Magistère L2 **Université Joseph Fourier**
Proposition de TP **Année 2007-2008**

Module « Analyse statistique de séquences biologiques »

Sujet n° 3 : Modèles markoviens cachés (Mathématiques)

Le but du TP est d'étudier certaines propriétés statistiques des modèles de Markov cachés et la robustesse des procédures d'estimation de ces modèles.

Préliminaire

Simuler un échantillon $\mathcal{E}$ de $n = 10^3$ nucléotides indépendants avec la distribution uniforme sur les nucléotides `a`, `c`, `g`, `t`.
On associe à $\mathcal{E}$ une séquence définie par $Z_k = 1$ si les nucléotides de numéros k et $k+1$ sont tous les deux `a`, et $Z_k = 0$ sinon.

- Calculer la proportion de 1 et la proportion de 0 parmi les $n - 1 = 999$ valeurs Z_k associées à l'échantillon $\mathcal{E}$.
- Calculer les proportions des quatre couples possibles de 1 et de 0 parmi les $n - 2 = 998$ valeurs (Z_k, Z_{k+1}) associées à l'échantillon $\mathcal{E}$.
- Déterminer si la règle d'indépendance de Z_k et Z_{k+1} est vérifiée. On explicitera les calculs utilisés.
- Utiliser les règles de calcul des probabilités conditionnelles pour calculer les fréquences théoriques des quatre couples possibles de 1 et de 0.
- Vérifier ce dernier résultat sur un échantillon $\mathcal{E}'$ de longueur $n = 10^5$.
- (Facultatif) Donner un argument simple montrant que $(Z_k)_k$ ne peut être décrit par un modèle Mm pour aucun ordre $m \geqslant 0$. Décrire $(Z_k)_k$ par un modèle M1M0.

Partie 1

L'organisme *Chimera feerica* est un procaryote. Les régions intergéniques (état `I`) suivent un modèle M0 dans lequel chaque nucléotide est présent à 25%. Les gènes (état `G`) suivent un modèle M0 dans lequel le nucléotide `c` est présent à 50% et les trois autres nucléotides sont équiprobables. Une séquence du génome de *Chimera feerica* est dans le fichier joint `chimera.txt`.

- Utiliser l'algorithme de Viterbi pour segmenter cette séquence selon un modèle M1M0 en supposant que la probabilité de transition de `I` vers `G` vaut $\frac{1}{40}$ et la probabilité de transition de `G` vers `I` vaut $\frac{1}{20}$. Décrire le programme utilisé et représenter par un graphe le résultat obtenu.
- Recommencer le même travail de segmentation par l'algorithme de Viterbi avec d'autres probabilités de transition de `I` vers `G` et de `G` vers `I`. Représenter par des graphes les résultats obtenus. Commenter les résultats.

Partie 2

Choisir deux distributions $\mu_{\mathtt{I}}$ et $\mu_{\mathtt{G}}$ sur l'espace des nucléotides, associées respectivement aux états `I` et `G`. Choisir une probabilité de sortie $q_{\mathtt{I}}$ de l'état `I` et une probabilité de sortie $q_{\mathtt{G}}$ de l'état `G`. Simuler une séquence d'états de longueur $n = 10^5$ décrite par ces probabilités de sortie $q_{\mathtt{I}}$ et $q_{\mathtt{G}}$, puis une séquence $\mathcal{S}$ d'observations émises par ces états selon les distributions $\mu_{\mathtt{I}}$ et $\mu_{\mathtt{G}}$.
On prendra soin de choisir des distributions $\mu_{\mathtt{I}}$ et $\mu_{\mathtt{G}}$ assez différentes et des probabilités de sortie $q_{\mathtt{I}}$ et $q_{\mathtt{G}}$ entre 10^{-2} et 10^{-4}.

- Utiliser l'algorithme de Viterbi pour segmenter $\mathcal{S}$ selon un modèle M1M0 pour différentes valeurs des probabilités de transition de `I` vers `G` et de `G` vers `I`. Décrire le programme utilisé. Représenter les taux d'erreur et les localisations des erreurs le long de la séquence $\mathcal{S}$ dans les segmentations prédites.

Commenter les résultats obtenus. On précisera en particulier les probabilités de transition donnant les meilleures prédictions et on tentera d'expliquer qualitatitivement l'apparition des erreurs de prédiction.

• Segmenter $\mathcal{S}$ en utilisant l'algorithme de Baum-Welch. Décrire le programme utilisé. Représenter en particulier l'évolution de la vraisemblance de $\mathcal{S}$ selon les modèles successifs proposés par l'algorithme et la segmentation finalement déduite de l'algorithme de Baum-Welch, pour chaque point de départ. On précisera la procédure de choix des différents points de départ de l'algorithme utilisés, on comparera entre eux les différents résultats obtenus et on les commentera. On proposera enfin une segmentation unique en précisant comment on l'a déduite des résultats précédents (plusieurs procédures sont possibles).

• (Facultatif) Expliquer de manière concise pourquoi l'algorithme de Baum-Welch peut converger vers plusieurs modèles différents selon les conditions initiales. Faire une recherche bibliographique sur cet aspect de l'algorithme en utilisant les ressources du web et décrire brièvement des solutions qui peuvent être employées pour remédier à ce problème.

Références

– Notes de cours.
– Sean Eddy. What is a hidden Markov model ? *Nature Biotechnology* 22, 1315-1316, 2004.
– Richard Durbin, Sean Eddy, Anders Krogh, Graeme Mitchison. *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.