CHAPITRE III. ANALYSE DE SÉQUENCES DE BIOMOLÉCULES

1. Introduction

En bioinformatique, la comparaison des séquences (ADN, ARN et/ou protéines : ARNm, régions 5'UTR, les EST, des clones, …) repose essentiellement sur la notion de l'alignement, et permet de déterminer le degré de ressemblance entre celles-ci (similitude ou identité en révélant des régions proches dans leurs séquences primaires). Cela peut alors indiquer que :

- La structure (primaire, secondaire ou tertiaire) des deux séquences est semblable,
- La fonction biologique est proche ou différente (dans le cas de la dissémilarité),
- L'origine des séquences alignées est commune ou éloignée (notion d'homologie), …

Cependant, la comparaison pour l'obtention d'un alignement optimal entre deux séquences biologiques, nécessite néanmoins la mise en œuvre de procédures de calcul (algorithmes) et de modèles biologiques permettant de quantifier la notion de ressemblance entre ces séquences.

2. Alignement des séquences

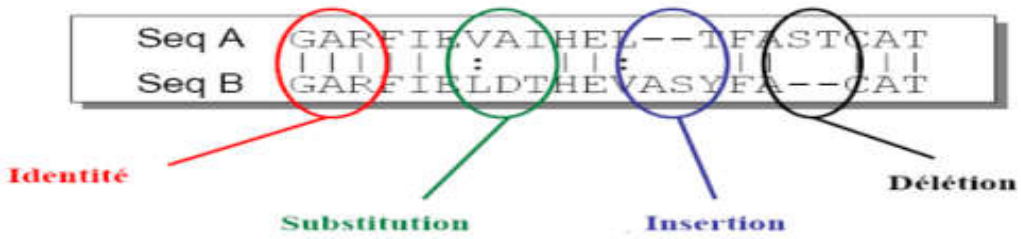
2.1. Qu'est-ce qu'un alignement

Est une manière de représenter deux ou plusieurs séquences de macromolécules biologiques (ADN, ARN ou protéines) les unes sous les autres, de manière à en faire ressortir les régions homologues ou similaires.

Rechercher le maximum d'appariements entre les résidus des séquences comparées. L'alignement est d'autant plus parfait qu'il n'y a pas de mésappariements et de brèches (insertions ou délétions).

❖Les caractères sont les mêmes : **identité = match** (en anglais)

❖Les caractères ne sont pas les mêmes : **substitution : mismach** (en anglais)



