République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Centre Universitaire Abdelhafid Boussouf – Mila

Institut des sciences et de la technologie

Département des Sciences de la nature et de la vie

Cours : Bioinformatique

3 ème Année Licence Microbiologie

Présenté par : Dr. Hicham BERRABAH

Année académique 2022-2023

Chapitre 1 : Rappel en biologie moléculaire

1.1. Définition

Les acides nucléiques sont des substances chimiques qui existent non seulement dans le noyau, mais aussi dans le cytoplasme des cellules. On en distingue deux types: l'ADN et l'ARN.

• L'A.D. N (acides désoxyribonucléique) est essentiellement localisé dans le noyau des cellules quand celui-ci est individualisé (chromosome) comme chez les eucaryotes et dans certains organites cellulaires tels que les chloroplastes et les mitochondries. Par contre, chez les procaryotes l'ADN baigne directement dans le cytoplasme.

• L'A.R. N (acides ribonucléiques), essentiellement retrouvé dans le cytoplasme des cellules. La plupart des êtres vivants possèdent simultanément les deux sortes d'acides nucléiques. Les virus font exception, ne renferment qu'un des deux acides nucléiques : l'ADN (virus de l'hépatite B) ou l'ARN (l'hépatite A, SIDA).

Les acides nucléiques doivent leur nom au fait qu'ils sont abondants dans les noyaux cellulaires (ils ont été d'abord isolés du noyau des cellules).

1.2 Composition :

Les acides nucléiques sont de très longues molécules, formées par la répétition de sous unités appelées « nucléotides ». Un nucléotide est lui-même constitué de trois éléments: un acide phosphorique, un ose (sucre) et une base.

Nucléotide = Acide phosphorique + ose + base

1.2.1 La base

Dans les nucléotides, il existe deux types possibles de base :

• Les bases dites pyrimidiques : possèdent un cycle pyrimidine et sont représentées par la Thymine (T), la Cytosine (C) (il y' a un ''y'' comme pyrimidine) et l'Uracile (U).

• Les bases dites puriques : possèdent toutes un noyau purine et sont représentées par l'Adénine (A) et la Guanine (G).

1.2.2 L'ose

On trouve deux types d'oses dans les acides nucléiques.

• Ribose : est un ribose en C5. Ce nom provient des initiales de l'institut ou il a été découvert ''Rockefeller Institute of Biochemistry'' à New York.

• Désoxyribose (2' - désoxy - D ribose) est un ribose dans lequel manque un O en 2'.

1.2.3 L'acide phosphorique

L'acide phosphorique est un triacide. Deux des trois fonctions acides seront estérifiées dans les ADN et ARN.

1.3. Lecture d'un acide nucléique

Une chaîne nucléique présente 2 extrémités:

• Une contenant le groupement phosphaté avec 2 fonctions acides libres, on l'appelle extrémité 5' P

• L'autre contenant un OH libre en 3' sur l'ose, on l'appelle extrémité 3' OH.

On lira toujours une chaîne d'acides nucléiques dans le sens 5' P vers 3' OH.

Par souci de simplification, on a maintenant pris l'habitude de faire figurer sur chaque séquence d'A.D.N les seuls chiffres 5' et 3'.

1.4 Caractéristiques de l'A.D.N

Trois caractéristiques sont propres à l 'A.D.N et vont le différencier des A.R.N.

1.4.1 L'ose

Comme les initiales "A.D.N" l'indiquent, l'ose entrant dans la constitution de l' A.D.N est du désoxyribose (et non pas le ribose comme ce sera le cas dans les A.R.N).

1.4.2 Les bases

Constituants les nucléotides de l' A.D.N sont: A G C T. On trouve :

• 2 bases puriques: adénine (A); guanine (G).

• 2 bases pyrimidiques: cytosine (C); Thymine (T).

Remarque : Il est à noter que dans l' A.D.N on ne trouve jamais d'Uracile (U) alors que dans les ARN, il y' aura de l'uracile à la place de la Thymine.

1.4.3 Les deux chaînes de nucléotides

Une molécule d' A.D.N est habituellement formée de 2 chaînes (on dit aussi 2 brins) de nucléotides (ou polynucléotides) alors qu'une molécule d' A.R.N n'en comprends qu'une (on note des exceptions chez certains virus).

Ces 2 chaînes ont 3 propriétés:

• **Antiparallèles:**

Signifie que les deux brins de nucléotides sont parallèles mais dans des directions opposées.

Pour un brin, la direction 5' vers 3' se trouve être, par exemple de haut en bas et pour le 2ème brin, la direction 5' vers 3' sera à l'inverse de bas en haut.

• **Complémentaires**

La règle de complémentarité est la suivante: en face de A on a T et en face de C on a G. En effet, la distribution des bases azotées de l'ADN n'est pas quelconque (observation faite par CHARGAFF en 1950), le rapport A+G/T+C est toujours égal à 1 (aux erreurs de mesures près). Il y' a donc autant de A que de T, autant de G que de C. Selon le modèle de Watson et CRICK : une purine (A,G) se lie avec une pyrimidine (C,T). Les bases complémentaires situées face à face sont liées entre elles par

des liaisons hydrogène. Le nombre de liaisons hydrogène pour le couple A - T est de deux et pour le couple C - G de trois.

