Statistique pour la bio-informatique

Séance 4 - 22 Octobre 2003

Matrices de scores (Brouillon provisoire)

1 Scores

Aligner deux séquences est une démarche utile pour analyser l'information biologique. Cela sert par exemple à détecter des protéines homologues, ou des régions conservées à l'intérieur de différentes séquences de protéines. Afin de quantifier la similitude entre les deux séquences, on calcule en général un score. Ce score repose sur un système qui permet d'attribuer une valeur élémentaire à chaque position lorsque deux séquences sont éditées l'une sous l'autre. Par exemple, considérons les deux séquences

$$\begin{array}{c} ggagac|tgtagac|agctaat|gctata \\ gaacgc|cctagcc|acgagcc|cttatc \end{array}$$

Donnons le score $+1$ si deux nucléotides en un site donné sont identiques et -1 sinon. En partant de la gauche vers la droite, le score cumulé se comporte comme une marche aléatoire (voir la figure 1). Le comportement de cette marche aléatoire sert de base pour le calcul d'un test statistique dans l'algorithme BLAST (Basic Local Alignement Search Tool) dont nous décrirons le principe plus tard.

Un score élémentaire est un élément d'une matrice qui rend compte de tous les états possibles en fonction de l'alphabet utilisé dans la description des séquences. Pour les acides nucléiques, la matrice identité est principalement employée (matrice 4×4). Elle rend compte de l'identité des résidus pour chacune des positions de la comparaison.

En général, le calcul du score associé à un alignement repose sur le choix d'une **fonction de score** appropriée. Le choix de cette fonction de score doit prendre en compte l'historique de l'évolution des séquences comportant des substitutions, insertions, délétions, etc. Dans le modèles les plus simples, on considère simplement les substitutions, et on suppose que leurs occurrences se produisent indépendamment en chaque site de la séquence (nucléotide ou acide aminé). Dans cette situation, la fonction de score est la somme des contributions associées aux différents événements de mutation.

Considérer l'insertion ou la suppression d'une ou plusieurs bases peut améliorer le score d'une comparaison et faire davantage ressortir les zones identiques ou très

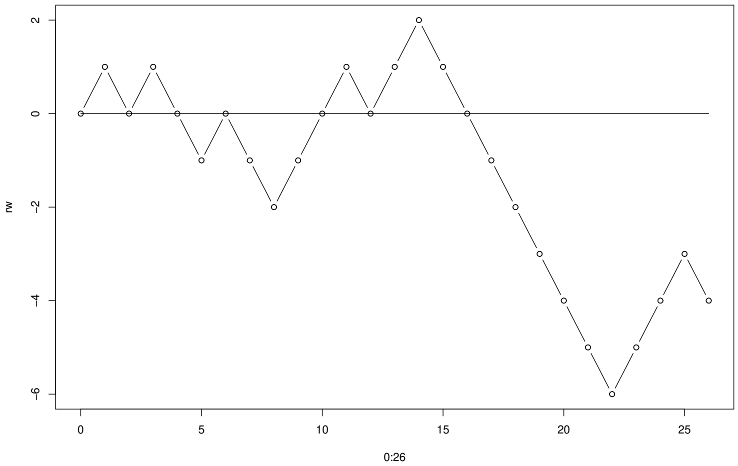

FIG. 1 – Marche aléatoire associée à la comparaison de deux séquences

proches. Ces brèches (en anglais *gap*) que l'on impose aux séquences seront typiquement pénalisantes dans le calcul du score. *Pour des raisons simplificatrices, nous ne considérerons dans la suite que le cas le plus simple ou les gaps sont absents.*

Dans cette séance, nous allons montrer comment on génère une fonction de score correspondant aux **substitutions** pour les acides aminés. Dans ce cas, les scores correspondants aux différentes substitutions peuvent être stockés dans une **matrice de substitution**. Cette matrice contient, pour toute paire (A, B), le score $s_{(A,B)}$ associé à l'alignement des deux symboles. Il existe plusieurs façons construire une matrice de substitution. La construction de matrices de substitution protéiques constitue une part importante de la modélisation biologique dans les programmes d'alignement ou de recherche de similitude en bio-informatique.

