots de* $\mathcal{A}$ *la longueur du sous-mot dans m.*

Exemple 3.5.7 *Soit l'alignement* $\mathcal{A} = \begin{pmatrix} a \\ a \end{pmatrix} \begin{pmatrix} b \\ \lambda \end{pmatrix} \begin{pmatrix} a \\ a \end{pmatrix} \begin{pmatrix} c \\ c \end{pmatrix} \begin{pmatrix} \lambda \\ a \end{pmatrix} \begin{pmatrix} a \\ a \end{pmatrix}$ *du modèle* $m = abaca$ *avec le mot* $u = aacaa$. *Les paires de sous-mots de longueur 3 de* $\mathcal{A}$ *sont les paires* (aba, aa), (bac, ac) *et* $(aca, acaa)$.

Définition 3.5.16 *Étant donné une valeur numérique t, un modèle m et trois entiers w,* e_s *et* e_t *avec* $1 \leq w \leq \mid m \mid$, $0 \leq e_s < w$ *et* $e_s \leq e_t$, *nous disons que m est w-présent dans s à la* *position i s'il existe un mot u situé à cette position dans s tel que l'alignement $\mathcal{A}$ de m et u ne comporte au total pas plus de e_t trous, tandis que celui de toute paire de sous-mots de longueur w de m de l'alignement $\mathcal{A}$ ne présente pas plus de e_s trous et a un score supérieur ou égal à t (où le score d'un alignement avec trou est celui donné dans la définition 3.2.20)*

Exemple 3.5.8 *En reprennant les données de la figure 3.30 avec de plus $e_s = 1$, $e_t = 2$ et w_λ (score d'un trou) $= 0$, le mot aacaa est une occurrence du modèle abaca car l'alignement $\mathcal{A} = \begin{pmatrix} a \\ a \end{pmatrix}\begin{pmatrix} b \\ \lambda \end{pmatrix}\begin{pmatrix} a \\ a \end{pmatrix}\begin{pmatrix} c \\ c \end{pmatrix}\begin{pmatrix} \lambda \\ a \end{pmatrix}\begin{pmatrix} a \\ a \end{pmatrix}$ possède deux trous en tout, celui de chaque sous-mot de longueur w de m présente au plus un trou et a un score égal à* 6 *pour la paire (aba, aa), à* 6 *pour la paire (bac, ac) et à* 9 *pour la paire $(aca, acaa)$.*

Exemple 3.5.9 *Un second exemple reprend celui de la figure 3.30 avec de plus $e_s = 1$, $e_t = \infty$ et* 0 *le score attribué à un trou. Le plus long modèle qui est w-présent dans les trois chaînes de la figure est alors le modèle $KQHINIVVICWG$ de longueur* 13 *(à la position* 6 *de s*1 *avec une insertion,* 3 *de s*2 *avec également une insertion et* 5 *de s*3 *sans trou).*

3.5.3 Résumé des mesures indirectes (et locales) de la ressemblance

Un résumé des mesures indirectes et locales de la ressemblance est donné dans le tableau ci-dessous.

Mesures Indirectes et Locales de la Ressemblance		
Types de modèles	Caractéristiques principales	Référence aux définitions
mots	distance de Levenshtein	définition 3.5.3
produits d'ensembles d'une couverture	distance ensembliste de Levenshtein	définition 3.5.8
produits d'ensembles d'une couverture combinatoire pondérée	appartenance au modèle	définition 3.5.12
mots	scores des sous-mots supérieurs à seuil	définition 3.5.15

3.5.4 Les mesures directes revues

Lorsque nous avons présenté les mesures directes dans la section 3.4, nous avons montré qu'une des approches possibles pour localiser des blocs de similarité dans un ensemble de chaînes consiste à travailler avec une relation R et avec les cliques maximales de cette relation. Aucun modèle n'est nécessaire et les mots sont en fait comparés directement entre eux. Implicitement cependant, les modèles sont présents.

En effet, si nous représentons un modèle comme une boîte avec une étiquette dans laquelle certains objets (des positions dans une chaîne s) sont déposés, nous pouvons voir les cliques maximales comme des boîtes noires qui n'ont pas besoin d'être étiquetées pour remplir leur fonction de regroupement des mots similaires. Et pourtant, si nous examinons la définition d'une clique maximale C_k, nous constatons qu'une telle clique possède par rapport à s un nom unique même si ce nom n'est pas utilisé. Ce nom est celui du k-produit des cliques de R de plus petite cardinalité qui le composent.