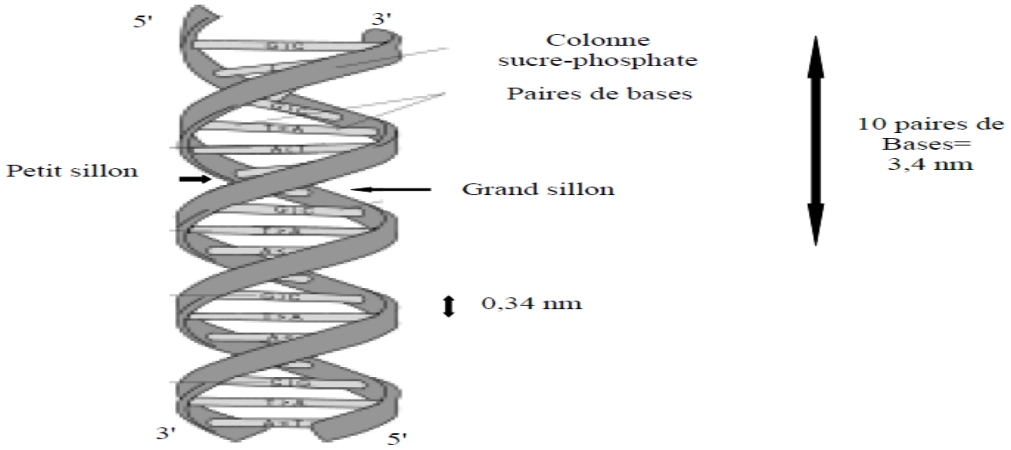
• **Hélicoïdales**

Cette distribution égale de A et T d'une part, de G et C d'autre part implique donc une structure particulière de l'ADN, structure qui a beaucoup intrigué les biologistes.

FRANKLIN et WILKINS, en étudiant l'ADN par diffraction aux rayons X, établissent que la molécule, présente une structure hélicoïdale. En 1953 WATSON et CRICK propose le modèle moléculaire de la structure en double hélice droite de l'ADN. Le modèle trouvé par WATSON et CRICK est conforme aux données de FRANKLIN et WILKINS et il respecte les règles de CHARGAFF.

Ainsi, la structure globale de la molécule d' A.D.N est celle d'une double hélice droite. Les 2 chaînes polynucléotidiques présentent dans l'espace une configuration hélicoïdale. Elles s'enroulent autour d'un même axe. Dans chacune des deux chaînes ‘’l'arête ou la corde vertébrale’’ est constituée par une alternance de molécules de sucres et d'acides phosphorique, alors que les bases azotées sont orientées latéralement et à l'intérieur des 2 hélices (elles sont empilées les unes au-dessus des autres). La distance entre chaque base est de 3.4 Å. il y' a 10 paires de bases par tour d'hélice; son pas est donc de 34 Å (3.4 nanomètre) (c'est à dire que la double hélice effectue un tour toutes les 10 paires de base environ).

Le diamètre de la double hélice est de 20 Å. La double hélice est une molécule relativement rigide (due aux contraintes structurales). Les arêtes créent par l'enchaînement des groupements phosphatés définissent 2 sillons: le petit et le grand sillon. Les bases ne sont accessibles aux protéines qu'au niveau du grand sillon.



Représentation de la double hélice d'ADN

1.5 L'A.D.N des différents êtres vivants

Il est vraiment très remarquable que l' A.D.N de tous les êtres vivants, qu'ils s' agissent d'un animal, d'une plante, d'une bactérie ou d'un virus, possède le même type de structure, soit 2 brins (sauf

exception rencontrée parfois chez certains type de virus) constitués chacun par une succession de plusieurs milliers de nucléotides.

Ce qui diffère d'une espèce à une autre, se sera:

• Le nombre de molécules d' A.D.N dans un virus ou une cellule animale ou végétale : une molécule d' A.D.N chez les virus ou chez Eschérichia coli, plusieurs dans une cellule animale et végétale.

• Leur longueur: quelques milliers de nucléotides, ou plusieurs milliards (répartis sur plusieurs chromosomes).

• Leur forme: linéaire ou circulaire.

• Leur localisation dans la cellule: A.D.N séparé ou non du cytoplasme par une membrane nucléaire.

• Mais, c'est essentiellement la séquence de base qui sera caractéristique de chaque molécule d' A.D.N. Ce sont ces séquences qui vont jouer un rôle biologique capital.

Des séquences de base différentes donneront des messages différents.

1.5.1 Les virus

Les virus sont des êtres vivants possédant les acides nucléiques les plus courts. L'A.D.N des virus est formé de quelques milliers à plusieurs dizaines de milliers de nucléotides. Leurs masses molaires varient de 10^6 à 10^7.

1.5.2 Les procaryotes (exemple: Eschérichia.coli)

L'A.D.N. n'est pas situé dans un noyau mais se trouve dans le cytoplasme où il constitue l'unique chromosome. Il a une forme circulaire (par rapport à la continuité de la chaîne d'A.D.N. et non à sa forme géométrique). L'A.D.N des procaryotes est plus long, approximativement mille fois plus que celui des virus. Ainsi l' A.D.N de Eschérichia coli comporte quatre millions de paires de nucléotides. La masse molaire de l'ADN des bactéries est de 10^9.