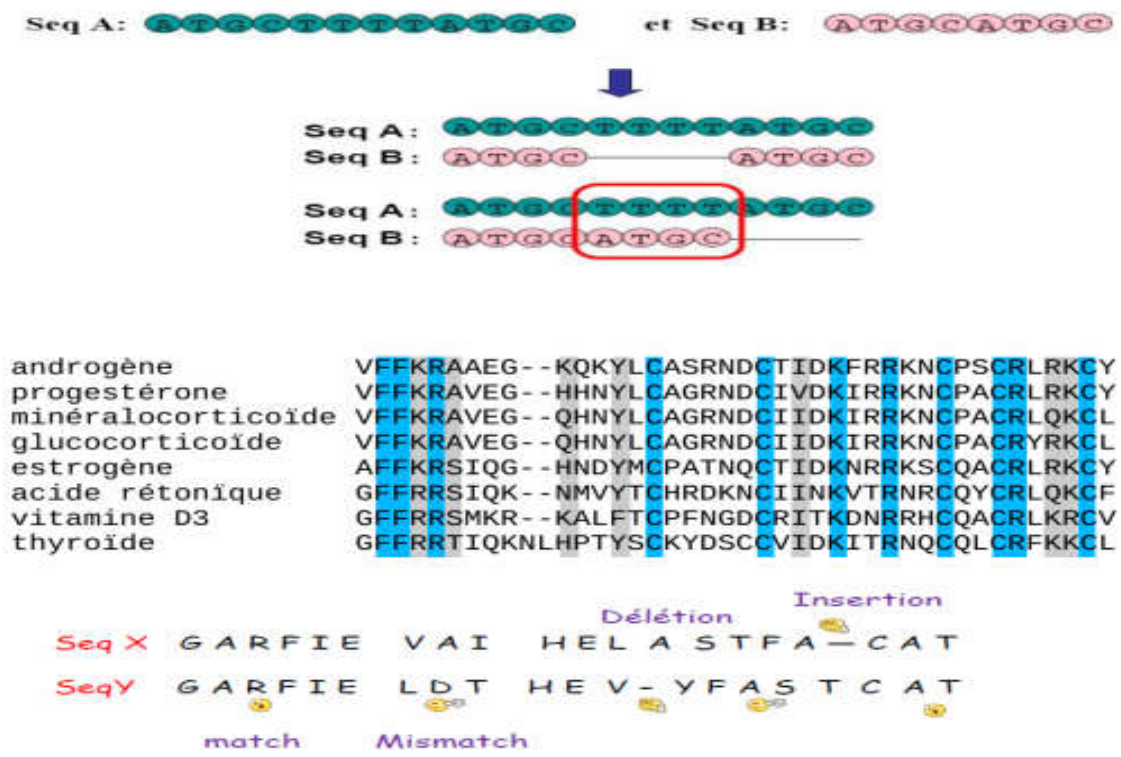


Figure 1: Exemple d'alignement

2.2. Les principes de bases pour identifier les ressemblances entre deux séquences

L'identité, la similitude et l'alignement

Les programmes de comparaison de séquences ont pour but de repérer les endroits où se trouvent des régions identiques ou très proches entre deux séquences et d'en déduire celles qui sont significatives et qui correspondent à un sens biologique de celles qui sont observées par hasard. En générale les algorithmes fonctionnent sur des segments de séquences (on parle de fenêtres, de motifs ou de mots) sur lesquels on regarde s'il existe ou pas une similitude significative. Si on ne prend en compte que des analogies entre sous-séquences sans traiter la possibilité d'insertion ou de délétion, on parlera alors de segments similaires. On distingue pour cette catégorie deux classes précises de similitude : la ressemblance parfaite ou identité et la ressemblance non parfaite que l'on qualifie de similitude (figure 2).

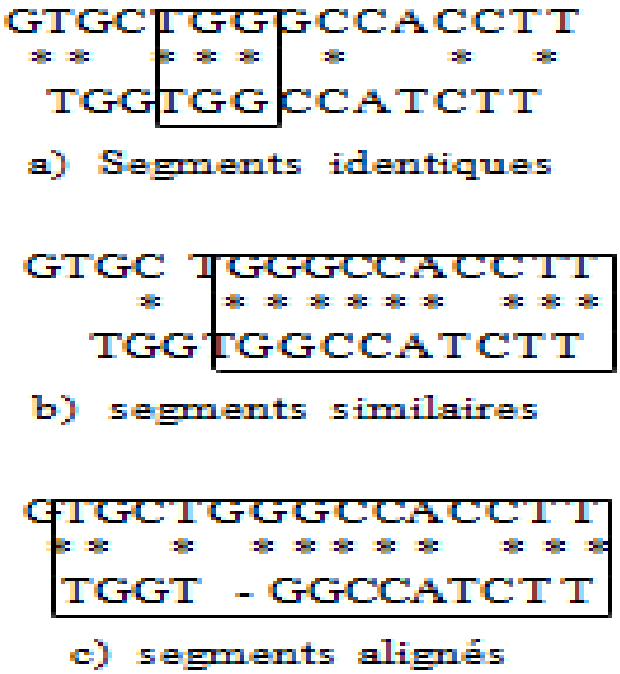


Figure 2: Les trois grandes classes de ressemblance issues de la comparaison de séquences.

3. Motifs dans les séquences

3.1. Pourquoi la recherche de motifs ?

Les motifs sont souvent recherchés dans des séquences car ils sont généralement impliqués dans des systèmes de régulation ou ils définissent des fonctions biologiques comme la détermination de la fonction d'une nouvelle séquence (par exemple en localisant un ou plusieurs motifs répertoriés dans des bases de motifs), l'identification dans une séquence nucléique de régions codantes, ou bien l'extraction à partir des banques de données des séquences possédant le même signal de régulation ou la même signature protéique pour effectuer des études comparatives ultérieures.

3.2. Qu'est-ce qu'un motif ?

Un "motif" (ou «pattern» en Anglais) est un segment court dans une séquence, il est continu et non ambigu. Il peut représenter une structure plus complexe lorsque lui-même est composé de différents "motifs" qui peuvent être plus ou moins éloignés les uns des autres et sa définition peut comporter des exclusions ou des associations de "motifs".

