Sujet Numéro 1 : Analyse d'une séquence nucléique (programmation)

Ces dix dernières années, on a assisté au séquençage des génomes entiers, la quantité de données que cela représente est très importante. Par exemple, le génome humain représente à lui seul 3 milliards de paires de bases. Actuellement, 667 génomes sont complètement séquencés (http://www.genomesonline.org/) et plus de 3000 génomes sont en cours de séquençage. Il est donc nécessaire de mettre en place des algorithmes reposant sur des modèles mathématiques afin d'aider au traitement et à l'analyse des données génomiques.

Ce projet de stage a pour objectif de mettre en place un logiciel de prédiction de gènes le long d'une séquence nucléique à partir d'une modélisation markovienne. Au cours de ce projet, nous considèrerons, de manière simpliste, qu'une séquence d'ADN est composée uniquement de gènes et de régions intergéniques (situées entre les gènes). De plus, schématiquement, le gène en lui-même sera considéré comme constitué d'une partie codante appelée exon et d'une partie non codante appelée intron.

À l'adresse suivante (http://www-fourier.ujf-grenoble.fr/~dpiau/magist07w.html) vous pourrez télécharger les séquences pour lesquelles on vous demande de prédire la présence ou non de gènes : Sequence_1_tetraodon.fa et Sequence_2_tetraodon.fa. Ce sont des séquences nucléiques du génome du *Tetraodon nigroviridis* [1] qui est un poisson vivant dans les mers chaudes d'Asie du sud-est. Son génome représente environ 400 millions paires de bases.

I Modélisation formelle de la séquence

Avant de commencer la construction d'un modèle HMM modélisation la séquence d'ADN nucléique du génome du *Tetraodon nigroviridis* commenter l'utilité d'étudier un tel organisme.

Construisez un modèle capable de décrire la structure d'une séquence nucléique (comme décrite ci-dessus).
Dans un premier temps, vous vous intéresserez à un modèle HMM ou au sein de chaque état, vous aurez affaire à un modèle de type M0.

- *Dessinez le modèle formel sous la forme d'un graphe représentant les changements d'états.*
- *Notez les paramètres de transition et les probabilités d'émission correspondantes.*

Leurs valeurs seront estimées dans la deuxième partie.

II Estimation des paramètres

L'estimation au Maximum de Vraisemblance nécessite le compte des mots de taille m+1 dans les séquences d'apprentissage. Pour la suite du problème, vous pouvez travailler sur les trois fichiers d'apprentissage présents à l'adresse suivante (http://www-fourier.ujf-grenoble.fr/~dpiau/magist07w.html) qui contiennent respectivement des séquences d'exons, d'introns et de régions intergéniques au format fasta (exon.fa, intron.fa et intergenique.fa). Un fichier au format fasta est un simple fichier **texte** de la forme suivante :

>nom sequence 1
AACCCTGGGC
>nom sequence 2
TTTTTTTTGCCCCGGTA
….

a) **Probabilités d'émission :** Effectuez l'apprentissage de vos paramètres d'émission pour chacune des régions. Pour cela, créer un programme ***Compte*** qui permet de compter les mots dans les séquences génomiques contenues dans un fichier au format fasta.
b) **Probabilités de transition** : Les probabilités de transitions entre états sont calculées de la même manière : pour une transition à partir d'un état A vers un autre état (ou vers lui-même) on compte le nombre de passages entre ces deux états sur des séquences déjà segmentées, et l'on divise ce nombre par le nombre total de transitions effectuées à partir de cet état A. Dans l'exemple étudié, ces facteurs peuvent être déterminés grâce aux longueurs moyennes des exons, introns et régions intergéniques ainsi que part le nombre moyen d'exons contenus dans un gène (voir tableau ci-dessous) :

	Introns	Exons	Intergéniques	Nombre d'exons par gène
Tetraodon (en bp)	200	150	600	4

Déterminer les probabilités de transitions entre états.

III Chemin optimal et discussion biologique

3.1 Programmer l'algorithme de Viterbi [2], qui est un algorithme de segmentation de séquences à partir d'un modèle HMM. Faites attention, les séquences étudiées peuvent être très grandes.

3.2 En utilisant l'algorithme de Viterbi, déterminez quelle est la structure des séquences tests (Sequence_1_tetraodon.fa, Sequence_2_tetraodon.fa) la plus probable étant donné le modèle estimé lors de la partie II.

3.3 Recommencer en utilisant un modèle HMM ou au sein de chaque état, vous aurez affaire à un modèle de type M3 afin de prendre en compte la structure en codon des exons. En effet, les exons sont constitués de triplet de nucléotides (mots de 3 lettres) qui vont coder pour des acides aminés qui constitueront ensuite la protéine.

3.4 Recommencer en utilisant un modèle HMM ou au sein de chaque état, vous aurez affaire à un modèle de type M5 afin de prendre en compte la structure en codon des exons, mais également la liaison entre deux codons (mots de 6 lettres).

3.5 Conclusion : Comparer entre eux les résultats obtenus lors des questions 3.2 à 3.4. Qu'en concluez vous ?
Vous pourrez également discuter de la pertinence de la base de données choisie, et du problème des séquences d'apprentissage en général.

IV Comparaison avec le logiciel Genscan (ref biblio)

a) À partir de l'article suivant [3] décrivez le type de modèle qui est utilisé dans l'algorithme Genscan. Préciser pour quels types d'espèces cet algorithme est optimal et pourquoi ?

b) Rechercher si vos deux séquences tests contiennent des régions codantes en utilisant le logiciel Genscan (http://genes.mit.edu/GENSCAN.html). Vous pouvez visualiser la structure du gène prédit au format pdf et enregistrer ce schéma. Notez bien qu'il s'agit d'une prédiction : la probabilité que les exons prédits soient réellement des exons vous donne une indication de la qualité de cette prédiction (noter ces valeurs de probabilité pour chaque exons prédit). Cette prédiction est faite avec une certaine sensibilité (capacité à prédire tous les exons) et une certaine spécificité (capacité à discriminer les vrais exons des faux exons) : vous pouvez trouver ces valeurs de sensibilité dans la documentation de Genscan.

c) Comparer les résultats obtenus par votre programme avec ceux de Genscan et ainsi qu'avec les annotations contenues dans les fichiers AF007218.annot et AJ251458.annot extrait de la banque de données HOVERGEN, qui contient respectivement les informations concernant Sequence_1_tetraodon.fa et Sequence_2_tetraodon.fa.

Qu'en concluez vous ? Quelles améliorations peuvent être apportées à votre modèle ?
Vous pourrez également discuter de la pertinence de la base de données choisie, et du problème des séquences d'apprentissage en général.

Référence bibliographique :

[1] Jaillon O *et al.* (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertabrate proto-karyotype. *Nature*, 431, 946-957.

[2] Rabiner L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, 2.

[3] Burge C, Karlin S. (1997) Prediction of complete gene structure in human genomic DNA. *Journal of Molecular Biology*, 268, 78-94.