3I005
Statistique en Bioinformatique : Analyse statistique d'une famille de protéines

L'objectif de ce projet est l'analyse statistique d'une famille de protéines donnée par un alignement de séquences:

- objectif 1: détection des positions conservées,
- objectif 2: détection de séquences qui appartiennent à la même famille,
- objectif 3: détection de corrélations entre colonnes différentes de l'alignement, et de leur rélation avec les distances entres acides aminées dans la structure 3D d'une protéine répresentative de la famille.

Le but génerale est "de faire parler" des séquences, ca veut dire d'extraire information statistique sur structure et fonction d'une famille de protéines homologues, en partant seulement des séquences.

A rendre (à ari.ugarte@gmail.com

- un rapport de 4 pages maximum, avec les réponses aux questions de l'énoncé et une guide d'utilisation de votre code,
- le code développé.

I. DONNÉES

Il y a 3 fichiers avec les données

- `Dtrain.txt`: C'est un alignement de $M = 5643$ protéines d'une seule famille en format FASTA, ex. :
  ```
  >IVBI5_BUNMU/30-82
  -CNLPPDPGPCHDNKFAFYHHPASNKCKEFVYGGCGGNDNRFKTRNKCQCTC-
  >C1IC53_WALAE/30-82
  -CHLPADPGPCSNYRPAYYYNPASRKCEEFMYGGCKGNKNNFKTRHECHRVCV
  >IVBI2_PSETT/30-82
  LCELPPDTGPCRVRFPSFYYNPDEQKCLEFIYGGCEGNANNFITKEECESTC-
  >IVBI1_OXYMI/30-82
  LCELPADTGPCRVGFPSFYYNPDEKKCLEFIYGGCEGNANNFITKEECESTC-
  ```
  Lignes qui commencent avec ">" contiennent des commentaires (nome de la protéine etc.). Elles n'ont aucune importance pour notre projet. Les autres lignes contiennent les séquences, que l'on va utiliser dans le projet. Les séquences sont alignées, elles ont toutes la même longueur ($L = 48$ positions dans notre fichier). Chaque position $i = 0, ..., L-1$ d'une séquence contient ou une acide aminé (A,C,...,Y, il y en a 20) ou un trou (–), que l'on considère comme 21ème lettre. Ensemble elles forment l'alphabet

  $$\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}, \mathtt{G}, \mathtt{H}, \mathtt{I}, \mathtt{K}, \mathtt{L}, \mathtt{M}, \mathtt{N}, \mathtt{P}, \mathtt{Q}, \mathtt{R}, \mathtt{S}, \mathtt{T}, \mathtt{V}, \mathtt{W}, \mathtt{Y}, -\}$$

  avec $q = 21$ lettres différentes.
- `testseq.txt`: Même format de `train.faa`, mais avec une seule séquence $b = (b_0, ..., b_{N-1})$ plus longue (longueur $N = 114 > L$). On va scanner cette séquence pour trouver une sous-séquence qui appartient à la famille definie par `Dtrain.txt`.
- `distances.txt`: Contient les distances entre paires d'acides aminées. Il y a trois colonnes: ID de la première position, ID de la deuxième position, distance (en Angstrom). ATTENTION: Les positions sont numerotées $i = 0, ..., L-1$.

II. MODÈLISATION PAR PSWM

A. Estimer une PSWM

On veut modèliser une famille de protéines, données par l'alignement (une matrice)

$$D_{train} = \begin{pmatrix} a_0^1 & a_2^1 & ... & a_{L-1}^1 \\ a_0^2 & a_2^2 & ... & a_{L-1}^2 \\ & & ... & \\ a_0^M & a_2^M & ... & a_{L-1}^M \end{pmatrix} .$$

Les entrées $a_i^m \in \mathcal{A}$ sont les acide aminées (ou un trou) en position $i \in \{0, ..., L-1\}$ de la séquence numero $m \in \{1, ..., M\}$, avec M = nombre total de séquences (notre cas: $M = 5643$). Les quantités importantes pour la modèlisation sont les

$$n_i(a) = \text{nombre d'occurences d'acide aminée } a \text{ en position (colonne) } i \ . \tag{1}$$

Ces nombres sont spécifiques pour chaque position (= chaque colonne de D_{train}). Ils satisfont $\sum_{a\in\mathcal{A}} n_i(a) = M$ pour chaque $i \in \{0, ..., L-1\}$.

