**CHAPITRE III : DEUX TYPES DE BIOINFORMATIQUES : LA SEQUENCE DES NUCLEOTIDES ET LA SEQUENCES DES ACIDES AMINES**

**Et aprés la séquence?**

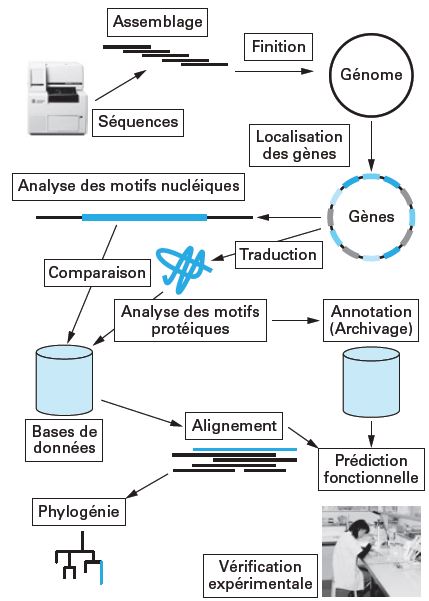
AGTAGTGTGTGCCCGTCTGTTGTGTGACTCTGGTAGCTAGAGATCCCTCAGACCCTTTGT  
GGTAGTGTGGAAAATCTCTAGCAGTGGCGCCCGAACAGGGACTTAAAAGCGAAAGTAAGA  
CCAGAGGAGATCTCTCGACGCAGGACTCGGCTTGCTGAAGTGCACTCGGCAAGAGGCGAG

1. **Annotation du génome**

**I.1. Définitions**

* La séquence du génome peut être comparée à un livre écrit dans une langue inconnue. Il faut donner un sens au texte du génome: trouver les mots, les phrases, le sens de chaque phrase et les liens entre elles.
* Annoter un génome, c’est donner un sens à sa séquence : trouver les mots, la ponctuation, le sens de chaque phrase et les liens entre elles …. Ces analyses font partie de la discipline nommée **Genomique**.
* Etablir un inventaire (par des méthodes bioinformatiques) de tous les éléments génétiques (emplacement de génes, leur traduction, localisation de promoteurs etc…) présents dans un génome, ainsi que leurs fonctions.

**I.2. Différentes étapes dans l’annotation des génomes**



**I.3. Types d’annotation**

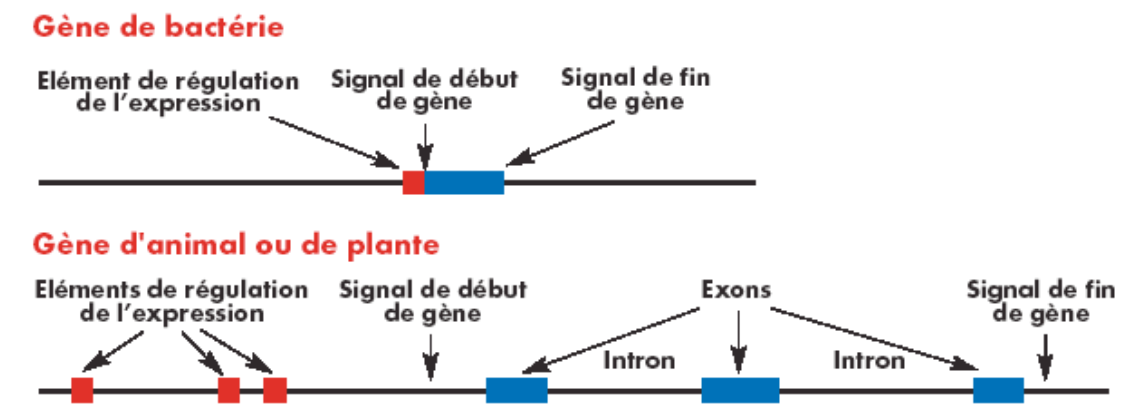
Il existe deux niveaux d’annotation des génomes :

* **Annotation structurelle:** Inventaire et analyse des éléments presents dans un génome :  
  • Identification de tous les génes codants pour des protéines et des ARN (ribosomique, de transfert).  
  • Identification de sites promoteurs, de terminaison de la transcription et de la traduction, d’epissage, intron, exons…etc
* **Annotation fonctionnelle**Identifier la fonction de tous les gènes détectés lors de la première étape

**I.3.1. Annotation structurelle**

* **Comment reconnaitre un gène?**

*Caractéristiques des gènes*



* **Cadre de lecture ouvert « ORF »**

Un «cadre de lecture ouvert » ou «Open Reading Frame (ORF)” est une region d’ADN entre un codon START et un codon STOP



