

Introduction

L'annotation est le travail d'analyse qui permet d'expliquer ou de proposer des hypothèses pour les propriétés biologiques 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 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 :

- A. **L'annotation syntaxique** : c'est l'étape qui permet d'identifier les objets génétiques présentant une pertinence biologique (séquences codantes, ARN, séquences répétées...etc.).
- B. **L'annotation fonctionnelle** : c'est l'étape qui permet de prédire la fonction potentielle des objets génétiques. Préalablement, identifier (similitudes de séquences, motifs, structures....etc.) et de collecter d'éventuelles informations expérimentales (littérature, jeux de données à grande échelle).
- C. **L'annotation relationnelle** : c'est l'étape qui permet de déterminer les interactions que les objets biologiques préalablement identifiés sont susceptibles d'entretenir (famille de gènes, réseaux de régulation, réseaux métaboliques.....etc.)

L'ensemble de ces informations seront ensuite stockées dans des bases de données consultables par l'expérimentateur. L'information joue un rôle essentiel dans l'annotation en raison de la puissance de calcul nécessaire pour effectuer les recherches et la masse énorme d'information générée la plupart des outils et bases des données présents ici sont d'accès libres par biais d'internet. D'autre part les outils bio-informatiques nécessaires aux différentes étapes de l'annotation peuvent être regroupés dans des plateformes d'annotation. Ces plateformes ont des interfaces conviviales utilisables par des biologistes non informaticiens.

1. Annotation syntaxique : la recherche d'objets génétiques

1.1.Principe

La recherche d'objets génétiques passe principalement par la recherche de gènes au sens large, c'est-à-dire toute séquence qui transcrit et/ou traduit. Peut avoir un rôle dans le fonctionnement biologique de la cellule cela recouvre donc les séquences codantes (c'est à

Chapitre N 6 l'annotation des génomes

(dire séquence traduites en protéines) les ARN on traduit ARNt, ARNr, petit ARN, ARN interférent, etc.)

La recherche de séquences codantes, bien qu'insuffisante pour la bonne compréhension du fonctionnement d'un génome est néanmoins celle qui est la plus développée et pour laquelle un grand nombre d'outils informatiques existe.

1.2. La recherche de signaux de séquences codants chez les eucaryotes

Chez les eucaryotes, l'annotation syntaxique est nettement plus compliquée. Pour les raisons suivantes :

- ✚ Les génomes eucaryotes ont une faible densité de codage. Il y a donc de larges régions génomiques sans séquence codante.
- ✚ Les gènes eucaryotes subissent des modifications post-transcriptionnelles : l'addition d'une coiffe 7-méthylguanilate à l'extrémité 5' du pré-ARN messager. Un clivage puis une polyadénylation à son extrémité 3'. L'épissage des ARN pré-messagers.
- ✚ Les gènes eucaryotes sont morcelés. Ils subissent des modifications de la séquence nucléotidique (épissage) du pré-ARN messager. L'épissage consiste en l'excision d'un ou plusieurs séquences (introns). Les séquences non excisées (exons). Forment après raboutage entre elles la séquence codante. Par conséquent de l'ORF.
- ✚ Enfin, l'épissage peut être alternatif. Différents profils d'épissage existent pour un même pré-ARN messager et par conséquent un gène peut produire différents CDS.

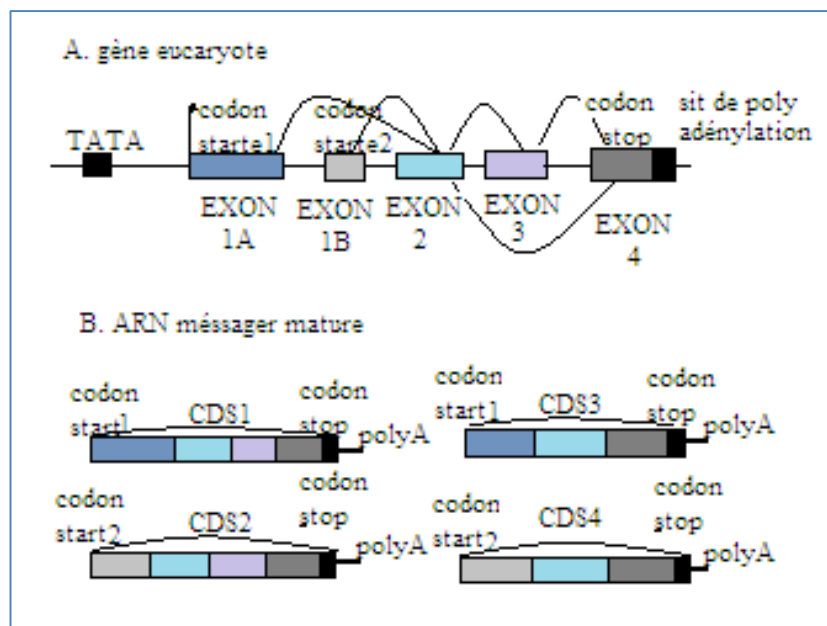


Figure 5.1. la notion de séquence codante chez les eucaryotes.