Plusieurs méthodes ont été imaginées pour identifier des éléments fonctionnels en utilisant leur conservation en séquence. La recherche de motifs conservés peut se faire à partir d'alignements multiples par recherche de blocs conservés dans l'alignement ou directement à partir de la séquence par des méthodes qui à la fois recherchent et déterminent des consensus. Ces dernières méthodes sont à la base des techniques d'alignement multiple "par blocs".

Dans notre travail, nous nous intéressons à la recherche de motif par la technique des MMC (HMM).

Exemple de recherche de motif:

Séquence : C T G T G T G T A C A T G T G de longueur 15

Motif : T G T G de longueur 4

Position : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Séquence : C T G T G T G T A C A T G T G

Solution**:** Ensemble de positions: *{2, 4, 12}*.

3.3. Qu'est- ce qu'un consensus ?

Chaîne de caractère indiquant les résidus conservés à chaque colonne d'un alignement multiple.

Le consensus est obtenu en retenant, pour chaque colonne d'un alignement multiple, soit un seul résidu (on parle alors de consensus strict, soit une combinaison de résidus représentatifs (consensus dégénéré). Les consensus dégénérés peuvent être représentés par des expressions régulières, combinées avec les spécifications IUPAC pour les résidus ambigüs.

Un consensus fournit une représentation compacte d'un motif séquentiel. Les consensus sont par exemple utilisés :

- pour les séquences nucléiques, afin de représenter les motifs de liaison de facteurs transcriptionnels sur les séquences d'ADN,

- pour des séquences peptidiques, afin de représenter les caractéristiques de domaines conservés au sein de familles de protéines homologues.
- Le consensus fournit une représentation compacte et intuitive d'un motif, mais souffre de quelques limitations.
- Les règles appliquées pour décider si l'on représentera une colonne de l'alignement par un résidu unique ou une combinaison de résidus sont floues, et varient selon les auteurs.
- En cas de positions ambigües, le consensus indique quels résidus sont acceptés, mais n'indique pas la fréquence relative de ces résidus dans les alignements initiaux (contrairement aux profils, ou matrices position-poids).

3.4. Expressions régulières

Une *expression régulière* est une chaîne de caractères qui décrit un *motif* (ou *pattern*) composé de différents types d'éléments :

- Caractères parfaitement déterminés (composés dans l'alphabet des acides aminés ou nucléotides selon le cas).
- Une série de lettres entre crochets [ ] indique une liste de résidus alternatifs acceptés à cette position (résidus partiellement déterminés).
- Un nombre entre accolades représente la répétition d'un caractère. Le nombre de répétitions peut être fixe (ex: A {4}=AAAA) ou variable (N {3,20} signifie "un nombre de nucléotides variant entre 3 et 20").
- Pour les motifs peptidiques, une série de résidus entre accolades signifie "tout sauf ces résidus". Par exemple {A, L} signifie "n'importe quel acide aminé sauf A et L".

Exemples:

- L'expression régulière: GAT [AT] AG signifie "la séquence GAT suivie soit d'un A soit d'un T, lui-même suivi de AG".
- Pour des séquences nucléiques, l'expression régulière CCGn {11} CCG signifie "la séquence CCG suivie de 11 résidus indifférents (A, C, G ou T), suivis d'une séquence CGG "
- Pour des séquences peptidiques, l'expression régulière [AV]-x-L {2} décrit un motif qui commence par soit Alanine soit Valine, ([AV]) suivi par n'importe quel acide aminé (x), suivi par 2 Leucines (L {2}).

4. L'alignement des séquences protéiques

4.1. Alignement de deux séquences

L'alignement de deux séquences permet d'identifier les mutations qui ont eu lieu lors de l'évolution. Ces évènements sont à l'origine de la divergence des séquences .

Un alignement de deux séquences (appelé souvent Alignement deux à deux) est une mise en correspondance entre les résidus avec une possible insertion des espaces (gaps) afin d'obtenir des séquences de longueur égales. Toutes les correspondances sont autorisées à condition que l'ordre des résidus soit respecté.

Trois situations sont possibles pour une position donnée de l'alignement:

- Les caractères sont les mêmes : identité
- Les caractères ne sont pas les mêmes : substitution
- L'une des positions est un gap (espace) : Insertion\Délétion .

