Chapitre I

Greg

25 septembre 2006

1 Introduction

Une grande quantité de connaissances biologiques a été accumulée à notre époque. Pour donner du sens à toute cette connaissance nous avons besoin de connaissances sur la biologie des organismes et des cellules. Une grande partie du challenge consiste à classer, organiser et parser l'immense quantité de données que nous avons à disposition.

1.1 Similarité de séquences, homologie et alignement

La Nature n'est pas un inventeur, mais un penseur. Ainsi les nouvelles séquences ne sont pas issues de rien, mais issues de séquences pré-existantes, ce qui est un bien pour l'analyse de séquences par ordinateur.

Définition 1 *Si on connaît quelque chose sur une séquence (sa fonction, structure, ...) et que l'on voit qu'elle est similaire à une autre séquence dont on ne sait rien, alors on dit que les deux séquences sont homologues et que l'on transfert par homologie les fonctions et la structure de la première séquence sur la séquence dont on ne sait rien.*

Pour savoir si deux séquences sont similaires, il faut d'abord commencer par faire un alignement. Pour une méthode de score définie, les algorithmes d'alignement trouvent toujours la solution optimale. Le problème est de définir une méthode de score qui soit biologiquement significative. En particulier, on voudrait prendre en compte le fait que deux séquences peuvent avoir une relation évolutionnaire, des structures 3D et autres qualités qui contraignent l'évolution de la séquence primaire.

Définition 2 *Les matrices de substitution servent à qualifier les préférences de substitution d'un acide aminé en en autre au travers de l'Evolution.*

1.2 Modèle probabiliste

Un modèle probabiliste, qu'est-ce que c'est ? Un modèle est un système capable de simuler l'objet en considération. Un modèle probabiliste est un modèle qui produit différentes sorties, chacune avec une probabilité qui lui est associée.

1.2.1 Exemple

Les séquences sont des chaînes de caractères composées d'un alphabet de 20 lettres (dans le cas des protéines) ou de 4 lettres (dans le cas des séquences d'ADN). Soit un résidu a et soit q_a la probabilité qu'il apparaisse, indépendement des autres acides aminés dans la séquence. Si on dénote la séquence par $x_1 \dots x_n$, alors la probabilité de la séquence entière sera $q_{x1}, q_{x_2} \dots q_{xn} = \prod_{i=1}^{n} q_{xi}$. C'est ce que l'on appelle le modèle de base (ou hypothèse nulle). Nous allons l'utiliser pour le confronter à d'autres modèles et ainsi savoir si ces autres modèles sont meilleurs ou moins bon.

1.2.2 Maximum de vraissemblance

Les paramètres des modèles probabilistes sont estimés à partir d'ensembles d'entraînement. Par exemple, la probabilité q_a est un paramètre dans le modèle de base. Ce paramètre peut être estimé en regardant la fréquence d'apparition de chaque acide aminé dans une base de données telle que Swiss-Prot. Ainsi, si une séquence n'est pas biaisée par un certain résidu, on peut dire que les q_{a_i} sont des estimateurs raisonables de notre modèle probabiliste. Cette manière d'estimr des modèles est appelé *estimation par le maximum de vraissemblance.*

1.2.3 Exemple

Lançons 80 fois une pièce prise au hasard parmi 3 pièces dont les probabilités respectives de donner face sont $\frac{1}{3}, \frac{1}{2}, \frac{2}{3}$. Après les 80 lancés, on compte 49 piles.

Conaissant la séquence (PFPFPFPFFFPPP), on peut essayer de trouver quelle pièce a été lancée ! On calcule alors :

$$P(49 \text{ piles}|p = \frac{1}{2}) = 0.000$$

$$P(49 \text{ piles}|p = \frac{1}{3}) = 0.012$$

$$P(49 \text{ piles}|p = \frac{2}{3}) = 0.054$$

La vraissemblance est maximisée par la valeur $\frac{2}{3}$ et est donc l'estimateur de maximum de vraissemblance pour le paramètre p.

La vraissemblance est en arrière-plan de la probabilité : Soit B donné, on utilise alors les probabilité conditionnelles $P(A|B)$ pour déduire A et, pour un A donné, on utilise la fonction de vraissemblance $L(A|B)$ pour déduire B. Cela est contenu dans le théorème de Bayes qui dit :

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} \tag{1}$$

où $P(A|B)$ et $\frac{P(A|B)}{P(A)}$ sont les fonctions de vraissemblance pour B étant donné A.

Quand on estime des paramètres pour un modèle sur un nombre limité de données, il y a danger d'overfitting, ce qui veut dire que le modèle devient très très bon sur ces données mais qu'il est très mauvais sur de nouvelles données.

1.2.4 Probabilité jointes

$$P(X, Y) = P(X|Y) \cdot P(Y) \tag{2}$$

1.2.5 Probabilités marginales

$$P(X) = \sum_{Y} P(X, Y) = \sum_{Y} P(X|Y) \cdot P(Y) \tag{3}$$

1.2.6 Théorème de Bayes

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)} \tag{4}$$

1.2.7 Exemple

Les protéines extra-cellulaires ont une composition en acide aminé qui est diffrente des protéines intra-celullaires. Supposons que la cysétine soit beaucoup plus présente dans les protéines extra-cellulaires. Utilisons ce fait pour déterminer si une séquence $x = x_1 \dots x_n$ est intra ou extra-cellulaire.

Premièrement, séparons les séquences de Swiss-Prot en deux : les séquences extra-cellulaires d'un côté, les séquences intra-cellulaires d'un autre. Calculons un ensemble de fréquences q_a^{int} pour les protéines intra-cellulaires et q_a^{ext} pour les protéines extra-cellulaires. Calculons entre la probabilité p^{int} qu'un séquence soit intra-cellulaire et p^{ext} qu'elle soit extra-cellulaire. On sait aussi que : $p^{\text{int}} = 1 - p^{\text{ext}}$. Ces 2 probabilités sont appelées probabilités à priori, car elles représentent la meilleure estimation que l'on puisse faire sur la séquence *avant* avant d'avoir vu des informations sur la séquence elle-même.

Maintenant, on peut écrire que $P(x|\text{ext}) = \prod_i q_{x_i}^{\text{ext}}$ et $P(x|\text{int}) = \prod_i q_{x_i}^{\text{int}}$. On sait aussi (car une séquence et soit intra soit extra-cellulaire) que : $p(x) = p^{\text{ext}} \cdot P(x|\text{ext}) + p^{\text{int}} \cdot P(x|\text{int})$. On souhaite maintenant calculer : connaisant x quelle est la probabilité qu'elle soit extra-cellulaire ? Pour cela on utilise le théorème de Bayes :

$$P(\text{ext}|x) = \frac{p^{\text{ext}} \cdot \prod_i q_{x_i}^{\text{ext}}}{p^{\text{ext}} \cdot \prod_i q_{x_i}^{\text{ext}} + p^{\text{int}} \cdot \prod_i q_{x_i}^{\text{int}}} \tag{5}$$