Supposons que nous travaillons avec la définition 3.5.3 de similarité. Soient $\Sigma = \{a, b\}$, $q = 4$ et $e = 2$ (au plus deux substitutions sont autorisées entre un modèle et ses occurrences). Soient les quatre chaînes :

$s1 = aaa$
$s2 = aaa$
$s3 = aaa$
$s4 = bbb.$

Alors les modèles aab, aba, baa, abb, bba et bbb sont tous des modèles de longueur 3 valables d'après la définition 3.5.3 (ce sont les seuls) mais le centre de gravité de l'ensemble d'occurrences de ces modèles est le mot aaa qui n'est pas un modèle valide (il n'est pas présent sur la dernière chaîne).

Figure 3.31: Un exemple indiquant pourquoi, étant donné un ensemble de mots similaires (i.e. d'occurrences), il n'existe pas toujours un modèle qui représente un centre de gravité ou un mot de Steiner pour cet ensemble.

Exemple 3.5.10 *En reprennant les données de la figure 3.23, le nom de l'unique clique maximale de R_2 par rapport à s est égal à $\{L, M\}\{I, L\}$.*

Nous pouvons donc considérer que la définition 3.4.10 de similarité entre mots n'est qu'un cas particulier de la définition 3.5.8 puisque les cliques maximales d'une relation R (et les classes) ne sont elles-mêmes que des cas particuliers d'une couverture de Σ. Il n'y a qu'une réserve à faire à cette dernière affirmation. Elle a trait à la propriété de maximalité dont nous avons parlé dans la section 3.5.2.1.2. Supposons que nous utilisons une couverture de Σ composée de cliques maximales, et qu'aucune erreur ne soit autorisée entre un modèle et ses occurrences. Étant donné un ensemble de chaînes $s1$,$s2$,...,sN et un modèle de longueur k qui est EL-présent avec 0 erreurs dans q de ces chaînes, nous ne pouvons garantir que son ensemble d'occurrences soit une clique maximale dans le sens où il ne contiendrait l'ensemble d'occurrences d'aucun autre modèle de même longueur. L'exemple de la figure 3.23 illustre ce point à nouveau puisque $\{I, L\}\{I, L\}$ et $\{L, M\}\{I, L\}$ sont tous deux des modèles de longueur 2 qui sont valables suivant la définition 3.5.8, mais en termes d'ensembles d'occurrences, le premier est complètement inclus dans le second et ne peut donc être une clique maximale. Notons qu'il s'agit ici d'une non-maximalité par rapport aux chaînes qui sont comparées et non intrinsèque (voir définition 3.5.6). En termes pratiques cela signifie cependant simplement que, si ce sont les cliques maximales qui nous intéressent, il faut pendant tout le processus d'identification des groupes de mots similaires d'un ensemble de chaînes, tester à chaque étape que les modèles obtenus soient maximaux en tant qu'ensembles d'occurrences. À part ce fait, nous pouvons considérer alors que les cliques maximales (et les classes d'équivalences) sont aussi des modèles, c'est-à-dire des produits d'ensembles d'une couverture de Σ.

Qu'en est-il des deux autres mesures directes, s'appuyant soit sur l'optimisation d'une fonction de score, soit sur des calculs statistiques? La première présente un point commun important avec les mesures indirectes et avec celle établie à partir d'une relation R. En effet, la définition de ressemblance entre mots qu'elle utilise conduit aussi à une construction par récurrence des blocs de similarité. Par ailleurs, l'idée d'une optimisation reste dissimulée derrière les mesures qui utilisent des modèles mais, d'une part, la façon de calculer les scores n'est plus la même et, d'autre part, les modèles ne sont pas nécessairement optimaux. Vérifient-ils une quelconque propriété de maximalité (ou de minimalité)? Le plus intéressant sans doute serait de pouvoir affirmer que les modèles sont des centres de gravité [Waterman, 1995] ou des mots de Steiner [Gusfield, 1995] pour leurs occurrences. Cela serait le cas si la somme des mesures de ressemblance entre un modèle et chacune de ses occurrences était maximale ou minimale. Toutefois cela n'est pas vrai pour tous les modèles. Il n'est même pas possible d'affirmer que, étant donné un ensemble de mots similaires (i.e. un ensemble d'occurrences), il existe toujours au moins un modèle représentant un centre de gravité ou un mot de Steiner. L'exemple de la figure 3.31 illustre ce point. Observons cependant que, au moins en ce qui concerne les définitions de ressemblance 3.5.3 et 3.5.8 présentées plus haut, les ensembles d'occurrences trouvés par les modèles peuvent être plus consistants (voir la section 3.3.2) que les ensembles de mots localement similaires repérés par programmation dynamique (voir la section 3.4.3.1). En effet, dans le premier cas la distance entre deux occurrences d'un même modèle est toujours d'au plus 2e, alors que dans le second la distance entre deux mots appartenant à un ensemble plus grand de mots localement similaires peut être faible puisque la similarité a pour base un score SP, c'est-à-dire la somme de tous les scores deux à deux.