1.5.3 Les eucaryotes

L'A.D.N est situé dans le noyau. Chaque chromosome contient une très longue molécule d'A.D.N., toute repliée, pelotonnée. Le nombre total de nucléotides dans une cellule humaine est très grand, approximativement mille fois plus grand que dans le cas des bactéries. On trouve environ trois milliards de paires de nucléotides dans les molécules d 'A.D.N constituant les 46 chromosomes humains. La masse molaire de l'ADN humain est de 10 13.

Remarque :

On trouve aussi, chez les eucaryotes, de l' A.D.N en dehors du noyau. Ainsi, les mitochondries et les chloroplastes contiennent de l' A.D.N.

L'A.D.N des mitochondries humain est composé de 2 brins, il est circulaire et clos. Il code pour des ARNr, ARNt et ARNm mitochondriaux. Un des deux brins contient d'avantage de gènes.

Le code génétique mitochondrial est légèrement différent du code nucléaire.

1.6 Structure et caractéristiques des A.R.N

Les A.R.N (acides ribonucléiques) sont caractérisés essentiellement par:

• L'ose: comme le nom l'indique A.R.N, l'ose est le ribose (à la différence de l' A.DN ou le sucre est du désoxyribose).

• Les bases: les bases rencontrées dans les A.R.N sont A, C, G et U à la place de T.

• Une seule chaîne de nucléotides (et non pas deux comme l' A.D.N). Cette chaîne est d'ailleurs plus courte que les chaînes d' A.D.N.

Remarque: L'appariement entre bases complémentaires pourra s'observer soit entre 2 molécules d' A.R.N différentes, soit une même molécule d' A.R.N, dans une région repliée en épingle à cheveux. Les règles d'appariement entre deux brins d' A.R.N seront les mêmes qu'entre deux brins d' A.D.N concernant C et G par contre U remplace T dans l'appariement avec A.

Les cellules contiennent essentiellement 4 types d' A.R.N :

• **ARNr** (ribosomique) : Les ARNr entrent dans la composition du ribosome (nécessaire à la synthèse des protéines). Un ribosome fonctionnel est lui-même formé de deux sous unités, chacune est constituée d'un mélange de protéines (r- protéines) et d'ARN (ARNr). Les ribosomes sont situés dans le cytoplasme et sont nécessaires à la synthèse des protéines. Ce sont de véritables "usines à protéines".

• **ARNt** (de transfert) : ils sont appelés ainsi car ils vont transférer, véhiculer les acides aminés qui se trouvent dans le cytoplasme jusqu'au ribosome, lieu de synthèse protéique. Un ARNt possède la structure générale des ARN. La chaîne d' ARNt se replie pour donner un aspect général en forme de trèfle. Deux sites sont importants dans un ARNt :

- L'extrémité 3'OH ou sera fixé l'acide aminé à transporter

- L'anticodon (triplet) situé sur une boucle de l'ARNt qui va jouer un rôle très important car il reconnaîtra le codon de l' ARNm. Cet appariement anticodon- codon se fait de manière antiparallèle et complémentaire entre les bases du codon et de l'anticodon.

• **ARNm** (messager) il est formé d'une seule chaîne de nucléotides comprenant les mêmes sortes de bases AUCG. On l'appelle messager car il porte l'information génétique contenue au niveau de l'ADN jusqu'au ribosome ou s'effectuera la synthèse protéique. La taille de la molécule d' ARNm dépend de la longueur de la ou les chaînes polypeptidiques pour laquelle il code. Les ARNm se renouvellent très vite, ils sont rapidement produits et rapidement dégradés. Ils ne durent que le temps d'un message. Un ARNm pourra cependant être lu plusieurs fois au niveau du ribosome.

• **SnRNA** (small nuclear)les plus petites molécules d'ARN, nous allons voir plus tard que ces ARN jouent un rôle important dans la maturation des Pré ARNm.

Chapitre 2 : Introduction à la bio-informatique et bases des données

1. Définition :

La bioinformatique est la discipline de l'analyse « *in silico* » de l'information biologique renfermée dans les séquences nucléotidiques (séquences de nucléotides) et protéiques (séquence des acides aminés).

- Son apparition, dans les années 1980. Coïncide avec la création des premières banques de données (EMBL et GenBank). À partir des années 1990, la bioinformatique devient indispensable avec l'accumulation des données de séquençage notamment les génomes complets. Fondée sur les acquis de la biologie, elle permet de produire de nouvelles connaissances et des suggestions pour de nouvelles expériences.

2. La bioinformatique propose des méthodes et des logiciels qui permettent :

- La collection, le stockage et la gestion des données biologiques et leur distribution à travers les réseaux.
- Le développement des outils (logiciels/algorithmes) pour analyser les problèmes de biologie moléculaire.
- L'analyse, la comparaison et la prédiction de la structure des gènes.
- La modélisation et la prédiction de la structure et de la fonction des protéines.
- Les études phylogénétiques et l'évolution moléculaire des êtres vivants.

3. Les banques des données biologiques

- Les bases de données contenant des informations biologiques et des données largement diffusées par le réseau Internet. Elles sont généralement reliées entre elles par des liens « links ».

