2. Alignement pair

Si une nouvelle séquence est obtenue à partir du séquençage génomique, la première étape est la recherche de similarités avec des séquences connues dans d'autres organismes.

Si la fonction/structure des séquences similaires/protéines est connue, très probablement (highly likely) la nouvelle séquence correspond à une protéine avec la même fonction/structure. En effet, il a été trouvé que seulement à peu près 1% des gènes humains n'ont pas de contrepartie dans le génome de souris et que la moyenne de similarité entre les gènes de la souris et de l'homme est de 85%.

Les similarités existent parce que toutes les cellules possèdent une cellule ancêtre commune (a mother cell). Donc, dans les différents organismes il pourrait avoir des mutations d'acides aminés dans certaines protéines parce que les acides aminés ne sont pas tous importants pour la fonction et peuvent être remplacés par des acides aminés qui ont des caractéristiques chimiques semblables sans changer la structure. Parfois les mutations sont tellement nombreuses qu'il est difficile de trouver des similarités.

La méthode du calcul des fonctions des gènes par similarités est appelée la *génomique comparative* ou la *recherche d'homologie*. Deux séquences sont homologues lorsqu'ils ont comme racine un ancêtre commun.

2.1. Les similarités de séquences et score

Après le séquençage, les biologistes n'ont habituellement aucune idée de l'utilité des gènes trouvés. En espérant découvrir un indice sur leurs fonctions, ils tentent de trouver des similitudes entre des gènes nouvellement séquencés et d'autres déjà séquencés dont ils connaissent les fonctions.

Le jeu suivant, transformer un mot anglais en un autre mot en passant par une série de mots intermédiaires, dans laquelle chaque mot ne diffère du suivant que d'une seule lettre. Pour transformer *head* en *tail*, on n'a besoin que de quatre intermédiaires :
head → heal → teal → tell → tall → tail.

Pour les séquences biologiques, il est connu comment une séquence peut mutée en une autre. Premièrement, il y'a les *points de mutation* ou un nucléotide ou acide aminé est changé en un autre. Deuxièmement, il y'a les *suppressions* ou un élément (nucléotide ou acide aminé) ou une subséquence entière d'un élément est supprimée de la séquence.

Troisièmement, il y'a les *insertions* ou un élément ou une subséquence est insérée dans la séquence. Un alignement peut s'interpréter comme le fruit d'un travail d'édition : trouver le nombre minimum d'opérations élémentaires d'édition qui permettent de transformer une

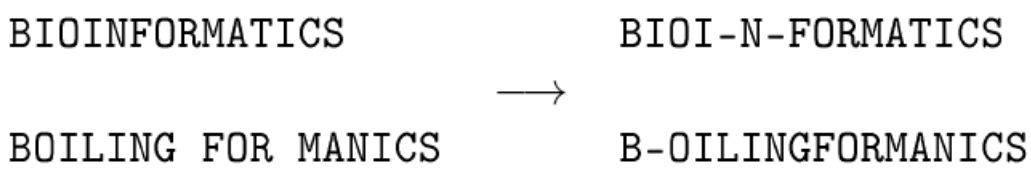
séquence en une autre. On considère les trois opérations suivantes :

(a) insertion : insertion d'une ou plusieurs lettres ;

(b) délétion : suppression d'une ou plusieurs lettres ;

(c) substitution : remplacement d'une lettre par une autre.

Dans une perspective évolutive ces trois opérations peuvent s'interpréter comme des mutations et le travail d'édition comme une tentative de reconstruction de l'histoire évolutive en considérant ces 3 mutations élémentaires. L'alignement suivant par exemple.



Le conte donne 12 lettres identiques sorties des 14 lettres de BIOINFORMATICS. Les mutations pourraient êtres :

(1) suppression I BOINFORMATICS

(2) insertion LI BOILINFORMATICS

(3) insertion G BOILINGFORMATICS

(4) changement de T en N BOILINGFORMANICS

Les deux textes semblent très similaires. Noter que l'insertion ou la suppression ne peuvent pas être distinguées si les deux séquences sont présentées (es que le I est supprimé de la première séquence ou inséré dans la seconde ?). Donc, les deux cas sont dénotées par "-".

La tache des algorithmes bioinformatiques est de trouver à partir de deux séries (la partie à gauche dans l'exemple au-dessus) l'alignement optimal (la partie à droite dans l'exemple au-dessus). L'alignement optimal est l'arrangement des deux séries d'une manière ou le nombre de mutations est minimal.

