**Chapitre 6 : Annotation des génomes**

**Introduction**

 Une succession brute de nucléotides n’à aucun sens ; l’annotationest le travail d’analyse qui permet d’expliquer ou de proposer des hypothèses pour les propriétés biologique d’un génome. Pour cela, il faut rechercher les objets génétiques présents dans le génome puis essayer de leur attribuer des fonctions. Ainsi, l’annotation est l’antichambre de l’expérimentation ; elle conduit à élaborer des protocoles expérimentaux qui valident ou invalident la fonction supposée de l’objet biologique. Classiquement on distingue trois étapes principales dans le processus d’annotation d’un génome :

* **L’annotation syntaxique :** c’est l’étape qui permet d’identifier les objets génétiques présentant une pertinence biologique **(**séquences codantes, ARN, séquence répétées**…** etc**.).**
* **L’annotation fonctionnelle :** c’est l’étape qui permet de prédire les fonctions potentielles des objets génétiques préalablement identifie (similitudes de séquence, motifs, structure…etc.) et de collecter d’éventuelles informations expérimentales.
* **L’annotation relationnelle :** c’est l’étape qui permet de déterminer l’interaction que les objets biologies préalablement identifie sont susceptibles d’entretenir (familles de gènes, réseaux de régulation).

 L’ensemble de ces informations seront ensuit stockées dans des bases de données consultables par l’expérimentateur.

**1. Annotation syntaxique : La recherche d’objets génétiques :**

**Principe :** La recherche d’objet génétiques passe par la recherche de gènes au sens large c'est-à-dire **:** toute séquence qui transcrite ou traduite, peut avoir un rôle dans les fonctionnements biologiques de la cellule cela recouvre**:** les séquences codantes (CDS : coding séquence), les ARNm et les ARN non traduits (ARNt , ARNr, RNAsn).

1. **La recherche de signaux de séquence codante chez les procaryotes :**
* L’annotation syntaxique des génomes procaryotes est relativement plus aisée que celle des génomes eucaryotes pour les raisons suivantes**:**

Les génomes procaryotes sont plus petits que les génomes eucaryotes et ont surtout une densité de codage bien plus importante de l’ordre de 80 **-** 90 % à quelle que pourcentage chez l’être humaine

* les gènes procaryotes sont fréquemment organisés en opéron, c’est-à-dire qu’une seule unité de transcription peut contenir plusieurs séquences codantes.
* les gènes procaryotes ne sont pas morcelés (fragmentés) contrairement à ceux des eucaryotes.

 **A\_1 : ORF et CDS chez les procaryotes :**

 La phase ouverte de lecture (ORF) est la région de l’ADN qui séparé deux codants de terminaison de traduction. Dans celle-ci, une séquence codante (CDS) débute toujours par un codon d’initiation de la traduction et se termine toujours par un codon de terminaison de la traduction c'est-à-dire : la séquence codante est parfois appelé ORF.

**A\_2 : la séquence de Shine-Dalgarno ou site de liaison au ribosome :**

 (Ribosome Banding Site ou RBS) se situe entre 3 à 10 nucléotides en amont du codon Start, c’est une région riche en purine qui permet au ribosome de se fixer spécifiquement sur les AUG correspondant à un véritable codon Start.

**A\_3 : le promoteur :**

 Région reconnue spécifiquement par le complexe ARN polymérase et le facteur sigma (Partie qui assure la spécificité de l’initiation de la transcription) le promoteur est constitué de deux éléments :

* la séquence conservée localisée environ 10 nucléotides avant le site d’initiation de la transcription : la boite TATA
* La boite -35

 Dans un opéron le promoteur ne se localise qu’en amont de la premier CDS puisque l’opéron est une unité de transcription.

**A\_4 : le terminateur de transcription :**

Séquence palindromique riche en GC suivies de séquences riche en A. c’est une séquence grâce à laquelle le complexe de transcription va se désassembler et ainsi terminer la transcription.