Il existe de nombreuses bases de données bibliographiques. Certaines sont consultables gratuitement ou après abonnement payant sur interne. Les bases de données bibliographiques sont des outils indispensables à tout travail de recherche et à toutes les étapes du parcours étudiant et professionnel. Les bases de données (ou banques de données, ou databases) sont des produits documentaires qui rassemblent:

▪ soit des documents immédiatement utilisables (articles, photos, chiffres), dans ce cas on parle d'information primaire,

▪ soit des informations sur ces articles, photos (auteur, titre, résumé etc.), l'information est alors qualifiée d'information secondaire.

- Il y a deux types de banques :

3.1 - Celles qui correspondent à une collecte des données l plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations. Traitent des thématiques générales, sont des banques de données ou bases de données généralistes.

On appelle banques généralistes, ou banques primaires, les ressources qui collectent, gèrent, archivent et mettent à disposition de la communauté scientifique un ensemble de données primaires.
Classiquement, on considère comme banques primaires les banques généralistes qui contiennent des séquences nucléiques et protéique obtenus par des méthodes expérimentales.
Bien qu'actuellement la plupart des séquences protéiques ne soient pas obtenues expérimentalement, mais à partir des données de séquence nucléiques. Ainsi que les banques qui gèrent les structures tridimensionnelles des protéines.

3.1.1. Les banques nucléiques

Les données stockées dans ce type de banques sont des données issues de séquençage d'ADN et ARN. Trois banques nucléiques connues, elles partagent des informations et donc contiennent des ensembles presque identiques de séquences. Ces trois banques s'échangent systématiquement leur contenu depuis 1987 et ont adopté un système de conventions communes : « DDBJ/EMBL/GenBank »:

• ***La banque EMBL***: créée en 1980 et financée par l'EMBO (European Moleculary Biology Organization), développée au sein du Laboratoire Européen de Biologie Moléculaire situé à Heidelberg (Allemagne), elle est maintenant diffusée par l'EBI:
http://www.ebi.ac.uk/embl/. En 24 février 2014, la banque contient 369.5 millions séquences.

• ***La banque GenBank (Genetic Sequence Databank):*** créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI (National Center for Biotechnology Information) : http://www.ncbi.nlm.nih.gov/. En février 2014 la banque contient 171.123.749 séquences. GenBank contient une sous-banque de protéines, traduction des séquences nucléiques, appelée GenPept.

• ***La banque DDBJ (DNA Databank of Japan):*** créée en 1986 et diffusée par le NIG (National Institute of Genetics, Japon), a enregistré un total de 81.994.905 de séquences ADN le mois de décembre 2019 (DDBJ 2019).

3.1.2. Les banques protéiques

Les données stockées dans ces bases sont issues d'une traduction de séquences d'ADN ou par le séquençage de protéines (rare car long et coûteux):

• ***La banque SwissProt :*** est une banque protéique crée en 1986 à l'Université de Genève et maintenue depuis 1987 dans le cadre d'une collaboration, entre cette université (via ExPASy, Expert Protein Analysis System) et l'EBI. Celle-ci regroupe aussi des séquences annotées de la banque PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL. En février 2014 la banque contient 542503 séquences compressant 192888369 acides aminés.

• ***La banque PDB (Protein Data Bank)*** créée en 1971, c'est la banque de référence des structures protéiques obtenues expérimentalement par cristallographie rayon X, spectroscopie RMN et cryo-microscopie électronique (technique la plus récemment utilisée). Les coordonnées des atomes

formant la structure d'une protéine, le détail de la séquence, les conditions de cristallisation sont les principales informations disponibles pour chaque structure de la banque. C'est à partir de cette banque que sont détectés les homologues structuraux. La Figure 1 représente l'évolution du nombre de structures protéiques enregistrées par année sur PDB, le mois de janvier 2020 a remarqué un total de 147.827 structures.

3.2 - Celles qui correspondent à des données plus homogènes et spécifiques. Traitent des thématiques Particulières, sont des banques de données ou bases de données spécialisées.

Les bases de données spécialisées sont d'intérêt divers et la masse des données qu'elles contiennent peut varier d'une base à une autre. Ces bases correspondent à des améliorations ou à des regroupements par rapport aux données issues des bases généralistes.

Donc on peut les appeler banques secondaires, ces bases contiennent des données homogènes et collecte des données établie autour d'une thématique particulière.

Exemple : bases spécialisée pour un génome spécifique, bases de séquences immunologiques, de voies métaboliques, de cartes génétiques, de motifs protéiques, d'expression de gènes, de structures...

3.2.1. Quelques exemples :

- **(LEAPdb) :** ***Late Embryogenesis Abundant Proteins database*** (G. Hunault & E. Jaspard) : cette base de données contient un grand nombre d'informations sur les protéines LEA impliqués dans la tolérance à de nombreux stress, notamment la déshydratation et le froid. Pour l'instant, on les a mises en évidence principalement chez les plantes.
- **RESID Database :** Base de données sur les acides aminés peu fréquents (sous-partie de la base de données PIR.
- **Map Viewer :** Map Viewer vous permet de visualiser et de rechercher génome complet d'un organisme, affichage des cartes chromosomiques. Le nombre et les types de cartes disponibles varient selon l'organisme
- **La base de données omim (*online mendelian inheritance in man*) :** Donne de nombreuses informations sur la classification des maladies génétiques, des présentations cliniques et la cartographie génomique de la localisation de la maladie. La base de données est mise à jour continuellement et offre probablement le meilleur lors de la recherche d'information sur les maladies héréditaires.