Chapitre N 6 l'annotation des génomes

A. Promoteur et signaux 5'

Outre les codons « start » et les codons « stop » différents type de signaux marquant la région 5' de la séquence codante peuvent être recherchés.

Les signaux de transcriptions sont les suivants :

- ✚ Les séquences promotrices : reconnues par les ADN polymérase chez les eucaryotes. Il y a 3 types d'ARN polymérase (ARNpolI, ARNpolII, ARNpolIII). Chaque ARN polymérase reconnaît un type de promoteur.
- ✚ Les promoteurs pol I se trouvent en amont des gènes d'ADN ribosomiaux.
- ✚ Les promoteurs pol II se trouvent en amont des gènes d'ADN messagers
- ✚ Les promoteurs pol III se trouvent en amont des gènes d'ARN ribosomiaux 5s et l'ARN de transfert.

En amont des séquences codant des protéines, on recherche donc une boîte TATA, une séquence conservée riche en A/T et de 8 nucléotides qui se trouvent dans les promoteurs pol II 25 à 30 nucléotides en amont du site d'initiation de la transcription (figure 19) :

- ✚ Les sites de liaison aux facteurs de transcription
- ✚ L'initiateur (INR), une séquence faiblement conservée qui se trouve près du site de début de transcription entre les positions -3 et +5.
- ✚ Les îlots CPG, ce sont des régions de 1 à 2 kb riches en di nucléotides CG qui sont fréquemment associées aux régions 5' des gènes de vertèbres et qui s'étendent sur le promoteur et le premier exon.

Pour les signaux de traduction, on recherche en particulier le site de liaison au ribosome ou séquence de KOZAK localisée en amont du codon « start ».

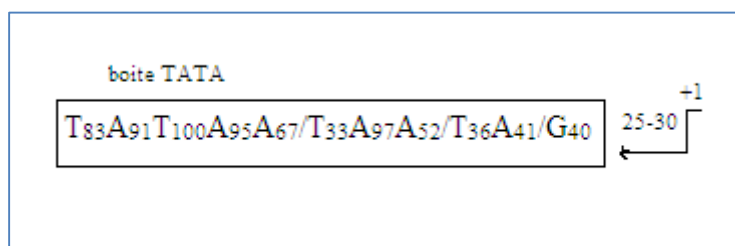


Figure 5.2. : Séquence consensus de la boîte TATA chez les eucaryotes.

Chapitre N 6 l'annotation des génomes

Le nombre en indice indiquent le pourcentage d'occurrence de chaque nucléotide dans la définition de la séquence consensus +1 indique le site d'initiation de la transcription.

B. Jonction exons-introns

Les introns possèdent quatre signatures importantes dont les deux premiers indiquent les jonctions exons-introns. (Figure 5.3)

- ✚ Le site donneur /GTRAGT à l'extrémité 5' de l'intron. Le dinucléotide GT est systématiquement exclu des ARNm matures.
- ✚ Le site accepteur NYAG/G à l'extrémité 3' de l'intron. Le di nucléotide AG est systématique exclu des ARNm matures.
- ✚ Les point de branchement CTRAY avec un A qui joue un rôle central dans le processus d'épissage.
- ✚ Une région riche en pyrimidine entre le point de branchement et le site récepteur.

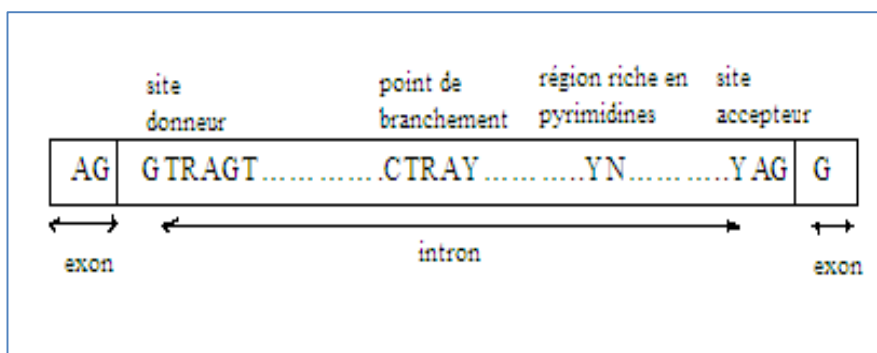


Figure 5.3 Signaux de jonction exon intron chez les vertèbres.

Y pyrimidine R purine.

C. Signaux3'

L'exon terminal contient un ensemble de signaux indicateurs de la terminaison de la transcription. Pour les ARN messagers transcrits par ARN polymérase II. Ces signaux sont également nécessaires à la polydénylation car ce dernier événement est couplé à la terminaison de la transcription :

- ✚ Le signal de polyadénylation : 5' AAUAAA3' ou 5'AUUAAA3'.
- ✚ Un signal de clivage. C'est un dinucléotide.CA assez mal conservé. Localisé 10 à 30 bases en aval du signal de poly dénylation. A ce niveau, l'ADN est clivé avant l'ajout de la queue poly A.
- ✚ Une région riche en GU de séquence variable, 20 à 40 bases après le site de clivage.