Pour les séquences protéiques, la construction de matrices de scores fait appel à des concepts sophistiqués et ne se contente pas d'indiquer la coïncidence des acides aminés. On considère qu'un acide aminé peut être substitué à un autre en fonction de certaines propriétés sans que la structure ou la fonctionnalité d'une protéine soit grandement altérée. Ainsi, on peut classer les acides aminés en familles et obtenir un système de scores qui rende compte de l'affinité des résidus protéiques entre eux. Les matrices protéiques, de dimension 20, peuvent être classées en deux catégories.

La première catégorie regroupe des matrices construites à partir des caractéristiques physico-chimiques des acides aminés (BLOSUM). Par exemple, si les scores sont calculés à partir de la matrice BLOSUM50, les scores élémentaires obtenus lors de la lecture de l'alignement protéique

$$TQLAAWCRMTCFEIECKV$$

$$RHLDSWRRASDDARIEEG$$

sont

$$-1, 1, 5, -2, 1, 15, -4, 7, -1, 2, -4.$$

À nouveau, le score cumulé peut être vu comme la réalisation d'une certaine marche aléatoire, plus complexe que celle décrite dans la figure 1.

La seconde catégorie, de loin la plus utilisée, se fonde sur un modèle statistique. Elle regroupe les matrices issues d'études montrant le caractère de substitution des acides aminés au cours de l'évolution (matrices PAM). C'est cette dernière méthode que nous présentons ci-dessous.

2 Scores et tests d'hypothèse

Une méthode raisonnable pour fabriquer un score consiste à comparer deux hypothèses différentes. Considérons deux acides aminés A et B dans un alignement. La première hypothèse, dite *hypothèse nulle*, est que les deux acides ne possèdent pas de lien évolutionniste (pas d'ancêtre commun) et se trouvent être alignés par hasard. Supposons que l'acide A apparaisse avec la probabilité p_A et que l'acide B apparaisse avec la probabilité p_B. Sous l'hypothèse nulle, l'alignement (A, B) apparaît alors avec la probabilité

$$\mathrm{P}(A, B \text{ alignes}) = p_A p_B \,.$$

L'hypothèse alternative est que A et B partagent un ancêtre commun. Il existe dans ce cas de nombreux modèles probabilistes permettant de calculer la probabilité pour que A et B se trouvent alignés connaissant l'acide aminé en ce site précis de la séquence ancestrale. Les plus connus de ces modèles concernant les acides nucléiques sont les modèles de *Jukes-Cantor*, ou de *Kimura*, etc. Il permettent de calculer

$$\mathrm{P}(A, B \text{ alignes}) = p_{AB} \,.$$

La théorie statistique des tests d'hypothèses permet de rejeter l'hypothèse nulle contre une alternative donnée. L'un des résultats de cette théorie est que les tests les plus puissants sont fondés sur les rapports de vraisemblance, ou **log-odds**. Ceci suggère de choisir la fonction de score suivant

$$s_{(A,B)} = \log p_{AB} - \log p_A - \log p_B.$$

Ainsi le calcul du score se ramène au choix d'un modèle d'évolution.

Ainsi, lorsque nous considérons l'alignement deux séquences r et s de même longueur ℓ. Le rapport de vraisemblance s'écrit

$$\frac{p_{rs}}{p_r p_s} = \prod_{i=1}^{\ell} \frac{p_{r_i s_i}}{p_{r_i} p_{s_i}}$$

et la fonction de score est égale à

$$\mathcal{S}(r,s) = \sum_{i=1}^{\ell} s_{(r_i,s_i)}.$$

Cette démarche statistique a donc pour objectif de minimiser le nombre faux négatifs lors de la comparaison, c'est-à-dire la nombre de sites se voyant attribuer de mauvais scores alors qu'ils sont "vraiment" alignés.