Chapitre 3 : Alignement des séquences

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%.

Alignement est un processus de comparaison de séquences permettant d'obtenir le maximum de correspondances entre les lettres qui les composent. Il est quantifié par un score de similarité.

Similarité: mesure du degré de ressemblance entre séquences, quantifié par un score, calculé à l'aide d'une matrice de score.

Homologie: parenté évolutive. Inférence déduite à partir du degré de similitude. Mais deux séquences similaires ne sont pas forcément dérivées d'un ancêtre commun.

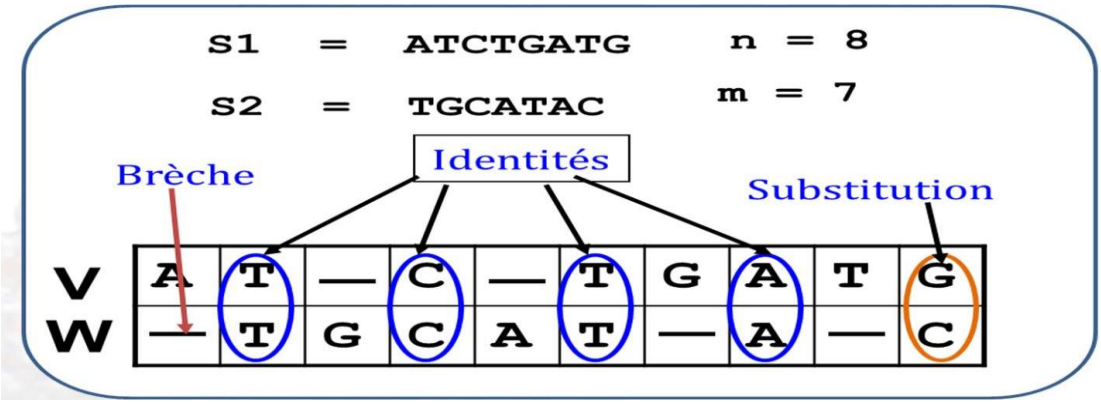
1. **Pourquoi avons-nous besoin de comparer des séquences?**

La comparaison des séquences est un aspect fondamental de la bioinformatique et très souvent la première étape de l'analyse de séquences. Il est nécessaire pour:

- La recherche de fonctions biologiques similaires.
- Construction d'arbres phylogénétiques.
- Identification de mutations dans des gènes.
- Prédiction des sites d'épissage dans les séquences eucaryotes.
- Détection du transfert de gène ....

2. **Alignement et détermination du score :**

Aligner deux séquence, c'est rechercher le maximum d'appariement entre les lettre qui les composent (nucléotides ou résidus d'acides aminés) avec le minimum de mésappariement et des brèches (gaps) (voir schéma).



Un alignement sera considéré comme bon s'il fait correspondre un nombre élevé d'identités, et un nombre minimal d'insertions, de délitions et de substitutions.

Ceci conduit naturellement a l'idée **d'évaluer la qualité d'un alignement** en lui attribuant une note :

Une prime à l'alignement pour chaque identité

Une pénalité pour chaque opération de modification (substitutions et brèches).

La notation de l'alignement **(score total)** peut ainsi être calculée en sommant les primes d'identité et les pénalités des brèches (d'insertions/délétions) et substitution effectuées.

La recherche de similitude entre séquences nécessite la détermination d'un score de similarité

Score Total = Σ Score élémentaires **- Σ** Score pénalités

Exemple de détermination de score avec la matrice unitaire (l'appariement vaut +1, le mésappariement vaut 0 et une brèche vaut -1)

Alignement sans brèches	**Alignement avec brèches**
Séquence 1 ATGACTGGGCCACT **Séquence 2 ATACTGGGACAACT**	
Séquence 1 ATGACTGGGCCACT **Séquence 2 ATACTGGGACAACT**	**Séquence 1 ATGACTGGGCC-ACT** **Séquence 2 AT-ACTGGGACAACT**
8 appariements (match) et 6 Mésappariement (mismatch). **Score 8 - 0= 8**	**12 appariements, 1 Mésappariement et 2 brèches.** **Score 12 - 2= 10**

Les matrices BLOSUM (de Steve Henikoff 1950) (BLOcks SUbstitution Matrix) sont déduites d'alignements de fragments (Blocks) de protéines très éloignées.

Par exemple: **BLOSUM62** est déduite à partir d'un alignement de séquences ayant 62% de similitude. Ces matrices sont bien adaptées aux recherches de séquences dans les banques de données (Blast, FASTA).

Chaque score donne le coût de remplacement d'un résidu par un autre. On note que :

-Les acides aminés rares ont un score élevés (Trp, Cys, His)

-Les acides aminés communs ont des scores faibles (Ala, Leu, Ile,…..)

-Les substitutions conservative entre acides aminés similaire sont peu pénalisantes. Ces substitutions peuvent se produire sans affecter l'activités de la protéine (ex : Lys↔Arg).