L'alignement peut être global (sur toute la longueur de la séquence) ou local (sur les parties les mieux conservées), selon la relation présumée entre les séquences. On définit un score d'alignement qui permet de définir le meilleur alignement de deux séquences et de quantifier leur ressemblance.

2.2. La matrice d'identité

La matrice d'identité ou matrice de dot (Dot Matrix) est un outil de représentation des alignements, ou une séquence est écrite horizontalement en haut et l'autre verticalement à gauche. Ce qui donne une matrice ou chaque lettre de la première séquence est couplée avec

chaque lettre de la deuxième séquence. Pour chaque correspondance de lettres un point (dot) est inscrit dans la position concordante dans la matrice. Quelles paires apparaissent dans l'alignement optimal ? On va voir ci-après que chaque chemin à travers la matrice correspond à un alignement

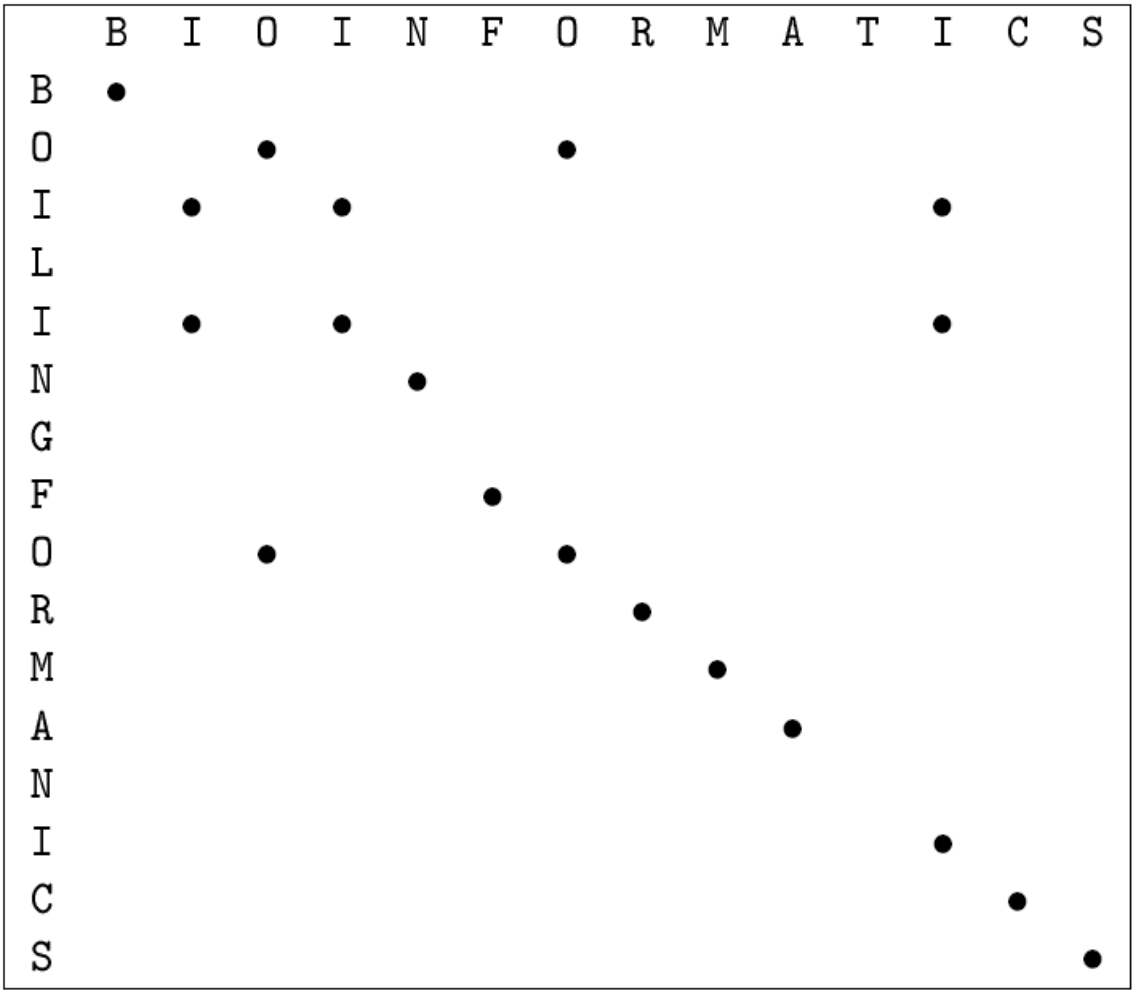


Figure 2a. Principe opérationnel de la matrice d'identité.

