TP3. Modèles de Markov cachés de type profils (Profile hidden Markov Models)

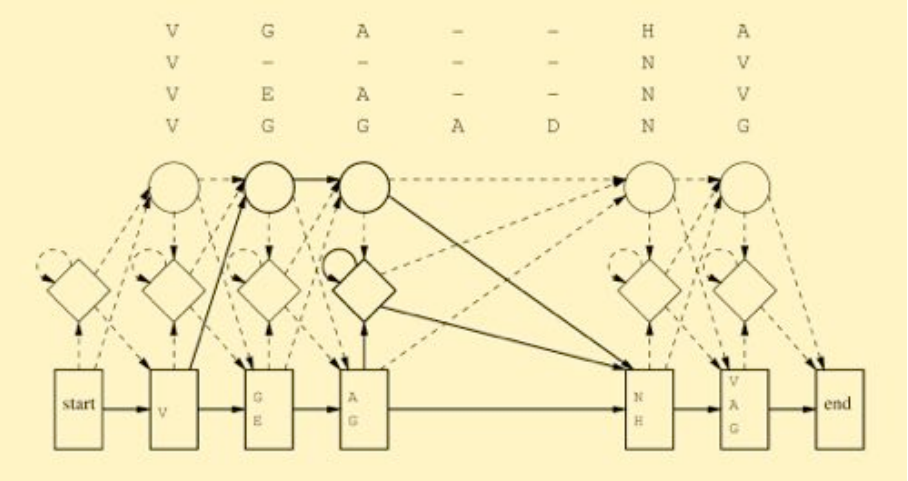
Nicolas Lartillot
nicolas.lartillot@univ-lyon1.fr

25 janvier 2017

L'objectif de ce TP est de découvrir le concept de Modèles de Markov cachés dit de profils (Profile hidden Markov Models), tels qu'implémentés par la suite de programme `HMMER`. Les modèles de profils cherchent à modéliser les domaines protéiques conservés. Ils sont utilisés pour identifier des homologues et pour effectuer des alignements multiples de domaines.

La démarche générale est la suivante. À partir d'un alignement de séquences, correspondant à un domaine protéique déjà bien identifié, on estime une HMM qui capture les conservations positionnelles en acides-aminés, ainsi que les patterns d'insertions et de délétions, qui caractérisent le domaine. Ce modèle est ensuite utilisé pour scanner des bases de données de séquences protéiques, afin d'identifier les occurrences du domaine protéique en question (sur la base de la probabilité marginale $p(x \mid M)$ que la séquence x soit une instance du domaine protéique). Enfin, les homologues putatifs sont rassemblés puis alignés entre eux, en utilisant les probabilités a posteriori fournies par le modèle (décodage a posteriori).

La structure typique d'un modèle à profil est représenté ci-dessous :



Vous sont donnés pour ce TP les fichiers suivants :

— `globins4.ali` : un petit alignement de 4 séquences de globines de vertébrés.

— `uniprothuman.dat` : un fichier contenant l'ensemble des séquences protéiques humaines, tirées de la base UniProt (`http://www.uniprot.org`).
— `uniprotCHICK.dat`, `uniprotDANRE.dat`, `uniprotPETMA.dat` : contenant respectivement les séquences du poulet (*Gallus gallus*), du poisson-zèbre (*Danio rerio*) et de la lamproie (*Petromyzon marinus*).
— `grepseq.pl` : un script en Perl permettant de récupérer des séquences protéiques des bases de données de type UniProt, à partir des accesseurs de ces séquences.
— `grepseq.pl` : un script en Perl permettant de récupérer des séquences protéiques des bases de données de type UniProt, à partir des accesseurs de ces séquences.
— `Pfamsmall.hmm` : un sous-ensemble de la base Pfam ((`http://pfam.xfam.org`))

Les globines de vertébrés

Le fichier `globins4.ali` contient un alignement multiple de quatre séquences de globines de vertébrés. Il est possible de visualiser cet alignement avec `seaview`, afin d'observer plus particulièrement les conservations positionnelles, ainsi que les occurrences d'insertion et de délétions dans certaines séquences de l'alignement.

Question 1 *À partir de cet alignement, construire un modèle de profil.*

La construction d'un hmm se fait en utilisant la commande suivante :

```
hmmbuild globins4.hmm globins4.ali
```

À l'issue de cette commande, un fichier `globins4.hmm` est produit, qui contient la spécification du modèle de profil appris sur l'alignement.

