BIN3002 automne 2012 — Devoir 2

Miklós Csűrös

21 décembre 2012

À remettre avant 23 :59 le 17 janvier 2013 par courriel (à csuros@iro. . .). Travaillez seul/e.

2 Structure secondaire et alignement (20 points)

Dans cet exercice, vous explorez la connexion entre alignement de séquences protéiques et la structure secondaire. On cherche en particulier la réponse à deux questions : (1) est-ce que l'annotation de structure secondaire se transfère par l'alignement de résidus, et (2) est-ce qu'il y a une correlation entre structure et évolution de séquences.

2.1 Données et outils

Séquences. On travaillera avec des protéines dans la banque de données PDB (*Protein Data Bank*) dont la structure 3D a été déterminée. L'identificateur PDB, reconnu par beaucoup de ressources bioinformatiques, comprend 4 caractères. Vous pouvez facilement retrouver les séquences à NCBI (`http://www.ncbi.nlm.nih.gov/`) par l'identificateur (p.e. `1ea7`).

Alignements structuraux. La base de donnnée HOMSTRAD contient des alignement de structures pour 1000+ familles de protéines. Le site principal est `http://tardis.nibio.go.jp/homstrad/`. Exemple : en cherchant «xylose isomerase», on retrouve une famille (famille «xla») avec 6 séquences (1dxia, 1xyaa, . . . — les 4 premiers caractères forment les identificateurs canoniques de PDB).

La page associé avec la famille (`http://tardis.nibio.go.jp/cgi-bin/homstrad/showpage.cgi?family=xia&disp=str`) donne l'alignement structural, où le coloriage encode la structure secondaire superimposée.

Structure secondaire. On utilise la base de données DSSP (*Dictionary of protein secondary structure* ; `http://swift.cmbi.ru.nl/gv/dssp/`) pour retrouver la structure secondaire des protéines. On peut télécharger directement le fichier correspondant à chaque séquence par son identificateur PDB. Par exemple, on trouve l'annotation structurale de la séquence 1ea7 dans le fichier `ftp://ftp.cmbi.ru.nl/pub/molbio/data/dssp/1ea7.dssp`. Le site explique bien le format de fichiers .dssp (`http://swift.cmbi.ru.nl/gv/dssp/DSSP_3.html`). En particulier, DSSP fournit l'annotation de traits suivants W(fr)

- ⋆ H hélice α (≥ 4 résidus)
- ⋆ B résidu isolé dans un pont β
- ⋆ E brin β étendu au sein d'un feuillet (≥ 2 résidus)
- ⋆ G hélice 3_{10} (≥ 3 résidus)
- ⋆ I hélice π (≥ 5 résidus)
- ⋆ T coude fermé par une liaison hydrogène (3–5 résidus)
- ⋆ S coude sans liaison hydrogène

Alignement statistique. On utilise le logiciel FSA (*Fast Statistical Alignment* ; `http://fsa.sourceforge.net`) pour calculer l'alignement de séquences.

2.2 Tâches

2.2.1 Installation de FSA

▶ Téléchargez et compilez le logiciel FSA. On n'aura besoin que l'exécutable `fsa` (sans exonerate et mummer), donc une installation complète n'est pas nécessaire (faire `configure` et `make`).

2.2.2 Familles de protéines

▶ Téléchargez les alignements structuraux pour deux familles : "xylose isomerase" (6 membres) et "subtilase" (11 membres) de HOMSTRAD. Vous pouvez télécharger directement en format pir, et convertir à format Fasta par un outil en ligne. (Notez que HOMSTRAD contient occasionnellement des séquences qui ont été retirés de PDB.)

2.2.3 Calcul de l'alignement statistique

▶ Lancez le logiciel `fsa` sur les séquences d'une famille avec l'option `--gui`. Capturer le texte affiché à la sortie standard (stdout) : c'est l'alignement AMAP en format Fasta. Visualisez l'alignement calculé et l'alignement structural (de HOMSTRAD) par l'outil `mad` distribué avec `fsa` :

```
% java -XmxMEM -jar FSAHOME/display/mad.jar monali homstrad &
```

Ici, `MEM` est la spécification de borne pour le mémoire de Java (mettez, disons 4096M), et `FSAHOME` est le répertoire racine de FSA. Enregistrez un image TIFF pour les deux alignements calculés.

Avec l'option `--gui`, le logiciel `fsa` écrit un fichier avec extension `.probs` qui contient les probabilités postérieures d'alignements de résidues entre paires de séquences.