Exemple d'alignement de deux séquences:

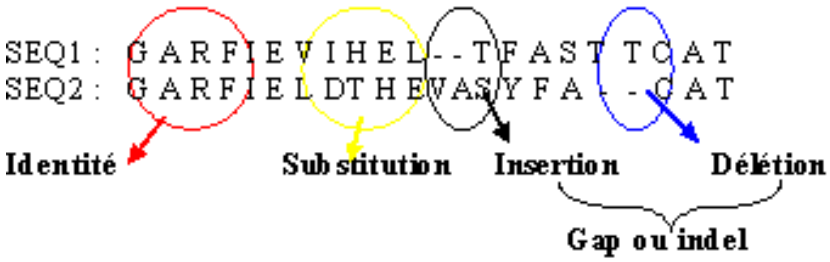


Figure 3: Alignement de deux séquences

4.2.Les méthodes d'alignement de deux séquences

Il existe deux types d'alignements de séquences: global et local.

Le premier prend en considération l'ensemble des résidus de chacune des séquences. Si les longueurs des séquences sont différentes, alors la plus courtes va subir des insertions de gaps afin d'arriver à aligner les deux séquences d'une extrémité à l'autre. Cependant dans un alignement global, si uniquement des segments courts sont très similaires entre deux séquences, les autres parties des séquences risquent de diminuer le poids de ces régions. C'est pourquoi d'autres

algorithmes d'alignements, dits locaux, basés sur la localisation des zones de similarité sont nés. Le but de ces alignements locaux est de trouver sans prédétermination de longueur les zones les plus similaires entre deux séquences. L'alignement local comporte donc une partie de chacune des séquences et non la totalité des séquences comme dans la plupart des alignements globaux.

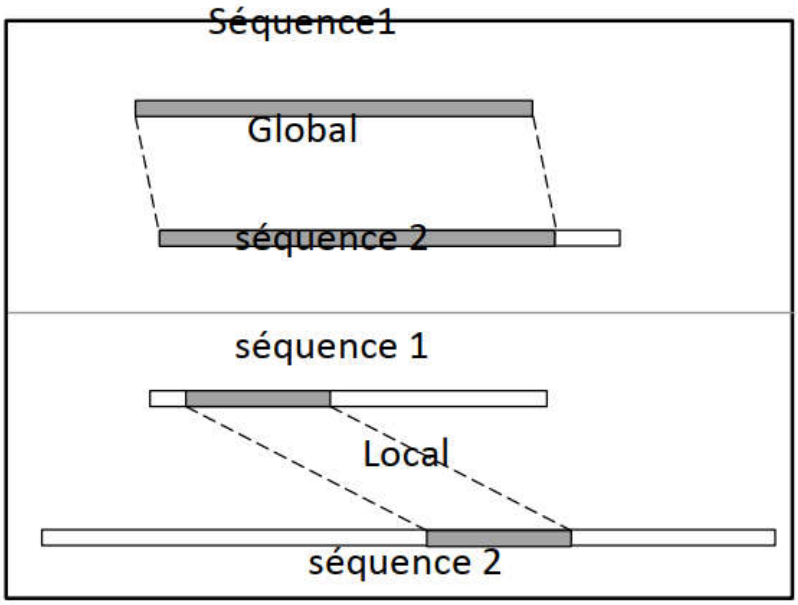


Figure 4 : Alignement local et global .

Cependant, il est clair que pour deux séquences données quel conques il y a plusieurs alignements possibles. Il est devenu alors nécessaire de pouvoir déterminer quel est le meilleur alignement ou plutôt l'optimal si possible.

4.3.Les alignements multiples

L'alignement multiple de séquences (Multiple Sequence Alignment: MSA) est une tâche cruciale et très importante en biologie moléculaire. MSA offre aux biologistes un moyen pour analyser des séquences d'ADN ou de protéines et de déterminer par la suite leur degré d'homologie ou de divergence. MSA est utilisé dans la construction des arbres phylogénétiques et identifier les motifs dans des familles de protéines, ceci permet de prédire leur aspect structurel et fonctionnel.