Question 2 *Visualiser le contenu du fichier `globins4.hmm`. Combien de position* match *contient le modèle de profil ? Combien de positions alignées y avait-il dans l'alignement de départ ?*

À noter que, afin d'avoir une sortie plus compacte du profil de conservation, il est également possible d'utiliser la commande de visualisation suivante :

```
hmmlogo globins4.hmm
```

Il est également possible de visualiser le logo sur la page web suivante : `http://skylign.org`.

Une fois qu'un modèle de profil a été construit, il est possible de l'utiliser afin de rechercher des homologues du domaine d'intérêt dans une base de données de séquences protéiques. Ceci se fait au moyen de la commande `hmmsearch`. Par exemple, si l'on veut chercher les domaines de globine dans les séquences protéiques humaines :

```
hmmsearch globins4.hmm uniprothuman.dat
```

Il est souvent plus pratique de stocker la sortie (assez abondante) de `hmmsearch` dans un fichier, par redirection :

```
hmmsearch globins4.hmm uniprothuman.dat > globinshuman.out
```

Question 3 *Examinez le fichier de sortie produite par* `hmmsearch` *Combien d'homologues de la globine trouve-t-on dans la base de séquences protéiques humaine ?*

Question 4 *Récupérer l'ensemble des séquences humaines pour lesquelles un domaine globine a été identifié, et sauvez-les dans un fichier séparé,* `globinshuman.dat`*.*

Pour effectuer cette sélection, un script `perl` vous est fourni, appelé `grepseq.pl`. La commande est la suivante :

```
perl grepseq.pl <database> <selection> <outfile>
```

Ici, `<database>` est le nom de la base de données de séquences dans laquelle effectuer la sélection (par exempe, `uniprothuman.dat`), `<selection>` est un fichier qui doit contenir la liste des identifiants des séquences à récupérer, et `<name>` est le nom du fichier dans lequel on veut écrire les séquences sélectionnées.

Question 5 *Parmi les séquences trouvées par* `hmmsearch`*, récupérer plus particulièrement la séquence protéique de la cytoglobine (*`CYGB_HUMAN`*), et conservez-là dans un fichier séparé (appelons ce fichier* `cygbhuman.dat`*)*

Le modèle de profil ne sert pas uniquement à détecter des homologues. Il peut également être utilisé pour aligner une séquence jugée homologue avec le profil. La commande est la suivante :

```
hmmalign globins4.hmm cygbhuman.dat
```

Observez la sortie de `hmmalign`, qui indique les positions matches et indels, ainsi que les scores de probabilité a posteriori. À noter qu'il est possible d'utiliser `hmmalign` directement sur un fichier contenant plusieurs séquences. Par exemple, si l'on veut aligner lensemble des séquences humaines récupérées précédemment :

```
hmmalign globins4.hmm globinshuman.dat
```

Dans un deuxième temps, il peut être plus pratique de ne récupérer que les positions 'match', et de sortir l'alignement en format `PSIBLAST` :

```
hmmalign --trim --outformat PSIBLAST globins4.hmm globinshuman.dat
```

Cet alignement peut-être sauvé dans un fichier :

```
hmmalign --trim --outformat PSIBLAST globins4.hmm globinshuman.dat > globinshuman.psi
```

puis traduit en format FASTA au moyen d'un petit script perl appelé `psi2fasta.pl`

```
perl psi2fasta.pl globinshuman.psi globinshuman.fasta
```

et enfin, visualié avec `seaview`

```
seaview globinshuman.fasta &
```

Cet alignement est 'propre' (uniquement les positions 'match'), encore que, idéalement, il faudrait ne récupérer que les positions qui matchent avec une probabilité a posteriori élevée (> 0.95, celles qui sont marquées d'une étoile dans la sortie par défaut). Il serait possible d'écrire un petit script en perl ou en python qui récupère uniquement ces positions jugées les plus fiables.

Question 6 *Identifier les homologues de globine chez les autres espèces de vertébrés suivantes : le poulet (Gallus), le poisson zèbre (Danio rerio), la lamproie (Petromyzon marinus). Rassembler toutes ces séquences dans un unique fichier, avec les séquences humaines, les aligner, récupérer les positions 'match', puis reconstruire et visualiser la phylogénie des globines pour ces 4 espèces.*

La base de données `Pfam`

`Pfam` (`http://pfam.xfam.org`) est une base de données qui enregistre une large collection de domaines protéiques. Pour chaque domaine, `Pfam` fournit un modèle de profil, qui a été construit en utilisant le logiciel `HMMER`.