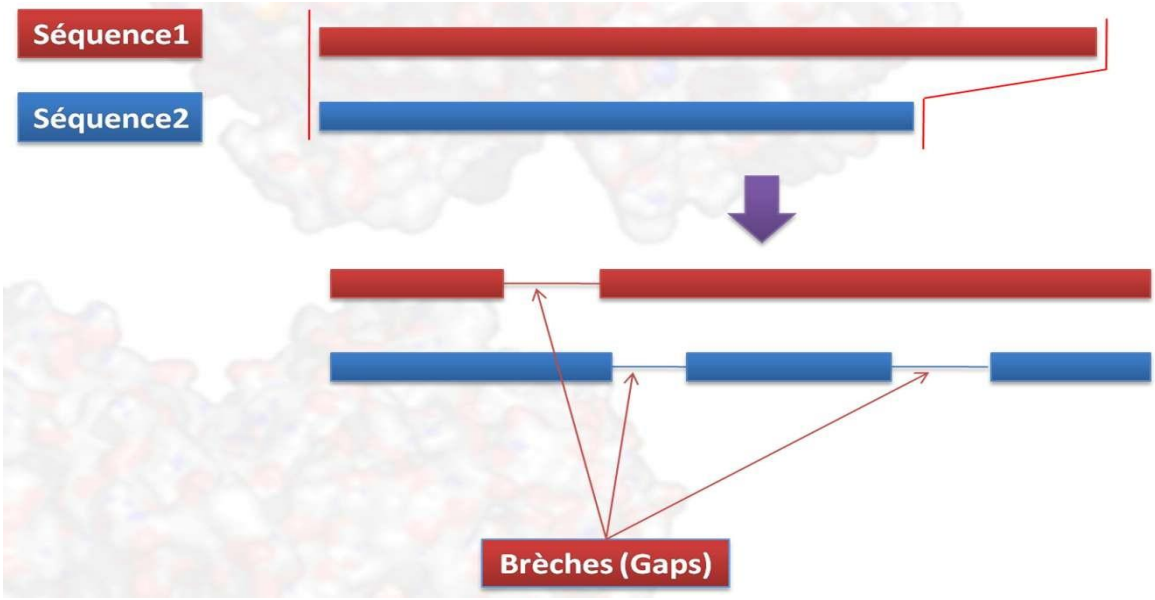
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Blusom 62

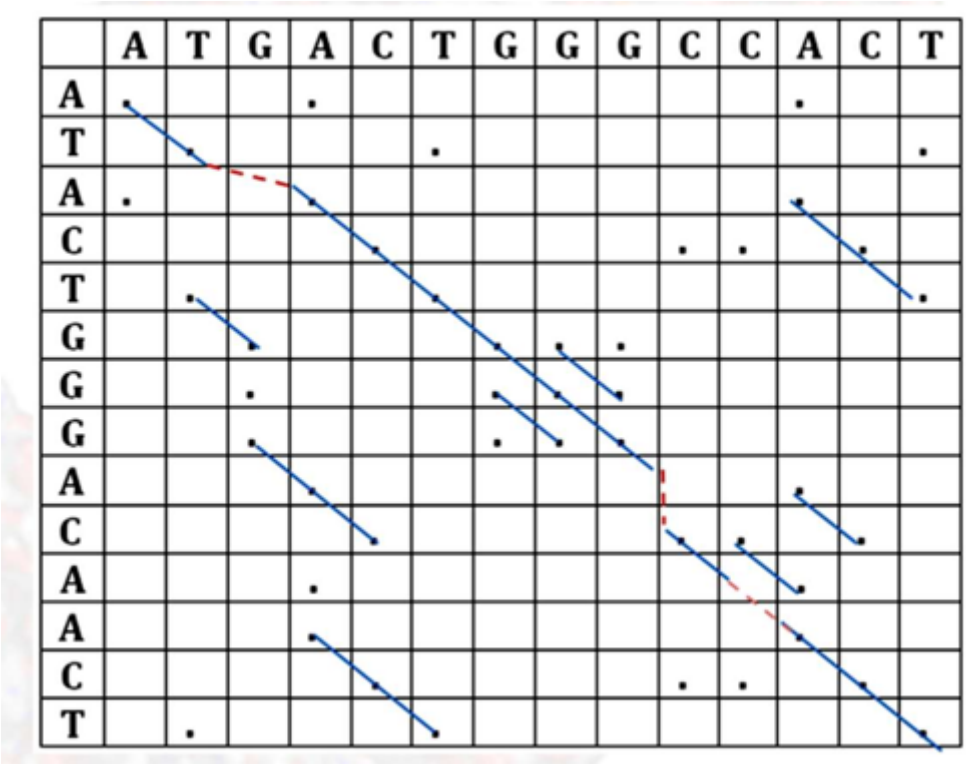
3. Il existe 3 types d'alignement :

3.1. <u>Alignement globale :</u>

Alignement de deux séquences sur la totalité de leur longueur en tenant compte de tous les résidus. Si les longueurs des séquences sont différentes des insertions / délétions sont introduites pour aligner les deux extrémités des deux séquences. Il permet de mesurer le degré de similitude entre deux séquences connues.