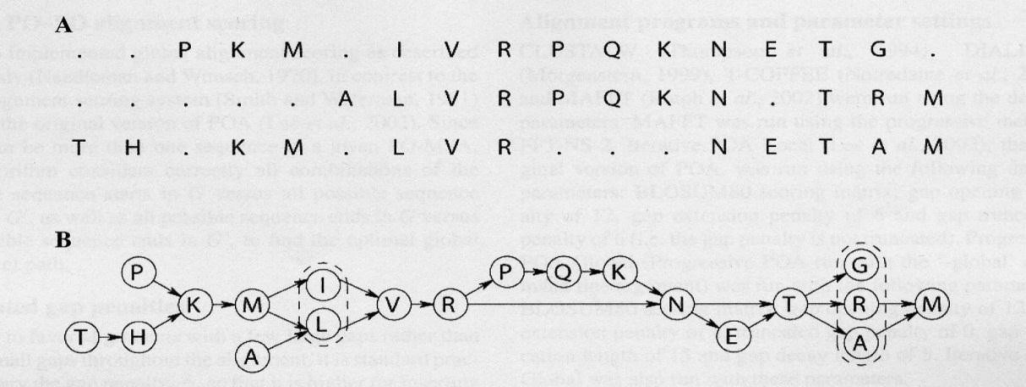


Figure 5: Représentation d'un alignement multiple sous forme linéaire (A) et sous forme de graphe (B).

4.4.Intérêt de l'alignement multiple de séquence

L'alignement multiple de séquences est un outil fondamental pour de nombreuses analyses en biologie. Il permet de comparer un groupe de protéines ou de gènes apparentés, afin d'établir des relations évolutives. Si deux séquences ont une similarité significative, il est fait l'hypothèse qu'elles partagent un ancêtre commun, elles sont donc homologues .Si deux séquences ont des motifs communs, il est fait l'hypothèse qu'elles sont soumises à une pression de sélection qui empêchent les mutations de se fixer, probablement parce que le motif est important pour assurer une fonction.

L'alignement multiple est principalement utilisé pour :

- Trouver des caractéristiques communes à une famille de protéines soit des régions conservées (des motifs), soit des acides aminés strictement conservés permettant de relier une séquence à une structure et à une fonction ;
- Construire l'arbre phylogénétique des séquences homologues considérées;
- Déduire des contraintes de structures pour les ARN.

Pour caractériser les régions conservées dans les séquences, il est souvent plus efficace d'utiliser plusieurs séquences et d'effectuer un alignement multiple. Récemment sont apparues des méthodes basées sur des stratégies itératives de raffinement d'un alignement initial, en utilisant soit des alignements locaux par programmation dynamique (Morgenstern et al. 1996), soit des alignements globaux par utilisation de chaînes de Markov cachés (Morgenstern et al. 1996) ou des algorithmes génétiques. Les algorithmes itératifs sont capables d'une plus grande

précision, mais ils sont plus gourmands en temps de calcul. La nature heuristique de ces programmes recommande la prudence dans l'interprétation des résultats et de préférence leur validation par l'utilisation de plusieurs programmes.

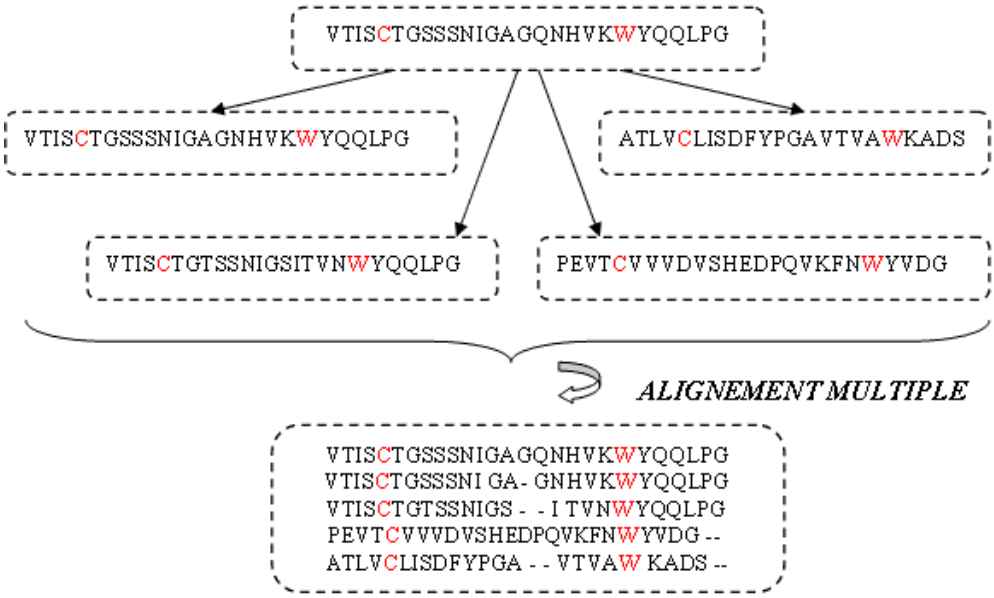


Figure 6: L'Alignement multiple de séquences protéiques.