* **Séquence codante « Coding sequence (CDS) »**

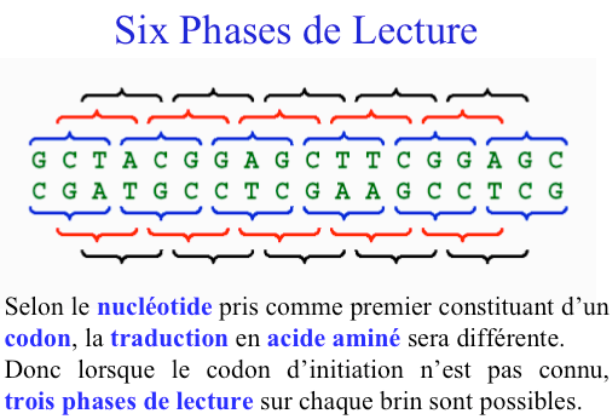
Si on établit qu’un ORF code pour une protéine ou un ARN on le désigne comme séquence codante (CDS) = gène



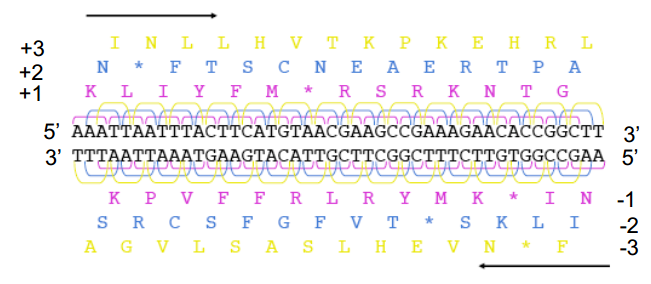
**Attention:** Un gène (CoDing Sequence CDS) est un ORF, mais un ORF ne correspond pas forcément à un gène

* **Recherche d’ORFs**

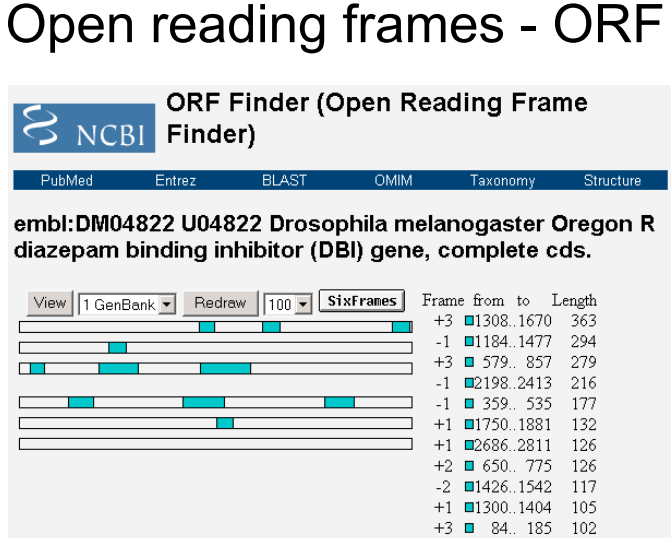
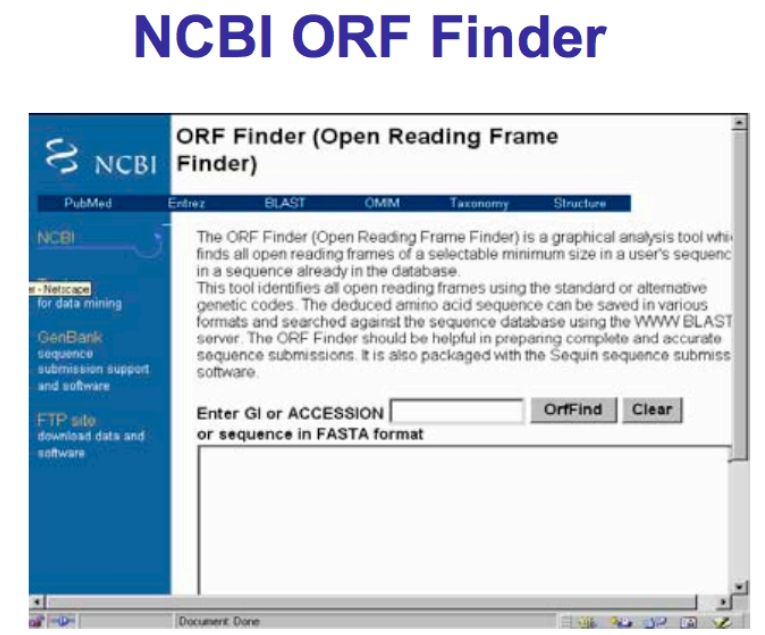
Chaque séquence d’ADN codante peut être traduite en six phases de lectures différentes: 3 dans un brin et 3 sur le brin complémentaire.