Un sous-ensemble de la base `Pfam` vous est fourni dans le dossier du TP, dans le fichier `Pfamsmall.hmm` (l'ensemble de la base peut facilement être téléchargé par ftp à partir du site web indiqué ci-dessus). Vous pouvez examiner le début du fichier `Pfamsmall.hmm` et observer que celui-ci consiste simplement en la concaténation de HMMs telles que celle construite précédemment pour la globine, en utilisant le programme `hmmbuild`.

Il est possible d'extraire de cette base de HMMs le modèle de profil correspondant à un domaine particulier, au moyen de la command `hmmfetch`. Par exemple, on peut trouver un modèle de profil correspondant au domaine de la globine :

```
hmmfetch Pfamsmall.hmm Globin
```

ou, avec redirection :

```
hmmfetch Pfamsmall.hmm Globin > globinpfam.hmm
```

À la différence du modèle que l'on a construit dans la section précédente (à partir d'un échantillon de 4 séquences), ce modèle a été appris sur un jeu plus large de séquences.

Question 7 *Effectuer une recherche des homologues de la globine chez l'humain en utilisant le modèle de globine fourni par `Pfam`. Comparer avec les résultats obtenus dans la section précédente*

Les protéines à doigt de zinc

Les protéines à doigt de zinc sont une famille très large de facteurs de transcription (essentiellement des répresseurs) qui s'est beaucoup diversifiée chez les eukaryotes, et plus particulièrement chez les vertébrés. Ces protéines se fixent à l'ADN au moyen d'un nombre variable de doigts de zinc. Le doigt de zinc représente un motif très simple, de petite taille, qui est donc généralement présent en plusieurs copies dans un facteur de transcription. Dans la suite, on considère plus particulièrement la sous-famille des doigts de zinc de type C2H2 (2 cystéines + 2 histidines). Cette famille est discutée dans l'article de revue de Tadepally et al, 2008, dont le pdf est fourni dans le dossier du TP. L'objectif de cette partie est d'utiliser HMMER afin de caractériser la structure multi-domaines des protéines à doigt de zinc, à l'instar de ce qui est présenté dans le revue de Tadepally et al, 2008.

Question 8 *Récupérer le modèle de profil correspondant au motif de doigt de zinc C2H2 dans la sous-base de `Pfam`. Le nom de ce domaine dans `Pfam` est `zf-C2H2`*

Question 9 *Analyzer la base de données humaines. Combien trouve-t-on de protéines contenant au moins un doigt de zinc C2H2 ? Combien de doigts de zinc ces protéines ont-elles en moyenne ? Combien de protéines n'ont qu'un doigt de zinc, et combien en ont plus de 10 ? Tracer l'histogramme décrivant la distribution du nombre de doigts de zinc par protéine.*

Question 10 *Récupérer l'ensemble des séquences protéiques humaines possédant des doigts de zinc.*

L'article mentionne que la majorité des protéines à doigt de zinc humaines contiennent un domaine KRAB. Une plus petite proportion ont le domaine SCAN (voir figure 2A de l'article).

Question 11 *Récupérer les deux modèles de profil correspondant aux domaines KRAB et SCAN.*

Question 12 *Évaluer le nombre de protéines à doigts de zinc qui possèdent chacun de ces deux domaines.*

Le programme `hmmsearch` identifie toutes les occurrences d'un domaine particulier à travers une banque de séquences. À l'inverse, le programme `hmmscan` permet d'annoter une séquence protéique particulière, en identifiant l'ensemble des domaines qu'elle contient. Prenons un cas particulier, la séquence humaine `ZF69B_HUMAN`.

Question 13 *Extraire la séquence de `ZF69B_HUMAN` et la sauvegarder dans un fichier séparé, `zf69b.dat`, puis annoter cette séquence en identifiant l'ensemble de ses domaines conservés, contre la base de Pfam.*

Pour des raisons d'efficacité de recherche, `HMMER` requiert que l'on compresse la banque de modèles de profils contre laquelle on va analyser la séquence protéique d'intérêt :

```
hmmpress Pfamsmall.hmm
```

Ensuite, on peut scanner la protéine d'intérêt :

```
hmmscan Pfamsmall.hmm zf69b.dat.
```

Question 14 *Récupérer toutes les protéines humaines possédant le domaine SCAN. Effectuez l'alignement du domaine SCAN à travers toutes ces protéines. Bien observer l'alignement stockholm, puis récupérer la partie bien alignée dans un fichier séparé. Utiliser cet alignement pour construire la phylogénie des protéines SCAN-zfC2H2.*