4.5.L'Alignement Progressif

La méthode la plus utilisée pour aligner plusieurs protéines est l'*alignement progressif*. Elle se décompose en plusieurs étapes:

1. Calcul d'une matrice de distance entre toutes les paires de séquences.
2. Construction d'un arbre-guide à partir de cette matrice de distances
3. Construction de l'alignement sur base de l'arbre-guide

1. Matrice de distances entre paires de séquences

La première étape d'un alignement progressif consiste à aligner chaque paire de séquences, et à calculer leur distance. On regroupe les résultats dans une matrice de distances, où :

- chaque ligne correspond à une séquence
- chaque colonne correspond à une séquence
- la valeur $d_{i,j}$ indique la distance entre la séquence i et la séquence j.

Tableau I : Distance entre les séquences à alignées

	seq 1	seq 2	...	seq n
seq 1	d1,1	d1,2	...	d1,n
seq 2	d2,1	d2,2	...	d2,n
...	...	...	...	...
seq n	dn,1	dn,2	...	dn,n

Les alignements par paires peuvent être effectués en utilisant la programmation dynamique (algorithme de Needleman-Wunsch) ou une heuristique plus rapide (fasta, blast).

2. Construction de l'arbre-guide

A partir de la matrice de distance, on peut construire un arbre-guide par la méthode du Neighbourjoining (*NJ*).

Le principe est d'établir en premier lieu un branchement qui relie les deux séquences les plus proches (celles qui ont la distance minimale dans la matrice de distances), puis les séquences un peu moins proches, et ainsi de suite jusqu'à avoir branché toutes les séquences.

Il s'agit uniquement d'un outil utilisé temporairement pour déterminer l'ordre d'incorporation des séquences dans l'alignement entre séquences multiples.

L'inférence phylogénétique nécessite des analyses plus poussées, qui ne pourront être effectuées qu'après avoir obtenu l'alignement multiple.

3. Construction de l'alignement progressif

Après avoir calculé la matrice de distance et construit l'arbre-guide, on construit l'alignement multiple en incorporant progressivement les séquences selon leur ordre de branchement dans l'arbre guide, en remontant des plus proches aux plus éloignées.

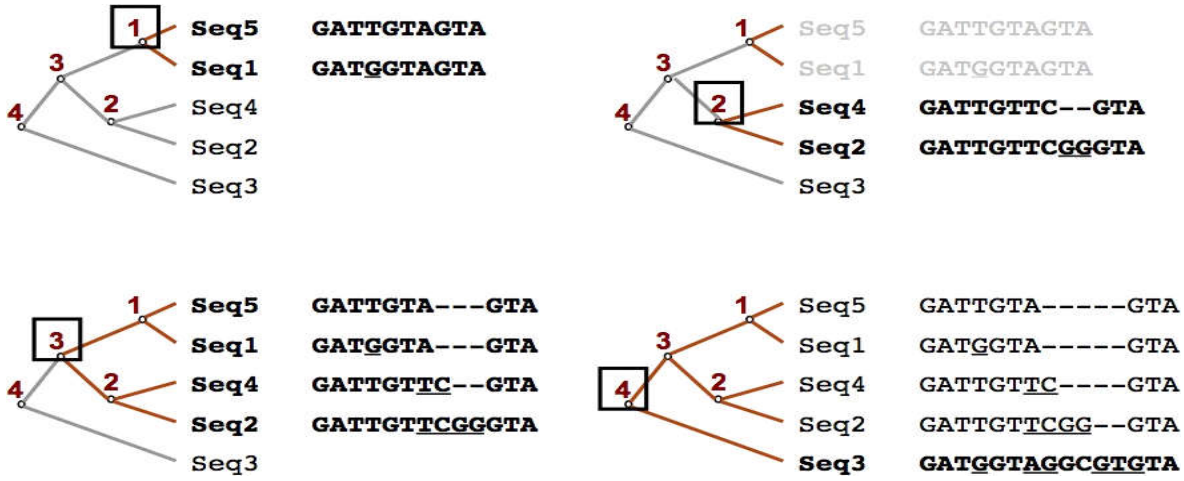


Figure 7 : Construction d'alignement progressif

5. Score d'un alignement multiple

Le score d'un alignement multiple doit rendre compte de la qualité de l'alignement. Les algorithmes utilisés cherchent à maximiser ce score, qui est une indication de l'alignement optimal.