Le modèle statistique est un modèle factorisé

$$P(a_0, ..., a_{L-1}|\omega) = \prod_{i=0}^{L-1} \omega_i(a_i) \ , \tag{2}$$

donc on suppose que les positions sont indépendantes. Les paramètres $\omega_i(a)$, qui forment une matrice $L \times q$, sont spécifiques pour la position i et l'acide aminée a. ω est appelé "position-specific weight matrix" (matrice de poids spécifiques des positions, PSWM ou PWM). Les poids sont calculé en utilisant les nombres d'occurence (et un pseudocount 1):

$$\omega_i(a) = \frac{n_i(a) + 1}{M + q} \ . \tag{3}$$

B. Conservation

On cherche des positions *conservées*, i.e. positions qui ont un poid très élévé pour une acide aminée. En géneral, des positions conservées sont biologiquement importantes, une mutations d'une telle position souvent interrompt le fonctionnement de la protéine.

Pour trouver ces positions, on calcule l'entropie rélative (aussi : information (en bit) sur l'identité de l'acide aminée donnée par la connaissance de la position) :

$$S_i = \log_2(q) + \sum_{a\in\mathcal{A}} \omega_i(a) \cdot \log_2 \left[\omega_i(a)\right] \ . \tag{4}$$

Elle prends une valeur entre zero (aucune information, donc position ne pas conservée) et $\log 2(21) \simeq 4.39$ (information maximale, donc acide aminée complètement conservée). Donc les positions conservées vont avoir des valeurs assez grandes de S_i. L'acide aminée plus probable dans une telle position est simplement donnée par la PSWM :

$$a_i^\star = \text{argmax}_{a\in\mathcal{A}} \omega_i(a) \ . \tag{5}$$

On va déterminer les positions plus conservées avec l'acide aminée présente.

C. Evaluer une nouvelle séquence

Comment est-ce qu'on peut décider si une nouvelle séquence $b = (b_0, ..., b_{L-1})$ (qui typiquement n'est pas contenue dans l'alignement) fait partie de la même famille de protéines? Selon équation (2) sa probabilité est

$$P(b_0, ..., b_{L-1}|\omega) = \prod_{i=0}^{L-1} \omega_i(b_i) \ . \tag{6}$$

Pour décider si cette probabilité est "assez grande" pour dire que b appartient à la famille donné par D_{train}, il faut la comparer avec un *modèle nul* qui n'est pas spécifique dans les positions:

$$P^{(0)}(b_0, ..., b_{L-1}) = \prod_{i=0}^{L-1} f^{(0)}(b_i) \tag{7}$$

avec

$$f^{(0)}(b) = \frac{1}{L} \sum_{i=0}^{L-1} \omega_i(b) \ , \tag{8}$$

i.e. $f^{(0)}(b)$ correspond à la fréquence de chaque $b \in \mathcal{A}$ dans l'alignement D_{train} entier (sans régarder la position i).

Pour comparer le modèle specifique (PSWM) avec le modèle nul, on introduit la *log-vraisemblance* (aussi "log-odds ratio")

$$\ell(b_0, ..., b_{L-1}) = \log_2 \frac{P(b_0, ..., b_{L-1}|\omega)}{P^{(0)}(b_0, ..., b_{L-1})} = \sum_{i=0}^{L-1} \log_2 \frac{\omega_i(b_i)}{f^{(0)}(b_i)} \ . \tag{9}$$

Pour $\ell(b_0, ..., b_{L-1}) > 0$, la séquence b est plus probable dans le modèle spécifique, donc on peut supposer qu'elle appartient à la même famille. Pour $\ell(b_0, ..., b_{L-1}) < 0$, la séquence b est plus probable dans le modèle nul. C'est une indication que b n'est pas de la famille donnée par D_{train}

Pour la séquence longue dans les fichier `testseq.txt`, il faut donc prendre chaque sous-séquence (positions consécutives) de longueur L (fenêtre glissante - sliding window), et déterminer la log-vraisemblance, pour trouver des sous-séquences qui appartiennent à la famille donnée par D_{train}.