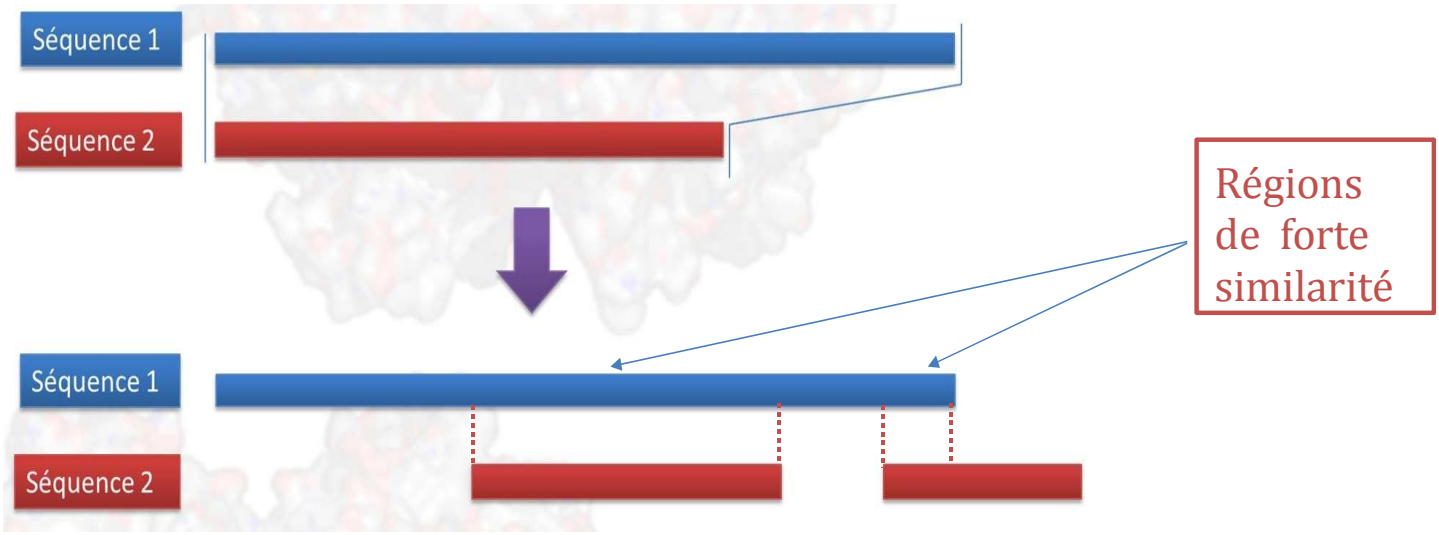
- **Dot-plot :** outil visuel de comparaison de séquences. Les séquences sont positionnées perpendiculairement dans un tableau et on met un point à chaque appariement. La multiplicité des points forme des diagonales. Les décalages correspondent des insertions / délétions et les segments parallèles indiquent des répétitions. Si on reprend l'exemple précédent cela donne le tableau suivant :



Avantage et inconvénients : simple et intuitif. Mais des problèmes de bruit de fond se posent pour les longue séquences. Cela nécessite l'utilisation d'un filtrage : on ne met un point que si **n** caractères sont identiques dans une fenêtre donnée, pour éliminer les segments de similitudes courtes.

3.2. Alignement local

Alignement de deux séquences portant sur des régions isolées et permettant de trouver des segments qui ont un haut degré de similarité. Outil efficace et rapide de recherche dans les bases de données en comparant une séquence inconnue à celles de la banque. (BLAST, FASTA)



Application :

Le fait d'ignorer segments d'ADN non-codant:

-Régions **non codantes** sont plus susceptibles d'être soumises à des mutations que les régions codantes.

-Alignement local entre deux séquences est susceptible d'être entre deux exons.

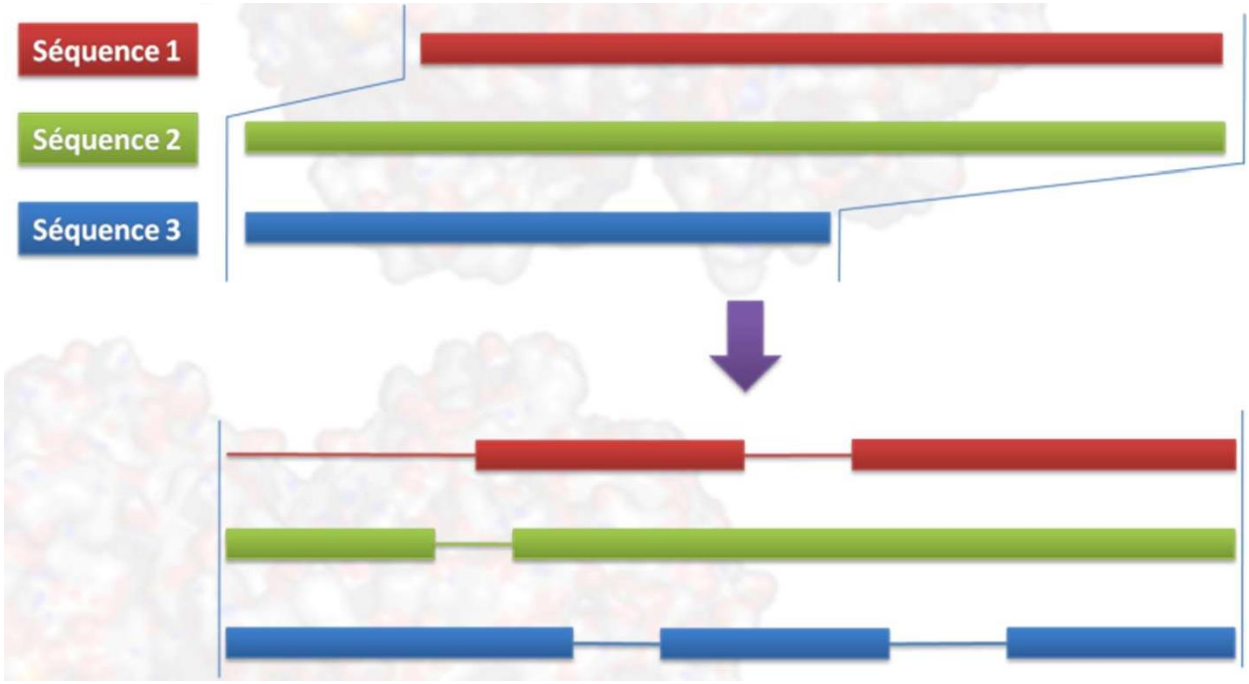
Localisation domaines protéiques:

-Les protéines de type différent et de différentes espèces présentent souvent des similitudes locales.

-Similitudes locales peuvent indiquer "sous-unités fonctionnelles ».

3.3. L'alignement multiple

Alignement portant sur plusieurs séquences à la fois et dans leur intégralité. Il permet de mettre en évidence les relations entre séquences que l'on ne peut pas visualiser en comparant les séquences 2 à 2.



Application :

- Caractérisation de régions conservées des protéines (motifs et domaines conservés).
- Prédiction des structures 2D ou 3D par comparaison avec des séquences et structures connues
- Construction des arbres phylogénétiques des séquences homologues.

4. BLAST

BLAST est l'abréviation de « Basic Local Alignment Search Tool » ou, en français, L'outil de recherche basique d'alignement local. BLAST, quant à lui, cherche les bases de données des protéines et ADNs pour des séquences (sujets) qui ressemblent à notre séquence (requête) utilisée comme mot clé.

En ce qui concerne BLAST, il utilise l'alignement local pour comparer les séquences. Il divise la séquence en question « requête » en morceaux composées de trois acides aminés (en cas des protéines) ou 11 nucléotides (en cas d'ADNs). Ces morceaux sont nommés mots. En cherchant les bases de données de séquences avec ces mots on trouvera plusieurs mots (mots voisins) qui ressemble à ceux de la requête. Les mots voisins appartiennent à un ou plusieurs séquences sujets.

L'unité fondamentale de BLAST est le HSP (High-scoring Segment Pair) (fragments similaires). C'est un couple de fragments identifiés sur chacune des séquences comparées, de longueur égale mais non prédéfinie, et qui possède un score significatif. En d'autres termes, un HSP correspond à un segment commun, le plus long possible, entre deux séquences qui correspond à une similitude sans insertion-délétion ayant au moins un score supérieur ou égal à un score seuil.

La stratégie de la recherche consiste à trouver tous les HSPs entre la séquence recherchée et les séquences de la base.

• Pour déterminer un HSP, des mots de longueur fixe sont identifiés dans **une première étape** entre la séquence recherchée et la séquence de la banque.

• Dans **une deuxième étape**, on cherche à étendre la similitude dans les deux directions le long de chaque séquence, à partir du mot commun, de manière à ce que le score cumulé puisse être amélioré.

• Dans **une troisième étape**, la signification des segments similaires obtenus est évaluée statistiquement. Le score de la similarité est normalisé et évalué en unité standard d'information (bit). Ensuite la probabilité (**E-value**) d'avoir un tel score au hasard est calculé pour cette longueur de segment (m) dans une banque contenant au total (n) nucléotides ou acides aminés. Seuls seront conservés et classés les HSP significatifs, c'est à dire ceux dont la probabilité est la plus faible.

Il existe en fait deux versions de l'algorithme une sans insertion délétion, BLAST 1.0 (1990) et l'autre avec insertion délétion, BLAST 2.0 (1997).

Ce logiciel possède en fait plusieurs programmes de comparaison avec les bases de données :

- BLASTN (pour comparer une séquence nucléique contre base nucléique),
- BLASTP (Pour comparer une séquence protéique contre base protéique),
- BLASTX (comparaison de séquence nucléique (traduite en 6 phases) contre base protéique),
- TBLASTN (comparaison de séquence protéique contre base nucléique (traduite en 6 phases)),
- TBLASTX (comparaison de séquence nucléique (traduite dans les 6 phases) contre base nucléique (traduite dans les 6 phases)).

E-value : La signification statistique des alignements produits par un BLAST est mesurée par E-value (expected-value). Elle indique le nombre d'alignements différents ayant le même degré de similitude et que l'on peut espérer trouver par hasard dans la banque, même s'il n'existait pas de vraie séquence similaire.

Si E=10-2 cela signifie que 1 alignement sur 100 sera trouvé par hasard.