Quelle que soit la méthode d'alignement multiple, le problème de la méthode de calcul du score se pose. La plus utilisée est le score somme des paires (SP)" sum of pairs ": somme sur chaque colonne de tous les scores entre acides aminés pris deux à deux (selon une matrice de substitution).En faisant la moyenne par paires ou la somme sur l'ensemble des colonnes, on obtient un score pour l'alignement. En outre, chaque algorithme implémente son propre calcul de score selon plusieurs critères, notamment:

- les modalités de prise en compte des pénalités de gap: ouverture, extension, fermeture;
- la prise en compte de la région concernée: dé favoriser les gaps dans les régions hydrophobes et les favoriser dans les régions hydrophiles. [9]

6. Matrices de score pour les protéines

6.1. Les matrices de substitutions

Deux grandes familles de matrices (log odds matrix)

- **Matrices PAM**

Les matrices PAM pour «Percent Accepted Mutation/Accepted Point Mutation », sont construites par étude de segments pris dans des séquences protéiques homologues (moins de 15% de différences).

PAM*x* : x% de mutations acceptées entre les séquences qui ont servi à construire la matrice.

Les fréquences de substitutions observées (ou probabilité conditionnelle: appelée "odd") sont transformées en logarithme de probabilité, normalisé en unité d'évolution. Le logarithme est utilisé pour que dans les programmes de recherche de ressemblance, la somme de ces éléments donne le logarithme de la probabilité pour la séquence entière (le modèle étant Markovien: indépendance de fréquences de substitution).

Les éléments diagonaux de la matrice indiquent une évolution sans substitution.

Pour PAM1, leur somme est telle qu'elle correspond à une probabilité de 99/100 (1 mutation pour 100 résidus: d'où le nom PAM : accepted point mutation)

L'indépendance des fréquences et les éléments de la matrice étant des logarithmes de fréquence,

On peut calculer PAM(N) en élevant PAM 1 à la puissance N, par exemple pour PAM 120, il faut multiplier PAM 1 par elle-même 120 fois.

➢ Matrice BLOSUM

Ces matrices BLOSUM (Blocks Substitutions Matrices) sont construites par analyse de séquences de protéines. Les séquences sont découpées en blocs (2000 résidus au total) par rapport au pourcentage d'acides aminés inchangés.

BLOSUM*x* : matrice obtenue à partir de séquences présentant au minimum x % d'identité (similitude) entre elles.

Une matrice "d'odds" est calculée à partir des blocs d'alignement pour chaque valeur de similitude, et ensuite chaque élément est transformé en unité d'information en prenant le logarithme du rapport de la valeur observée à la valeur qu'on obtiendrait au hasard. Cette matrice est ensuite normalisée. Les correspondances entre BLOSUM et PAM, basées sur la théorie de l'information sont:

- PAM250--->BLOSUM45
- PAM160--->BLOSUM62
- PAM120--->BLOSUM8

Les matrices : sont calculées selon le principe du log-odd-ratio.

$$\textbf{odd-ratio} = \frac{\Pr(\text{observé})}{\Pr(\text{attendu})}$$

$$\text{odd-ratio (a,b)} = \frac{Pa,b}{Pa \times Pb} \text{ pour deux lettres a, b avec :}$$

Pa,b : probabilité d'observer a aligné à b sur deux séquences homologues.

Pa × Pb : probabilité attendu d'aligner a à b sur deux séquences non-homologues.

log-odd-ratio = log (odd-ratio)