Règles : vous pouvez bouger horizontalement "→", verticalement "↓", et vous pouvez bouger seulement diagonalement " " si vous êtes dans la position de dot.

Tache : faite le plus possible de mouvements diagonaux quand vous bougez du coint le plus haut à gauche au coint le plus bas à droite.

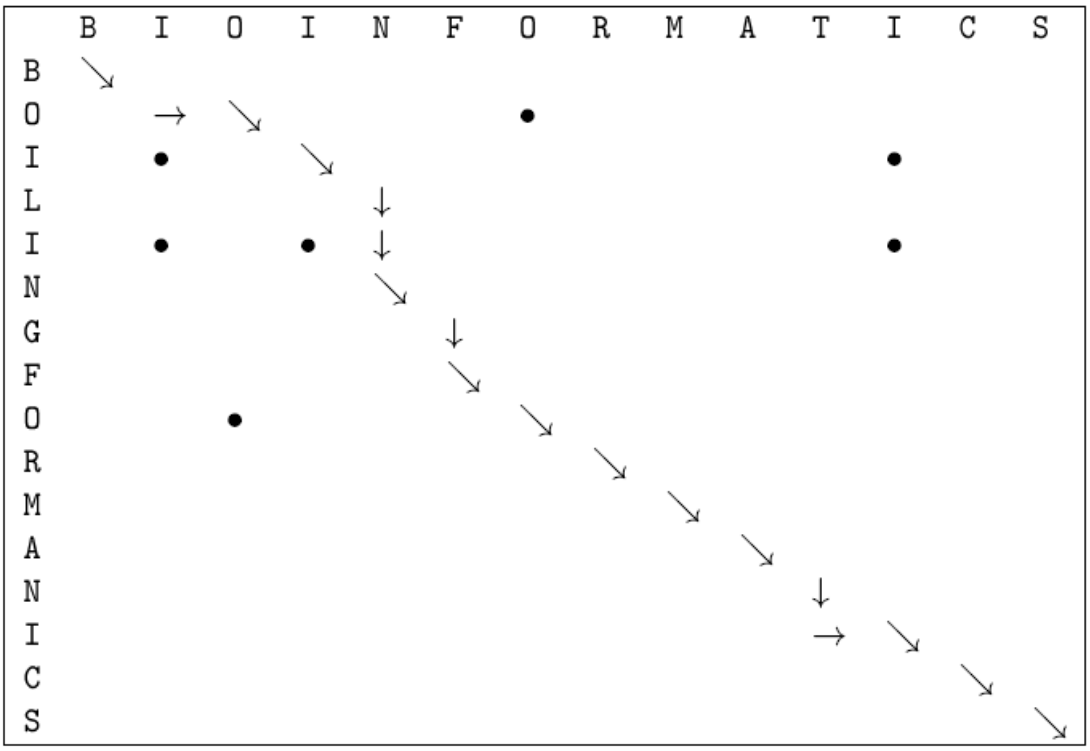


Figure 2b. Principe opérationnel de la matrice d'identité.

Le nombre de mouvements diagonaux“ ” représente les correspondances et le nombre de scores, “→” correspond à “-” dans la séquence verticale, “↓” à “-” dans la séquence horizontale et la combinaison “→↓” ou “↓→” correspond à une divergence. Donc, chaque chemin à travers la matrice correspond à un alignement et chaque alignement peut être exprimé par un chemin dans la matrice.

Dans la Figure 3 les dot sur les diagonales correspondent aux régions de correspondances (similarités). Elle représente des Matrices Dot pour la comparaison de la protéine triosephosphate isomérase (TIM) humaine avec celle de la levure, *E. coli* et *Archaeon.* Pour la levure la diagonale est complète et pour *E. coli* de petits trous « gaps » sont visibles, mais *Archaeon* ne montre pas une diagonale étendue. Donc, la TIM humaine correspond le plus avec la TIM de la levure, suivie par la TIM d'*E. coli* et possède la similarité la plus faible avec la TIM d'*Archaeon*.

