IFT3290 H05 — Devoir 2

Miklós Csűrös

23 février 2006

À remettre le 15 mars.

1 Segmentation par HMM (100 points)

Le but de ce travail pratique est de développer un outil pour diviser une séquence d'ADN en segments selon la fréquence de nucléotides. On utilisera un HMM à deux états q_{AT} et q_{GC}. L'état q_{AT} produit des segments riches en $\mathtt{A}+\mathtt{T}$ et l'état q_{GC} produit des segments riches en $\mathtt{G}+\mathtt{C}$. En conséquence, les probabilités d'émission de `A` et `T` sont plus grandes en q_{AT} qu'en q_{GC}. (L'idée de ce TP vient de Sean Eddy.)

Implantation de l'HMM

Implantez ce modèle de Markov caché dans une classe de Java. (Si vous préférez un autre langage de programmation, consultez avec moi.) Les paramètres du modèle sont les probabilités de transition entre les deux états $\tau(q_{\mathrm{AT}}, q_{\mathrm{GC}})$ et $\tau(q_{\mathrm{GC}}, q_{\mathrm{AT}})$, et les probabilités d'émission $p(q_{\mathrm{AT}}, \mathtt{A})$ et $p(q_{\mathrm{GC}}, \mathtt{A})$. Les autres probabilités sont calculées à partir de ces valeurs :

$$\tau(q_{\mathrm{AT}}, q_{\mathrm{AT}}) = 1 - \tau(q_{\mathrm{AT}}, q_{\mathrm{GC}}) \qquad \tau(q_{\mathrm{GC}}, q_{\mathrm{GC}}) = 1 - \tau(q_{\mathrm{GC}}, q_{\mathrm{AT}})$$
$$p(q_{\mathrm{AT}}, \mathtt{T}) = p(q_{\mathrm{AT}}, \mathtt{A}) \qquad p(q_{\mathrm{AT}}, \mathtt{G}) = p(q_{\mathrm{AT}}, \mathtt{C}) = \frac{1}{2} - p(q_{\mathrm{AT}}, \mathtt{A})$$
$$p(q_{\mathrm{GC}}, \mathtt{T}) = p(q_{\mathrm{GC}}, \mathtt{A}) \qquad p(q_{\mathrm{GC}}, \mathtt{G}) = p(q_{\mathrm{GC}}, \mathtt{C}) = \frac{1}{2} - p(q_{\mathrm{GC}}, \mathtt{A}).$$

Le HMM démarre toujours en état q_{AT}.

Algorithme de Viterbi

Implantez l'algorithme de Viterbi. La méthode [dans la classe de l'HMM] en question a la signature

```
public int[] cheminViterbi(String sequence);
```

La vecteur v retournée contient des 0s et des 1s : $v[i] = 0$ si le chemin passe par état q_{AT} pour générer le i-ème caractère ($i = 0, 1, \ldots,$ `sequence.length()` $- 1$). Travaillez avec les logarithmes des probabilités pour éviter l'underflow. (La longueur de la séquence peut être dans les millions.) La séquence contient les caractères `A`, `C`, `G`, `T` (tous majuscules) ; autres caractères sont à ignorer. Plus précisement, calculez les récurrences avec $p(q, \sigma) = 1$ quand σ est un caractère différent des quatre nucléotides.

Implantez un algorithme d'apprentissage en utilisant la méthode de pseudo-compteurs Laplace. La méthode `apprentissage` [dans la classe de l'HMM] en question prend deux arguments, `int[] chemin` et `String seq`. Elle compte les fréquences des différentes transitions et émissions et recalcule les probabilités de transition et émission.

L'outil

En combinant les classes pour le HMM et la lecture du fichier Fasta du Devoir 1, implantez une classe executable. La méthode `main` prend cinq arguments : fichier, $\tau(q_{\mathrm{AT}}, q_{\mathrm{GC}})$, $p(q_{\mathrm{AT}}, \mathtt{A})$, $\tau(q_{\mathrm{GC}}, q_{\mathrm{AT}})$, et $p(q_{\mathrm{GC}}, \mathtt{A})$ [dans cet ordre]. Après avoir lu le fichier, et initialisé le HMM, le chemin Viterbi est calculé, et les paramètres de l'HMM sont optimisés.

```
...
SimpleHMM M=new SimpleHMM(t01,p0A,t10,p1A);
int[] chemin=M.cheminViterbi(sequence);
for (int rep=0; rep<5; rep=++){ // répéter 5 à 10 fois
    M.aprentissage(chemin, sequence);
    chemin=M.cheminViterbi(sequence);
}
... // (afficher chemin)
```

Le logiciel doit afficher (sur `System.stdout`) la liste des segments où le chemin reste dans l'état q_{GC} pour au moins 50 positions consécutives. Le format de la liste est illustré ici (avec l'exemple d'un autre génome).

```
358764..358940  177     61.0
359972..360044  73      65.7
402967..403055  89      62.9
```

La première colonne donne le début et la fin (séparés par `..`) de chaque segment, la deuxième colonne donne sa longueur, et la troisième colonne donne le pourcentage de $(\mathtt{G} + \mathtt{C})$ dans le segment (arrondi jusqu'à 0.1).

Recherche de gènes ARN

L'organisme *Nanoarchaeum equitans* possède le plus petit génome jamais séquencé (en excluant les viruses). Il est une archéobactérie hyperthérmophyle. Téléchargez la séquence de son génome (nombre d'accession `NC_005213` à GenBank), dans deux formats : Fasta (`.fna`) et GenBank (`.gbk`). En utilisant les paramètres $\tau(q_{\mathrm{AT}}, q_{\mathrm{GC}}) = 0.001$, $p(q_{\mathrm{AT}}, \mathtt{A}) = 0.3$, $\tau(q_{\mathrm{GC}}, q_{\mathrm{AT}}) = 0.01$, et $p(q_{\mathrm{GC}}, \mathtt{A}) = 0.2$, compilez la liste de segments riche en $(\mathtt{G} + \mathtt{C})$. Comparez cette liste à la liste de gènes non-codant dans le génome.

Il y a une forte correlation. En fait, Klein et al. (2002) ont utilisé un HMM de deux états pour identifier des gènes ARN dans le génome du *Methanocaldococcus jannaschii*. (Lisez l'article.)

Vous avez besoin de la liste des gènes ARN dans le génome. Le logiciel `tRNAscan-SE` les trouve pour vous (`http://www.genetics.wustl.edu/eddy/tRNAscan-SE/`) : installez une copie et annotez le génome ou utilisez la liste que j'ai préparée dans le fichier `NC_005213-tRNA-SE.txt`.

Identifiez les tARNs trouvés par segmentation. Compilez un tableau qui montre la correspondance entre les segments de votre logiciel et les gènes ARN selon l'annotation. Pour chaque segment qui se chevauche avec un gène ARN, donnez le nom du gène et le début et la longueur du gène. Essayez d'identifier les segments riches en `G` + `C` qui ne correspondent pas à des ARNt en se servant de l'annotation dans le fichier `.gbk` et/ou de la recherche de similarités par BLASTN (n'oubliez pas de séléctionner «Archaea» parmi les organismes pour limiter la recherche).

Remise de travail

Il faut remettre votre code, la liste de segments trouvés par le logiciel dans le génome de *N. equitans*, ainsi que le tableau de comparaisons (dans un fichier) entre les gènes et les segments, et vos théories concernant la reste des segments.

Références

Klein, R. J., Z. Misulovin, and S. R. Eddy (2002). Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proceedings of the National Academy of Sciences of the USA 99*, 7542–7547.