3 Matrices PAM

3.1 Présentation

Les matrices que nous décrivons sont liées à un modèle très simplifié de l'évolution. Les matrices de type PAM (Percent Accepted Mutation), la matrice de mutation de Dayhoff (1978), sont sans aucun doute celles qui ont été les plus utilisées dans les programmes de comparaison de séquences protéiques. Elles représentent les échanges possibles ou acceptables d'un acide aminé par un autre lors de l'évolution des protéines (Dayhoff et al., 1978). Elles ont été déduites de l'étude de 71 familles de protéines (de l'ordre de 1300 séquences) très semblables (moins de 15 % de différence) que l'on pouvait facilement aligner. De ces alignements, une matrice de probabilité a été calculée où chaque élément de la matrice donne la probabilité qu'un acide aminé A soit remplacé par un acide aminé B durant une étape de l'évolution. Cette matrice de probabilité de mutation correspond en fait à une substitution acceptée pour 100 sites durant un temps d'évolution particulier, c'est à dire une substitution qui ne détruise pas l'activité de la protéine. On parle ainsi de matrice 1PAM. Si l'on multiplie la matrice par elle-même x fois, on obtient une matrice xPAM qui donne des probabilités de substitution pour des distances d'évolution plus grande.

Cette section explique le principe détaillé de construction de la matrice PAM. Les ingrédients principaux de son succès reposent sur la théorie des rapports de vraisemblance et l'estimation d'une chaîne de Markov. L'idée générale est de former des blocs de séquences alignées similaires, puis d'estimer la fréquence des substitutions lors de l'évolution.

3.2 PAM

La démarche de test statistique rappelée dans la section précédente restreint donc la construction de score à un problème de choix d'alternative, c'est à dire de choix d'un modèle d'évolution pertinent pour les proteines. Dans cette approche, nous devrons être capable de calculer la probabilité p_{AB} que deux acides descendent d'un même ancêtre dans un alignement de paire et/ou estimer cette probabilité à partir de deux séquences r et s.

En utilisant l'hypothèse d'indépendance faite au paragraphe précédent, il est naturel d'estimer la probabilité p_{AB} à l'aide de la fréquence d'occurrence du couple (A, B)

$$p_{AB} \approx \frac{n_{A,B}(r, s)}{\ell}$$

où r et s sont deux séquences alignées de longueurs ℓ et $n_{A,B}$ représente le nombre d'occurences. Par ailleurs cet estimateur correspond aussi au maximum de vraisemblance pour ce problème d'estimation.

L'approche décrite ci dessus correspond à la démarche adoptée par Dayhoff et ses collègues en 1978 afin de calculer les célèbres matrices de substitution de type PAM. Toutefois, un problème fondamental se posait pour appliquer cette méthode. En effet, la seule manière disponible pour trouver des séquences r et s alignées ou homologues était d'appliquer un algorithme de recherche d'alignement, reposant lui même sur une matrice de score. Cela ressemblait donc bien à un cercle vicieux.

Pour contourner ce problème, Dayhoff a considéré l'étude de 71 familles de protéines (de l'ordre de 1300 séquences) très semblables (moins de 15 % de différence) que l'on pouvait facilement aligner sans algorithme reposant sur les scores. Cette approche conduit à un biais important, car la matrice de substitution ainsi construite est uniquement valable pour des séquences suffisament homologues, c'est à dire proche dans l'évolution. Que dire alors de séquences dont les relations généalogiques (phylogénétiques) sont plus éloignées ? L'idée majeure fut d'introduire un paramètre supplémentaire t représentant le temps à l'échelle de l'évolution et de considérer que les probabilités mesurées au présent sont liées à celles observables dans le futur ou le passé

$$p_{AB} = p_{AB}(t).$$

En supposant que les événements de substitutions possèdent une dynamique en équilibre et réversible dans le temps (on observe des transitions de A vers B ou de B vers A

dans les mêmes proportions), on obtient

$$M_{A,B} = M^t_{A,B} \equiv \frac{p_{AB}(t)}{p_A}.$$

Le bénéfice de cette approche est que l'on peut simplement extrapoler les probabilités de transition sur une période plus longue, par exemple kt, en calculant la puissance k^{e} de M