***Exemple***



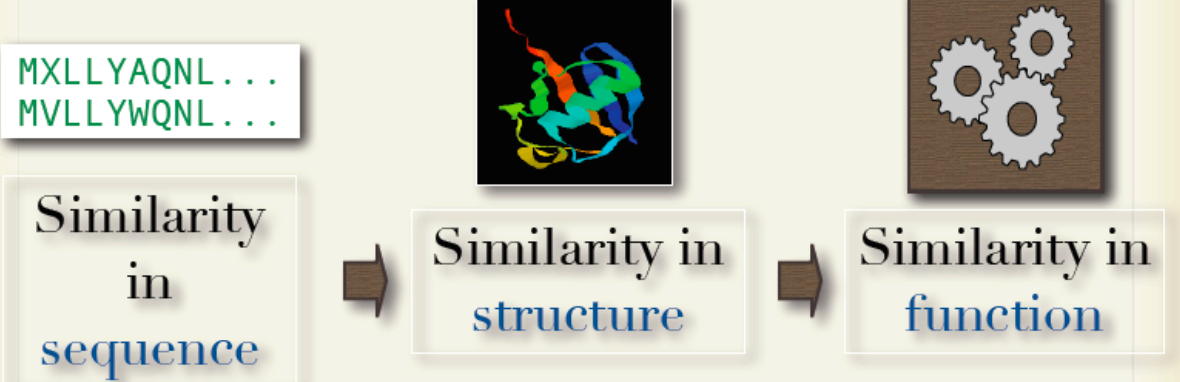
**Figure 8: Representation des six cadres de lecture ORFs d’un polypeptide**

**Figure 9: Obtention des six cadres de lecture ORFs d’un polypeptide dans le portail NCBI**

**I.3.2. Annotation fonctionnelle**

Toute protéine inconnue ayant un pourcentage de similarité suffisamment élevé avec au moins une protéine connue dont la fonction est identifiée se verra attribuer cette fonction.

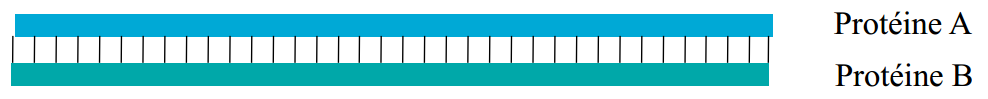


* **Comparaison des séquences**

Pour comparer deux séquences il faut les aligner afin de :  
•  Identifier les points communs ou les différences entre deux séquences.  
•  Retracer l’histoire évolutive des séquences en simulant les mutations:  
- Changement d’un nucléotide par un autre (substitution).  
- Ajout ou suppression d’un ou plusieurs nucléotides (insertion ou deletion)  
 **Exemple:** 

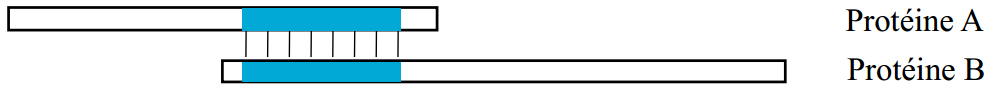
* **Types d’alignements**

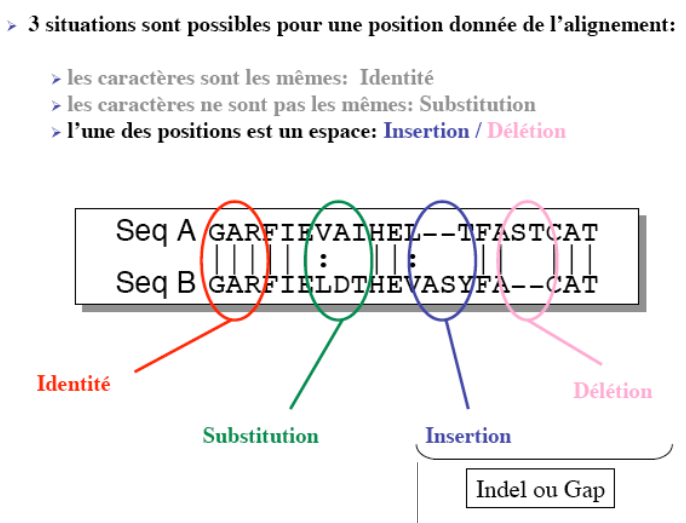
• **Alignement global:** prise en compte de la totalite de la séquence