**B- recherche de signaux de séquences codantes chez les eucaryotes :**

Chez les eucaryotes, l’annotation syntaxique est nettement plus compliquée pour les raisons suivantes :

* les génomes ont une faible densité de codage (régions répétées sans séquences codante)
* les gènes eucaryotes subissent des modifications post-transrationnelles.
* les gènes eucaryotes sont morcelés.

 **B.1. promoteur et signaux 5’ :**

 Les séquences promotrices reconnues par l’ADN polymérases : chaque types d’ADN polymérase reconnue un type spécifique de gènes

* les sites de liaison aux facteurs de transcription.

 L’initiation (INR) : une séquence faiblement conservée qui se trouve prés de site de début de transcription entre les positions -3 +5.

 Pour les signaux de traduction, on recherche en particulier le site de liaison au ribosome ou séquence de KOZAK localisé en amant du codant Start.

 **B.2. jonctions exons- introns :**

 Les introns possèdent quatre signateurs importants dont les deux premiers indiquent les jonctions exons-introns :

* le site donneur d’épissage GT à l’extrémité 5’ de l’intron
* le site accepteur d’épissage AG a l’extrémité 3’ de l’intron
* le pont de branchement (A)
* la région riche en pyrimidine entre le point de branchement et le site récepteur

 **B.3. les signaux 3’ :**

L’exon terminal contient un ensemble de signaux de terminaison de la transcription

* le signal de poly adénylation :5’\_AAUAAA3’ ou 5’AUUAAA3’
* le signal de (un di nucléotide C A, localisé 10 à 30 base en aval du signal de poly adénylation.
* une région riche en G U

 **Des méthodes informatiques pour l’annotation syntaxique :**

 Les meilleures méthodes sont celles qui combinent la détection de signaux de gènes et l’analyse du contenu des gènes. Elles utilisent des algorithmes qui ; après une phase d’apprentissage sur un ensemble de données permettent de différencier les régions génétique des régions inter génique. Ces méthodes permettent de prédire des caractéristiques locales de séquences moléculaires tout en intégrant des connaissances provenant de recherchse antérieurs dans les analyses.

**2 annotations fonctionnelles :**

 Permet d’attribuer à des objets génomiques prédits par l’annotation syntaxique des fonctions potentielles. En aucun cas l’annotation ne permet d’avoir accès à la fonction réelle ; seule l’expérimentation l’autorise. L’annotation fonctionnelle est fondée sur la recherche de similarité avec des séquences nucléotidique, des séquences d’acides aminés ou éventuellement des structures déjà décrites dans les bases de données.

 En générales l’annotation s’effectue en deux étapes une phase automatique qui s’effectué grâce à des programmes informatiques de comparaison et une phase manuelle au cours de laquelle l’annotateur peut corriger le cas échéant la première phase.

 Les outils peuvent être compares avec des programmes comme FASTA ou BLAST. Ces instruments de recherche de similarité reposent sur la notion d’alignement local.

 Les algorithmes d’alignement local recherchent dans des paires de séquences des régions isolées qu’ont un haut degré de similitude.

 **EXP :** l’utilisateur fournit une séquence requête qui est alors comparée à toutes les séquences d’une base de données choisie.

 Le résultat du programme classe les résultats de similitude en fonction d’un indice de significativité appelé E value (valeur attendue), le résultat le plus significatif étant le premier de la liste (la valeur E se rapproche à 0, moins la similitude est due au hasard)

 C- **Annotation relationnelle :**

 Vise a déterminé comment les gènes et protéines interagissent entre eux pour accomplir une tache spécifique dans la cellule pour cela, des programmes informatiques utilisent les annotations des gènes codant des enzymes dans le génome étudie, de même, à partir des réseaux d’interaction connus, des bases de données sont spécialisées dans la reconstitution de réseaux d’interactions entre des gènes, des protéines, et des procédées cellulaire. Tous ces approches de reconstitution reposent sur l’établissement de similarités de séquences et ne peuvent donc pas prédire de nouvelles voies de synthèse ou de nouveaux réseaux d’interaction.