Enfin, existe-t-il un point commun du même genre entre la définition 3.4.11 de similarité établie à partir de critères statistiques et les quatre dernières utilisant des modèles? En général non, bien que cela puisse dépendre la formule utilisée pour estimer la similarité d'un groupe de mots de manière statistique. Les deux mesures sont très différentes par contre dans la façon d'obtenir les blocs de similarité ainsi que nous l'avons vu. En particulier, les méthodes stochastiques ne construisent pas les blocs par récurrence et ne sont pas exactes.

3.5.5 Diagramme résumant toutes les définitions de ressemblances

Le diagramme de la figure 3.32 montre les rapports entre toutes ces mesures de ressemblance.

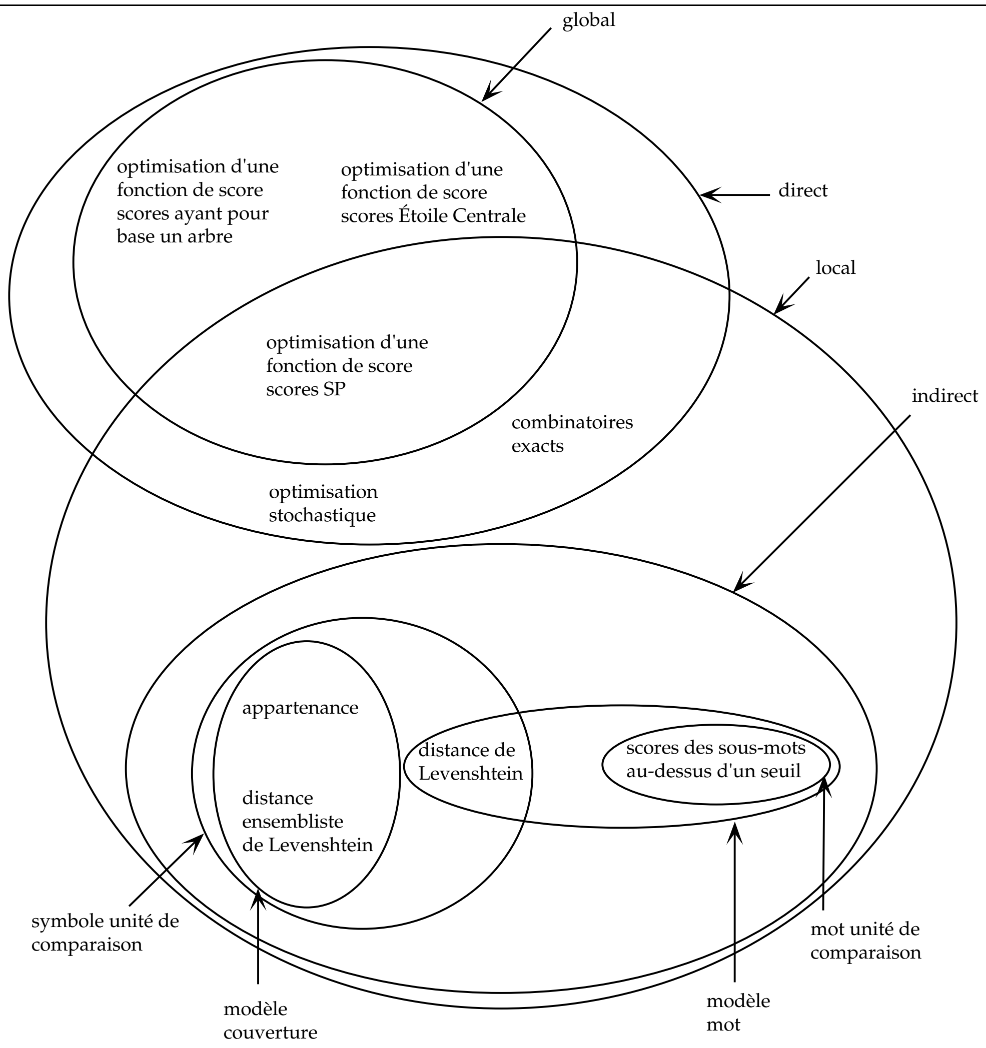


Figure 3.32: Diagramme résumant toutes les définitions de ressemblance.