• **Alignement local:** recherche de la région de plus forte similarité entre deux séquences

***Domaine***

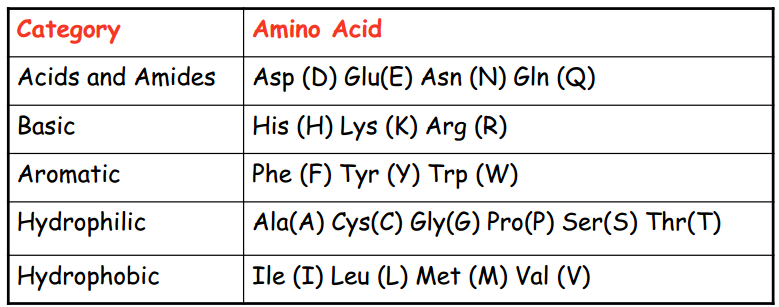




1. **Propriétés physico-chimiques des acides aminés**

* Les acides aminés composant une proteine peuvent avoir des propriétés physico-chimiques similaires. On sait que la structure 3D dépend de ces caractéristiques
* Une similitude au niveau de ces propriétés sera suffisante pour permettre la substitution d’un acid aminé en un autre sans perturber la fonction de la protéine (par exemple, échange de l’acide aminé hydrophobe valine en leucine).
* Lors de la comparaison de deux séquences protéiques, nous devons prendre en compte ces similitudes et pas seulement les identités.
* **Logiciels d'alignement de deux séquences**•  BLAST

•  FASTA (voir chapite 8)



1. **Analyse des séquences nucléiques et protéiques**

L’analyse des séquences est à l’origine même du terme « *bioinformatics* ».

**III.1. Analyse de séquences nucléiques**

* Recherche de phase de lecture ouverte (gène) et de signaux de régulation de la transcription et de la traduction, détection de bornes introns/exons.
* R[echerche de régions transcrites](http://biochimej.univ-angers.fr/Page2/COURS/9ModulGenFoncVeg/5MethEtudGenFonc/4EST/1EST.htm) (EST) - profil d'expression des gènes (puces à ADN, analyse d'images).
* Sélection d'amorces (PCR), calcul de la température de fusion (Tm).
* Etablissement de cartes de restriction (recherche de sous-séquence).
* Recherhce de régions riches en (CG par exemple).
* Recherche de répétitions.
* Recherche de motifs à l'aide d’expressions régulières.
* Assemblage à partir de fragments séquencés.
* Détection de polymorphismes de nucléotide simple ou d'insertion / délétion.
* Reconstruction d'arbres phylogèniques.
* Analyse de génomes entiers (génomique structurale, synténie) - réseaux de gènes

**III.2. Analyse de séquences protéiques**

* Traduction *in silico.*
* Régulation de la traduction.
* Taux de synthèse des protéines ([protéomique](http://biochimej.univ-angers.fr/Page2/COURS/9ModulGenFoncVeg/6Proteomique/1Proteomiq.htm)).
* Prédiction de [modification post-traductionnelles](http://biochimej.univ-angers.fr/Page2/COURS/7RelStructFonction/2Biochimie/2ModifPOSTtraduc/1ModPostTrad.htm).
* recherche de motifs structuraux : détection de [sites actifs](http://biochimej.univ-angers.fr/Page2/COURS/4EnzymologieLicence/1COURS1/111Cours.html#Specificite) (enzymes), de domaines, de [types de repliement](http://biochimej.univ-angers.fr/Page2/COURS/7RelStructFonction/4Repliement/1Introduction.htm) (famille de protéines).
* prédiction de structures secondaires et modélisation de structures tridimensionnelles.
* Composition, masse moléculaire.
* Calcul du coeeficient théorique d'absorption.
* Calcul du pH du point isoélectrique (pI) théorique.
* Fragments peptidiques obtenus par hydrolyse enzymatique.

**III.3.** **Prédiction sur les séquences**

**III.3.1. Prédiction sur les séquences nucléiques**

* Prédiction de structure d'un gène.
* Prédiction de signaux.
* Prédiction de structure secondaire d'un ARN.

**III.3.2. Prédictions sur les séquences protéiques**

* Famille de protéines.
* Famille et signature.
* Localisation subcellulairE.
* Segments nucléaires.
* Segments transmembranaires.
* Sites spécifiques (O-glycosilation, antigénicité, etc..).
* Peptide signal.
* Structure secondaire.
* Structure 3D.