# TP Bioinformatique - M1 SPGF

Etude d'une famille de protéines, les $\beta$-lactoglobulines

Novembre 2007

## 1 Annotation génomique

Vous disposez pour votre étude d'une séquence génomique issue du chien (*canis familiaris*). Le but de cette partie est de déterminer la présence de gène(s) dans cette séquence d'environ 50 Kb par des méthodes bioinformatiques classiques.

### 1.1 Tracer un dot-plot

`Dotter` est un outil permettant de tracer un dotplot. L'utilisateur peut ainsi comparer visuellement deux séquences ou une séquence contre elle-même dans le but de mettre en évidence des sous-séquences répétées. L'utilisation du filtre permet d'affiner le résultat.

*Tracer un dotplot de la séquence qui vous est fournie contre elle-même, que remarquez-vous ?*

### 1.2 Elimination des séquences dupliquées

A l'aide de l'outil `Repeatmasker`, nous vous demandons d'éliminer les séquences répétées qui sont connues pour ne pas contenir de régions codantes. Notez que ce programme est aussi disponible sous la forme de service web sur http://www.repeatmasker.org/

*Quel est le pourcentage de séquences éliminées par* `Repeatmasker` *? Quelles sont leurs natures ?*

Refaites un dotplot de la séquence générée par `Repeatmasker` contre la séquence initiale.

*Que pouvez conclure sur les différentes "bandes" qui apparaissent sur le graphe ?*

A partir de maintenant, nous ne travaillerons plus qu'avec la séquence issue de `Repeatmasker`, où les régions répétées sont masquées.

### 1.3 Détection des gènes

#### 1.3.1 Ab initio

**Détection des cadres de lectures ouverts.** Dans un premier temps, on se propose de recenser toutes les ORF présentes dans la séquence. Utilisez l'outil `ORFfinder`, adaptez les paramètres. Notez que ce programme est aussi disponible sous la forme de service web sur http://www.ncbi.nlm.nih.gov/projects/gorf/

*Combien d'ORF sont présentes dans la séquence ? Quelle est la plus longue ?*

**Recherche de sites fonctionnels, détection des séquences signaux.** Ces méthodes recherchent les sites fonctionnels présents dans les gènes, soit par homologie à partir d'une base de données de signaux connues, soit en modélisant le signal grâce à un profil de type PSSM ou HMM.

Utilisez tout d'abord l'outil POLYAH sur le site http://softberry.com/berry.phtml qui permet la reconnaissance des régions de polyadénylation et de coupure à l'extrémité 3', puis réalisez un BLAST contre la banque de données EPD (*Eukaryotic Promoter Database*) sur le site http://www.ch.embnet.org/software/aBLAST.html.

**Méthodes d'apprentissage automatique.** Les méthodes de type GENSCAN (http://genes.mit.edu/GENSCAN.html) ou FGENESH (http://mendel.cs.rhul.ac.uk/mendel.php?topic=fgen-file) prennent en compte de multiples signaux (épissage, polyA, régions promoteur, contenu en codons et dicodons ...) et les intègrent par utilisation d'une modélisation probabiliste de type modèle de Markov. Ces méthodes modélisent la structure complète des gènes en utilisant toutes les informations qui les caractérisent (par exemple les biais de composition entre les séquences codantes et non codantes). Elles sont performantes et donnent des résultats satisfaisants sur les gènes de structure canonique.

A l'aide des deux méthodes, réalisez une première cartographie de la séquence.

*Combien de gènes sont détectés ? Quelles sont leurs tailles ? Existe t-il des différences entre les deux prédictions ?*

### 1.3.2 Par homologie

**Homologie seule** Réalisez un BLASTN sur le site du *NCBI* (http://www.ncbi.nlm.nih.gov/) contre la banque EST (Expressed Sequence Tags) qui donne une information tissulaire (ubiquitaire ou tissu spécifique ?).

*A quoi correspondent les EST ? Quelle banque devriez vous choisir ?*

Pour répondre à la question d'un épissage alternatif ou non, on vous demande également de réaliser un BLASTN contre une banque de mRNA.

Pour une recherche d'homologues sur un génome complet, il faudrait utiliser PMATCH ou EXONERATE qui sont des outils plus rapides que BLASTN, mais au détriment de la sensibilité. Nous pouvons néanmoins réaliser un BLASTN contre la banque CHROMOSOME chez l'homme.

*Que remarquez-vous ?*

Réalisez enfin un BLASTX contre une banque de séquences protéiques (typiquement SWISSPROT) en limitant votre recherche à un seul organisme (chien, homme, bovin, ...).

*Quelles différences observez-vous entre ces différentes requêtes ? Sur quel(s) chromosome(s) sont situés les différents gènes codants pour les protéines homologues retenues ? Que pouvez-vous conclure ?*

**Méthodes combinées, homologie et signaux.** Ces méthodes utilisent l'information obtenue par homologie et *ab initio*, elles sont très utiles dans le cas d'étude de génomes entiers ou de grandes séquences nucléiques. A l'aide des outils SIM4 et GENEWISE http://www.ebi.ac.uk/Wise2/, réalisez une dernière cartographie de votre séquence.

*Quels sont les signaux utilisés par ces algorithmes ?*

**Comparaison génomique.** L'idée primordiale est que les régions codantes sont plus conservées que les régions non codantes à travers l'évolution. Ainsi, en comparant les génomes de deux espèces,

les régions d'homologies devraient indiquer les régions codantes. C'est ce que réalise le programme EXOFISH en prenant pour référence le génome de *Tetraodon nigroviridis* qui présente l'avantage d'être très pauvre en régions non codantes.

