TP analyse des séquences biologiques Utilisation d'une librairie C++ pour estimer les paramètres de modèles de Markov.

Franck Picard

L'objectif de ce TP est de déterminer quel serait le meilleur modèle markovien décrivant au mieux une séquence d'ADN. Pour cela, on utilise `seq++` une libraire `C++` destinée spécialement à l'estimation des modèles markoviens pour les séquences biologiques. Cette librairie peut être téléchargée à l'adresse suivante `http://stat.genopole.cnrs.fr/seqpp/index.html`.

1 Installation de la librarie

Téléchargez et installez la librairie `seq++`. On peut se référer aux instructions données en ligne `http://stat.genopole.cnrs.fr/seqpp/inst.html`

Vérifications. l'installation de `seq++` nécessite les compilateurs C++, F77, ainsi que la librarie GSL.

Installation.

```
./configure
make
make install
make docs (if you want to generate the html doc)
```

2 Etude de la séquence du génome d'HIV

- Télécharger la séquence du génome du virus de l'immunodéficience humain sur le site du NCBI `http://www.ncbi.nlm.nih.gov/genome`. Télécharger le fichier au format FASTA (`send`,`complete record`, `file`, `FASTA`), et le sauver sous le nom `HIV.FASTA`.

- Estimer un modèle markovien d'ordre 1 pour cette séquence `estim_m -d 1 -o hiv-m1.txt hiv.fasta`
- Etudier la structure du fichier de sortie `hiv-m1.txt`. On trouve l'ordre d'enchainement des lettres avec la commande `estim_m -h` (`--dna Use DNA alphabet 1:AGCT, default setting`).
- Estimer un modèle markovien d'ordre 2 jusqu'à l'ordre 5 pour la séquence de HIV (`hiv-m2.txt`...`hiv-m5.txt`). Les matrices de transition sont structurées comme suit (pour l'ordre 2 par exemple):

	A	G	C	T
AA				
AG				
AC				
AT				
GA				
GG				
GC				
GT				
...				

- Obtenir les valeur des log-vraisemblances pour les modèles markoviens d'ordre 1 à 5 à l'aide de `seq++`. Tracer la log-vraisemblance en fonction de l'ordre du modèle. Rappel:

$$\log \mathcal{L}_{s,m}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\pi}}) \quad = \quad \log \widehat{\boldsymbol{\mu}}(x_1, ..., x_m) + \sum_{x_1,...,x_{m+1}} N(x_1, ..., x_{m+1}) \log \widehat{\pi}(x_1, ..., x_{m+1})$$

- Tracer cette log-vraisemblance en fonction de l'ordre du modèle, et calculer le critère BIC pour chaque modèle. On rappelle que la taille d'un modèle Markovien d'ordre m est :$(|\mathcal{A}| - 1) \times |\mathcal{A}|^m$. Recalculer le BIC à partir des vraisemblances données par `seq++`. NB: ici le BIC calculé s'écrit:

$$BIC\left(\mathcal{M}_\theta\right) = -2 \log \mathcal{L}\left(\mathcal{M}_\theta\right) + \log(n) \times |\mathcal{M}_\theta|$$

- Choisissez le meilleur modèle au sens du BIC pour décrire la séquence du génome d'HIV.
- Procédez de même avec la séquence de la bactérie E. Coli (K12). Quel modèle choisiriez vous dans ce cas ?