$$M^k_{A,B} = M^{tk}_{A,B}.$$

Afin de se débarrasser du paramètre t qui est inconnu, on effectue la normalisation définie pour $i \neq j$, par

$$m_{ij} = cM_{ij}$$

et pour $i = j$

$$m_{ii} = 1 - c\sum_k M_{ik}$$

La matrice m, appelée 1PAM, est une matrice de transition pour laquelle on attend un pour cent de mutation en moyenne lors d'une transition. La constante c peut être déterminée pour respecter cette contrainte. Supposons que la fréquence de l'acide j est p_j, alors nous avons

$$0.01 = \sum_i p_i \sum_{j\neq i} m_{ij} = c\sum_i p_i \sum_{j\neq i} M_{ij}$$

et ainsi

$$c = \frac{0.01}{\sum_i p_i \sum_{j\neq i} M_{ij}}.$$

3.3 Variante

L'idée sous-jacente à PAM est de former des blocs de séquences alignées similaires, puis d'estimer la fréquence des substitutions lors de l'évolution. Pour cela, on construit un arbre phylogénétique pour chaque bloc (décrivant l'évolution des différentes séquences). Une méthode possible est celle de *parsimonie*, qui représente les arbres ayant le plus petit nombre de substitutions. Il n'y a d'ailleurs pas forcément unicité de tels arbres. Considérons par exemple un bloc constitué des trois séquences

$$AA \quad AB \quad BB.$$

Il y a dans ce cas 5 arbres possibles pour la méthode de parsimonie. Chaque arbre présente deux substitutions. Par exemple, un arbre parmi les 5 possibles est celui décrit ci dessous

$$\begin{array}{ccccc}
 & & AB & & \\
 & / & & \backslash & \\
 & AB & & & \backslash \\
/ & & \backslash & & \backslash \\
AA & & AB & & BB
\end{array}$$

On examine ensuite les substitutions le long des arbres obtenus. Pour un symbole différent du symbole parent, on compte +1. Dans le cas contraire, on compte +2. Dans l'arbre ci-dessus par exemple, le décompte correspondant à $A \to B$ ou $B \to A$ est +2, le décompte correspondant à $A \to A$ est +6, le décompte correspondant à $B \to B$ est +6.

Lorsqu'il y a plusieurs arbres, on effectue ce décompte sur la totalité des arbres, puis on divise par le nombre d'arbres. Sur la totalité des 5 arbres, nous obtenons (exercice), la matrice de comptage

$$M = \begin{array}{c|cc}
 & A & B \\
\hline
A & 6 & \\
B & 2 & 6
\end{array}$$

Cette matrice est convertie en matrice de transition par normalisation. En numérotant de 1 à 20 les acides aminés, nous avons

$$\forall i, j = 1, \ldots, 20, \quad a_{ij} = \frac{m_{ij}}{\sum_k m_{ik}}$$

et, pour $c > 0$ suffisamment petit, nous pouvons considérer la matrice de transition P, définie pour $i \neq j$, par

$$p_{ij} = c a_{ij}$$

et pour $i = j$

$$p_{ii} = 1 - c \sum_k a_{ik}$$

La matrice 1PAM est une matrice de transition pour laquelle on attend un pour cent de mutation en moyenne lors d'une transition. Cela permet de déterminer la valeur de la constante c. Supposons que la fréquence de l'acide j est p_j, alors nous avons

$$0.01 = \sum_i p_i \sum_{j \neq i} p_{ij} = c \sum_i p_i \sum_{j \neq i} a_{ij}$$

et ainsi

$$c = \frac{0.01}{\sum_i p_i \sum_{j \neq i} a_{ij}}.$$

Pour obtenir la matrice XPAM, il suffit d'élever 1PAM à la puissance X. Enfin, la matrice de scores S sera définie comme

$$S(i,j) = \log(\frac{p_{ij}}{p_j})$$

où de manière équivalente

$$S(i,j) = \log(\frac{p_i p_{ij}}{p_i p_j}).$$