*Que révèle ce programme sur notre séquence d'intérêt ?*

### 1.4 Bilan

Comparer vos résultats avec l'annotation automatique sur ENSEMBL. Si vous avez le temps, utilisez ARTEMIS pour récapituler l'ensemble de vos résultats.

## 2 Analyse des séquences protéiques

### 2.1 Analyse de la famille des $\beta$-lactoglobulines

A l'aide de l'outil SRS (*Sequence Retrieval System*) disponible sur le site http://www.ebi.ac.uk/, établir une liste **informative** des $\beta$-lactoglobulines (protéines d'environ 180 AA) par interrogation sur la banque UNIPROT.

*Est-ce une banque "redondante" ? Chez quelques organismes les $\beta$-lactoglobulines sont-elles présentes ?*

Réaliser une recherche d'homologues avec les programme BLAST et PSI-BLAST sur les séquences que vous venez d'extraire. Utilisez différentes matrices de substitution (BLOSUM62 / PAM70).

*Que pouvez conclure quant à la glycodéline humaine ?*

### 2.2 Recherche de motifs essentiels

Réaliser un alignement multiple de ces différentes séquences en utilisant les programmes CLUSTALW puis 3D-coffee[1].

*Analyser les éventuelles différences entre* CLUSTALW *et* 3D-coffee*. Dégagez les résidus conservés et commentez.*

A l'aide du programme PRATT, qui recherche des motifs conservés pour une liste de séquences protéiques, extraire les signatures des $\beta$-lactoglobulines. Vous utiliserez tout d'abord le programme sur la liste des séquences non alignées, puis sur les alignements multiples que vous avez obtenus.

*Où sont placés les résidus conservés sur les signatures générées par* PRATT *?*

Les $\beta$-lactoglobulines appartiennent à une famille plus large dont les protéines sont présentes dans de nombreux organismes, tant procayotes qu'eucaryotes.

*Déterminez cette superfamille. Existe t-il un motif* PROSITE *pour celle-ci ? Comparer les signatures définies précédemment avec le motif consensus de la superfamille.*

Afin d'évaluer la pertinence des signatures générées par PRATT, utilisez l'outil PATTINPROT pour des valeurs de similarité comprises entre 60% et 100%. A l'aide des formules suivantes :

$$Sp = \frac{T^-}{T^- + F^+} \qquad Se = \frac{T^+}{T^+ + F^-} \tag{1}$$

Calculez les valeurs de sensibilité et de spécificité pour chaque sortie PATTINPROT.

[1] Utilisez la structure dont le code PDB est 1BEB pour réaliser votre alignement

*Que pouvez-vous conclure ?*

### 2.3 Construction des arbres phylogénétiques

A partir de la liste établie par PROSITE, construire l'arbre phylogénétique de la superfamille décrite ci-avant avec le programme NJPlot.

*Que pouvez-vous observer concernant la place des $\beta$-lactoglobulines au sein de cette famille ? Que pensez-vous des valeurs de bootstrap portées sur le dendogramme ? Où est située la glycodéline humaine ?*

A partir de l'alignement multiple des séquences protéiques des $\beta$-lactoglobulines, générez l'arbre phylogénétique.

*Quelles observations pouvez-vous faire ? Essayer d'établir le schéma évolutif de cette famille de protéines. Conclure.*

## 3 Analyse structurale

### 3.1 Prediction des structures secondaires

Pour la suite de notre étude, nous nous intéresserons plus spécifiquement à la $\beta$-lactoglobuline bovine[2]. Vous utiliserez l'outil Consensus secondary structure prediction disponible sur le site http://npsa-pbil.ibcp.fr ainsi que du serveur PSIPRED.

*Comparez les prédictions obtenues par rapport aux structures secondaires de la protéine résolue par cristallographie aux rayons X. Quel phénomène pouvez-vous observer ? Quelles hypothèses pouvez-vous formuler pour l'expliquer ?*

### 3.2 Etude de la structure tridimensionnelle

A l'aide de l'outil de visualisation de votre choix, observez la structure de différentes protéines appartenant à la famille des lipocalines.

*Quel motif structural les définit ?*

En reprenant l'alignement réalisé en début d'étude, représentez sur la structure 1GX8 les résidus conservés.

*Que pouvez-vous remarquer ? Le résidu Glu89 semble jouer un rôle déterminant dans la l'ouverture du site de liaison, où est-il situé et comment est orientée sa chaîne latérale ? Visualiser les ponts disulfures.*

A l'aide de l'outils CE SERVER, réaliser un alignement structural entre deux structures de $\beta$-lactoglobulines (2BLG et 3BLG).

*Quelle est la différence entre ces deux structures ? La poche est-elle bien conservée ?*

Représenter sur la structure tridimensionnelle les motifs mis en évidence précedemment ainsi que le motif PROSITE de la famille des lipocalines.

*Conclure quant à la fonction de la protéine.*

[2]code PDB : 1BEB pour le dimère ; 1GX8 en présence de rétinol

# 4 Programmation Python

L'objectif de cet exercice est de transformer un fichier au format *genbank* en un fichier au format *fasta*.

La première étape consiste à récupérer sur le site `http://www.ncbi.nlm.nih.gov/Genbank/`, le fichier *genbank* dont le numéro d'accession est `AF107201`.

*A quel gène ce fichier fait-il référence, chez quel organisme ? Que signifie la mention "complete cds" ?*

Un fichier *fasta* répond à une syntaxe précise. La séquence nucléique (habituellement 60 bases par lignes) est précédée d'une première ligne de commentaires, généralement sous la forme :

```
> définition | organisme | type de génome | numéro d'accession
```

*A l'aide des éléments vus en TD et des différents outils qui vous serons communiqués durant la séance de TP, écrivez un programme capable de transformer le fichier `AF107201.genbank` au format fasta.*