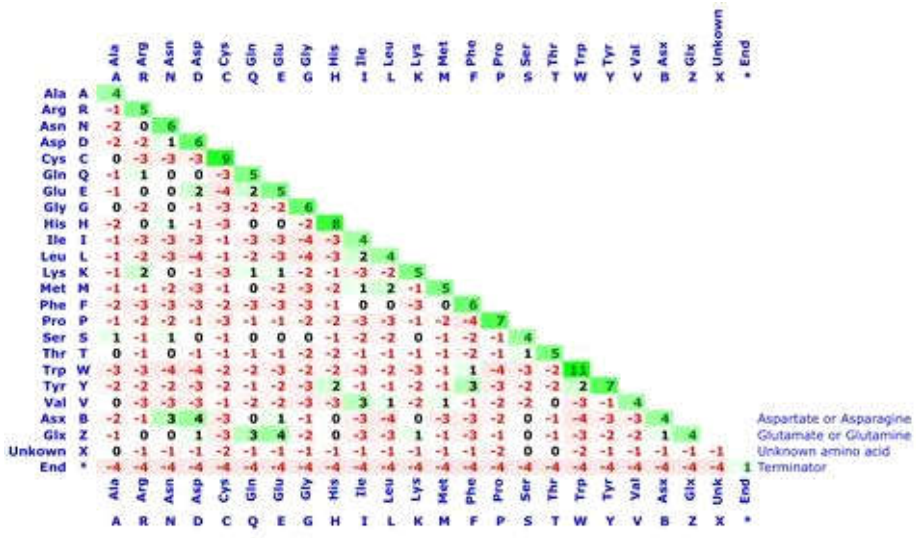


Figure 8: Représentation d'un exemple de la matrice BLOSUM

6.2.Choix de la Matrice Protéique

Le choix d'une matrice dé pend du type d'analyse que l'on veut faire. Il n'y a pas une matrice idéale et un grand nombre d'études comparatives sur les matrice sont mis en évidence (de manière schématique) que :

- Pour des séquences similaires et courtes, il est préférable d'utiliser une matrice BLOSUM élevée ou PAM faible.
- Pour des séquences divergentes et longues, il est préférable d'utiliser une matrice BLOSUM faible ou PAM élevée.
- La matrice BLOSUM 62 semble être la matrice la plus utilisée pour la comparaison avec les banques de données, et pour un grand nombre de logiciels d'alignement de séquence, elle semble être la matrice par défaut.

6.3. Comparaison des matrices PAM et BLOSUM

Plusieurs études comparatives ont été réalisées entre ces matrices performantes, largement utilisés pour réaliser des recherches dans les banques ou des alignements. La figure3.3 établit la correspondance entre ces deux types de matrice. La plupart des études ont montré que les matrices BLOSUM donnaient de meilleurs résultats que les matrices PAM. Deux raisons expliquent ce résultat:

- ces matrices sont réalisées à partir d'alignements locaux (régions structurellement conservées) alors que les matrices PAM incluent des régions très divergentes.
- les matrices BLOSUM sont plus et donc ont bénéficié d'un plus grand nombre de données à disposition.

Les algorithmes de recherche dans les banques FASTA [LipmanandPearson, 1985] utilise par défaut la matrice BLOSUM 62.

Évaluation d'un Alignement

Évaluer un alignement revient alors à mesurer sa qualité en déterminant la distance qui sépare les deux séquences. Le score d'un alignement est la somme des scores de toutes les positions de bases (résidus) prises deux à deux [13].

Exemple d'évaluation:

On peut attribuer une valeur positive à des symboles alignés identiques et une pénalité (valeur négative) à une substitution ou à un gap.

Si l'on considère l'exemple
précédent: Score (identité)=2
Score(substitution)=-1
Score (gap)= -2

Le score de cet alignement serait alors:

SEQ1: G A R F I EV H EL - - TF AT T C A T

SEQ2: G A R F I EL T HEVAS Y F - - C A T

score total = 2+2+2+2+2+2-1-1-1-1-2-2-1-1-1-2 -2+2+2+2 =**+3**

Nous présenterons tout d'abord les méthodes permettant de mesurer, de donner un score à ces mutations. Ce score nous permettra alors de déterminer le score d'un alignement, défini comme la somme des scores des évènements mutationnels (illustrés sur la figure II.1) étant survenus entre deux séquences. Ensuite, nous décrirons les alignements de deux séquences et enfin l'alignement multiple de séquences.

7. La fonction profil

Cette fonction a pour objectif de calculer le profil d'un alignement A.

Ce profil est une représentation numérique d'un MSA qui représente les caractéristiques communes d'une famille de protéines. La fonction Profil est utilisée pour déterminer le degré d'appartenance d'une protéine à une famille. On peut signaler qu'il est utile dans l'alignement des séquences pas trop divergentes .Il permet de déterminer des régions conservées dans une séquence ou plusieurs. C'est la somme des fréquences d'apparition de chaque résidu dans chaque colonne de l'alignement.