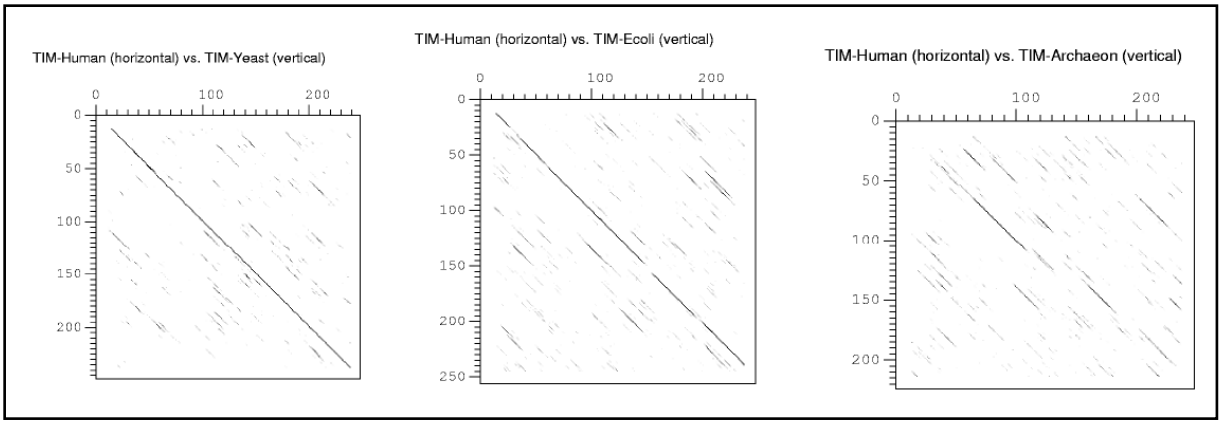


Figure 3. Matrice de dot de la triosephosphate isomérase humaine avec la même protéine dans la levure, *E. coli* et *Archaeon.* La levure donne la meilleure correspondance car la diagonale est presque complète. *E. coli* a quelques fractures dans la diagonale. *Archaeon* montre la similarité la plus faible. Cependant, la structure 3D et la fonction est la même pour toutes les protéines.

3. Alignement multiple

Le but de la comparaison des séquences protéiques est de découvrir des similitudes «biologiques » (i.e. structurelles ou fonctionnelles) parmi les protéines. Des protéines biologiquement similaires peuvent ne pas exhiber une forte similitude de séquences et l'on aimerait reconnaître la ressemblance structurelle / fonctionnelle, même lorsque les séquences sont très différentes.

La comparaison simultanée de nombreuses séquences permet souvent de trouver des similitudes invisibles dans la comparaison de séquences par paires « l'alignement par paires chuchote... l'alignement multiple crie ».

L'alignement multiple est la base de l'étude de familles de protéines et de domaines fonctionnels. Son but est de révéler des similarités de séquence ou de structure dans une famille de séquences voisines dans l'évolution ou par la fonction.

Il convient de bien analyser le résultat de l'alignement multiple avant de passer à la construction de l'arbre phylogénétique et de bien régler les paramètres du logiciel. Nous allons procéder à l'alignement multiple du jeu de séquences en utilisant l'outil ClustalW.

Ces séquences appartiennent à la famille des facteurs de transcription du type "Basic Leucine Zipper". Ce sont des gènes qui codent pour des protéines qui régulent la transcription des ARNm.

Le résultat d'une partie l'alignement multiple de cette série de séquences est le suivant :

Cet alignement présente beaucoup de gap qui faussent l'interprétation. Ceci est dû au fait que nos séquences appartiennent à des individus dont la taxonomie est totalement différente. Nous avons aligné des séquences de grenouille, de blé, etc.

Exemple d'alignement de séquences par BLAST/NCBI

La Figure 4 représente le résultat d'un alignement de la séquence partiel du gène ARNr16S d'*Aeromonas veronii* obtenue sur GenBank, via le programme BlastN.

>*Aeromonas veronii*

Figure 4. Analyse bioinformatique des séquences d'ADNr16s sur GenBank, via le programme BlastN.

GenBank, via le programme BlastN réalise un alignement en utilisant ses propres séquences et propose celle qui présente la meilleure identité avec la nôtre en calculant un score qui correspond au nombre de nucléotides identiques chez les deux séquences. Ce score peut être traduit sous forme de pourcentage d'identité (%id). La valeur calculée de E-value indique la probabilité que le résultat de cet alignement a eu lieu par hasard. Donc plus cette valeur est proche du zéro et mieux c'est. Or tous les alignements ont abouti à des valeurs nulles de la E-value ; ce qui exprime que les identités retrouvées entre nos séquences et celles proposées par GenBank ne sont pas dues au hasard.