D. Réalisation

- Première fonction: Pour chaque position (colonne) $i = 0, ..., L-1$ et chaque acide aminée $a \in \mathcal{A}$ (le trou compris), calculer le nombre d'occurence $n_i(a)$ (équation (1)) et le poid $\omega_i(a)$ (équation (3)).
- Deuxième fonction: Pour chaque position $i = 0, ..., L-1$, déterminer l'entropie rélative S_i (équation (4)), et pour les trois positions plus conservées aussi les acides aminées conservées (équation (5)). Tracer l'entropie rélative en fonction de la position i.
- Troisième fonction: Déterminer les paramètres $f^{(0)}(a)$ du modèle nul (équation (7)).
- Quatrième fonction (à appliquer à `testseq.txt`): Déterminer $\ell(b_i, ..., b_{i+L-1})$ (équation (9)) pour chaque sous-séquence de longueur L. Déterminer si il y a des sous-séquences de la famille definie par D_{train}. Tracer la log-vraisemblance en fonction de sa première position $i = 0, ..., N-L$.

Aide: Pour tester les fonctions, comparer vos résultats avec les valeurs suivantes :

$$\begin{aligned} \omega_0(\text{`}-\text{'}) &\simeq 0.31 \\ S_0 &\simeq 1.85 \\ \ell(b_0, ..., b_{L-1}) &\simeq -116 \end{aligned}$$

III. CO-ÉVOLUTION DE RÉSIDUES EN CONTACT

ATTENTION: Terminer la première partie du projet (PSWM) avant de vous consacrer à la deuxième.

Dans un alignment, il y a de l'information au-délà de ce que l'on peut détecter avec le simple modéle factorisé des PSWM. Comme discuté au TDs, la co-évolution de deux positions, qui sont en contact dans le repliement d'une protéine, induit des corrélations entre les occurences des acides aminées dans ces positions. Pour détecter ces corrélations, il faut calculer les nombres de co-occurences:

$$n_{ij}(a, b) = \text{nombre de séquences avec acide aminée } a \text{ en position } i \text{ ET avec acide aminée } b \text{ en position } j \ . \tag{10}$$

Maintenant on peut déterminer les poid respectifs

$$\omega_{ij}(a, b) = \frac{n_{ij}(a, b) + 1/q}{M + q} \ . \tag{11}$$

Le pseudocount est choisi pour garantir $\sum_b \omega_{ij}(a,b) = \omega_i(a)$.

Les corrélations sont quantifié par l'*information mutuelle*

$$M_{ij} = \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{A}} \omega_{ij}(a,b) \log_2 \frac{\omega_{ij}(a,b)}{\omega_i(a)\omega_j(b)} \tag{12}$$

pour chaque pair de positions $0 \leq i < j \leq L-1$. L'information mutuelle vaut zéro si et seulement si les deux positions sont statistiquement indépendantes ($\omega_{ij}(a,b) = \omega_i(a)\omega_j(b)$). Pour chaque dépendance statistique, M_{ij} prends une valeur positive.

Il est maintenant possible de comparer les paires de positions les plus corrélées (valeurs plus hautes de l'information mutuelle) avec les distances en `distances.txt`.

A. Réalisation

- Première fonction: Identique à la première fonction ci-dessus (calcul $\omega_i(a)$).
- Deuxième fonction: Pour chaque paire de positions $1 \leq i < j \leq L$ et chaque combinaison d'acides aminées $a, b \in \mathcal{A}$ (le trou compris), calculer le nombre d'occurence $n_{ij}(a,b)$ (équation (10)) et le poid $\omega_{ij}(a,b)$ (équation (11)).
- Troisième fonction: Pour chaque paire de positions $0 \leq i < j \leq L-1$, calculer l'information mutuelle M_{ij} (équation (12)).
- Quatrième fonction: Trier ("sort") les M_{ij}, sélectionner les 1,2,3,4,5,...,50 paires de positions avec les valeurs M plus grandes, chercher les distances correpondentes. Calculer la fraction des paires sélectionnées qui ont une distance plus petit que 8 (= qui sont des contacts). Tracer cette fraction en fonction du nombre de paires considerées. Est-ce que vous pouvez vérifier que les paires les plus correlées ont une probabilité élévée d'être en contact?

Test: $M_{0,1} \simeq 0.404$.