2.2.4 Transfer de l'annotation de structure

Examinez si l'annotation de structure secondaire se transfère d'une séquence à une autre. Supposons qu'on compare séquence i avec séquence j. Soit $S_i[k]$ le k-ème résidu en séquence i, et $A_i[k]$ la classification du même résidu dans la structure secondaire (H, B, E, G, I, T, S ou aucune). L'alignement statistique nous fournit la probabilité postérieure d'alignement de résidus et de trous :

$$p_{ij}(k \diamondsuit k') = \mathbb{P}\big\{ S_i[k] \text{ et } S_j[k'] \text{ sont homologues} \big\}$$
$$p_{ij}(- \diamondsuit k) = \mathbb{P}\big\{ S_j[k] \text{ n'a pas de homologue} \big\}$$

Pour transférer l'annotation A_i vers la séquence j, on calcule

$$\hat{A}_j^{(i)}[k, \alpha] = \sum_m p_{ij}(m \diamondsuit k) \cdot \{A_i[m] = \alpha\} \tag{2.1}$$

pour chaque catégorie

$$\alpha \in \mathcal{S} = \{\mathsf{H}, \mathsf{B}, \mathsf{E}, \mathsf{G}, \mathsf{I}, \mathsf{T}, \mathsf{S}, \emptyset\}.$$

Dans d'autres mots, $\hat{A}_j^{(i)}[k, \alpha]$ donne la probabilité postérieure que $A_j[k] = \alpha$, en assumant que deux résidus homologues ont le même rôle dans la structure secondaire.

On mesure le succès en comparant la vraie annotation A_j et les annotations inférées $\hat{A}_j^{(i)}$. En particulier, pour chaque catégorie α, on compte les vraies positives : $\mathsf{TP}_\alpha(i \to j) = \sum_k \hat{A}_j^{(i)}[k, \alpha] \cdot \{A_j[k] = \alpha\}$. Après normalisation par le nombre de sites α, on obtient une mesure de sensitivité

$$\mathsf{sens}_\alpha(i \to j) = \frac{\mathsf{TP}_\alpha(i \to j)}{\sum_k \{A_j[k] = \alpha\}}.$$

De façon similaire, on mesure la spécificité par le nombre de fausses positives

$$\mathsf{spec}_\alpha(i \to j) = \frac{\sum_k \hat{A}_j^{(i)}[k, \alpha] \cdot \{A_j[k] \neq \alpha\}}{\sum_k \{A_j[k] \neq \alpha\}}.$$

▶ Calculez la sensitivité et spécificité de transfer d'annotation pour une séquence j fixe (de votre choix) dans chaque famille, à partir de chaque autre séquence i. Pour cela, vous devez écrire un script qui parse les probabilités postérieures (fichier `.probs`) et l'annotation DSSP. Compilez les résultats dans un tableau (rangée=séquence de référence i, colonne=catégorie α) pour chaque famille.

2.2.5 Évolution de structure

Examinez si les contraintes structurales influencent l'évolution des séquences. On s'intéresse en particulier à la fréquence de substitutions de résidus, ainsi qu'à celle d'insertions et de suppressions. On définit la conservation de résidu k en séquence j par

$$\rho_j[k] = \frac{\sum_{i \neq j} \sum_m p_{ij}(m \diamondsuit k) \cdot \{S_i[m] = S_j[k]\}}{\sum_{i \neq j} \sum_m \{S_i[m] = S_j[k]\}}.$$

De plus, on définit le *flottement* de chaque résidu par $\gamma_j[k] = \frac{\sum_{i \neq j} p_{ij}(- \diamondsuit k)}{n-1}$, où n est le nombre de séquences dans la famille. Les comptes de catégories se calculent à

travers les résidus qui y appartiennent :

$$\rho_j[\alpha] = \frac{\sum_k \rho_j[k] \cdot \{A_j[k] = \alpha\}}{\sum_k \{A_j[k] = \alpha\}}$$
$$\gamma_j[\alpha] = \frac{\sum_k \gamma_j[k] \cdot \{A_j[k] = \alpha\}}{\sum_k \{A_j[k] = \alpha\}}$$

► Calculez ρ et γ pour une séquence fixe j de votre choix dans chaque famille. Compilez les résultats dans un tableau pour chaque famille et classification de structure secondaire. Est-ce que vous pouvez déduire des caractérisations générales sur l'évolution de séquences selon la structure secondaire ?

2.3 Soumission de travail

Soumettez un document PDF qui montre les alignements coloriés (§2.2.3), ainsi que les tableaux qui résument les résultats de §2.2.4 et de §2.2.5. Expliquez dans le même document quels outils additionnels vous avez développé (p.e., parser DSSP), et toutes ressource additionnelle que vous avez utilisée. Il n'est pas nécessaire de soumettre votre code.