Vendredi 23 janvier 2004

Démo 2 – Programmation Dynamique 2

Jean-Eudes Duchesne

1. Programmation dynamique avec « trous » (gap)

Étant donné la présence d'introns et d'exons, dans l'ADN (modèle « Block-indel »), comment transformer les règles de l'alignement global entre deux séquences pour pouvoir aligner une séquence d'ADN avec son homologue d'ARN ? (autres motivations : duplications, insertions par rétro-virus, translocations)

Alignement global usuel entre s et t :

$$D(i, j) = \text{MAX}[\ D(i-1, j) + p(s_i, -),\ D(i, j-1) + p(-, t_j),\ D(i-1, j-1) + p(s_i, t_j)\ ]$$

Problème ? La matrice D qui contient les résultats de la distance d'édition ne contient pas assez d'information. Entre autre, elle ne contient pas d'informations sur l'état courant de la récurrence. Sommes-nous dans un trou (dans la solution optimale) présentement ou non ? Ceci influence la décision à prendre donc les règles de bases ne sont pas suffisantes.

Modèle 1 (simple) : **Pondération constante**

W_t est le coup associé à l'ouverture d'un gap. Poids d'un trou de taille q est donc W_t peu importe q. Les matrices E et F représentent les deux cas associés aux indels. Comme à chaque opération il est nécessaire de considérer l'ouverture d'un gap ou l'extension d'un gap, il est nécessaire de déterminer préalablement la solution optimale entre ces deux choix.

$$D(i, j) = \text{MAX}[\ E(i-1,j-1),\ F(i-1,j-1),\ D(i-1, j-1)\ ] + p(s_i, t_j)$$

$$E(i,j) = \text{MAX}[\ E(i-1,j), D(i-1, j) + W_t]$$
$$F(i,j) = \text{MAX}[\ F(i,j-1), D(i, j-1) + W_t]$$

conditions initiales :

$$D(i, 0) = E(i, 0) = W_t$$
$$D(0, j) = F(0, j) = W_t$$

Le reste est supposé = ∞ (« **nil** » dans le script gap.rb,).

Complexité en temps ? O(nm)

Modèle 2 : **Pondération linéaire ou affine**

W_t est le coût associé à l'ouverture d'un gap et W_e est le coût associé à l'extension d'un gap. Le poids d'un trou de taille q est donc $W_t + qW_e$ peu importe q.

$$D(i, j) = MAX \begin{bmatrix} E(i-1,j-1), \\ F(i-1,j-1), \\ D(i-1, j-1) \end{bmatrix} + p(s_i, t_j)$$

$$E(i,j) = MAX [\quad E(i-1,j), D(i-1, j) + W_t] + W_e$$
$$F(i,j) = MAX [\quad F(i,j-1), D(i, j-1) + W_t] + W_e$$

conditions initiales :

$$D(i, 0) = E(i, 0) = W_t + iW_e$$
$$D(0, j) = F(0, j) = W_t + jW_e$$

Complexité en temps ? O(nm)

Modèle 3 : **Pondération quelconque**

W() est une fonction quelconque qui retourne le coût associé à la présence d'un gap de taille q. Ainsi, le poids d'un trou de taille q est W(q), pour tout q. Les différences avec les modèles précédents sont : Absence de W_t ce qui implique que l'ouverture d'un trou n'est pas pénalisée, seule sa taille importe ; Comme la fonction n'est pas nécessairement croissante, la taille de q importe (il faut donc maximiser la fonction sur toutes les tailles possibles de q).

$$D(i, j) = MAX \begin{bmatrix} E(i-1,j-1), \\ F(i-1,j-1), \\ D(i-1, j-1) \end{bmatrix} + p(s_i, t_j)$$

$$E(i,j) = MAX_{0<k<i} [\quad D(k,j) + W(i-k)]$$
$$F(i,j) = MAX_{0<k<j} [\quad D(i, k) + W(j-k)]$$

Conditions initiales :

$$D(i, 0) = E(i, 0) = W(i)$$
$$D(0, j) = F(0, j) = W(j)$$

Complexité en temps ? $O(nm^2 + n^2m)$

2. Exercice : Alignement entre séquence et protéine

Devoir 1, H2003 :Étant donné une séquence nucléique S et une séquence protéique P, concevoir un algorithme qui calcul le meilleur alignement entre S et P en utilisant les match/mismatch et indels.

$D(i, j) = MAX [$ $D(i-1,j) + Wd$, (suppression)
$D(i,j-1) + Wi$, (insertion)
$D(i-3, j-1) + p(s_i, t_j)$] où p est une matrice donnée.

conditions initiales :

$D(i,0) = iWd$ si i est positif et 0 sinon
$D(0,j) = jWi$

Pour rester équivalent, on peut fixer $Wd = 1$ et $Wi = 3$.

Problème ? Cet algorithme considère seulement le cas ou l'on insère un acide aminé complet dans la séquence. C'est donc dire qu'une insertion ajoute obligatoirement 3 nucléotides ! Comment régler ce problème ? Ceci vous est laissé en exercice.

3. Matrices de scores

PAM (Point Accepted Mutation ou Percent Accepted Mutation)

Matrice de substitution pour les acides aminés. Essentiel pour les utilisateurs de recherche de séquences de protéines. Les scores doivent permettre de comprendre la signification biologique d'un alignement.

Unité PAM : Deux séquences S et T divergent d'une unité PAM si la série des mutations ponctuelles (substitutions) qui a converti S à T est telle qu'en moyenne une seule mutation est survenue tous les 100 acides aminés.

Il n'y a pas de correspondance entre les divergences locales en acides aminés et les unités PAM. Des séquences divergeant 200 PAM peuvent quand même avoir 25% d'identité entre les séquences.

Matrice PAMn : Étant donné deux acides aminés A_i et A_j, la case (i, j) de la matrice PAM contient la fréquence avec laquelle A_i est remplacé par A_j dans les séquences qui divergent de n unité PAM.

Les matrices PAM sont destinées à comparer des séquences d'acides aminés qui divergent d'un nombre spécifique d'unités PAM. Par exemple, la matrice PAM250 est très utilisée.

Dans un monde idéal :

1) Aligner un ensemble de séquences divergeant de n unités PAM deux à deux.
2) Pour chaque couple (A_i, A_j) calculer $f(i, j) = nb\ (A_i, A_j) / nb$ total d'appariements.
3) Normaliser par la fréquence de substitution due au hasard. $P(i,j) = \log (f(i,j) / f(i)f(j))$ où $f(x)$ est la fréquence d'apparition de A_x dans les séquences.

Problèmes : alignements corrects ? Il faut une matrice pour faire l'alignement, mais il faut un alignement pour obtenir une matrice … boucle sans fin !

Dans un monde un peu moins idéal :

1) Utiliser des séquences très homologues, donc faciles à aligner (85%+ de correspondances) pour faire comme la méthode idéale, mais avec un très faible nombre PAM (on obtient une matrice PAM 1 qui sera ensuite utilisée pour extrapoler les autres matrices PAM)..
2) Calculer : $\log (f(i)\ M^n(i,j) /\ f(i)f(j)) = \log (M^n(i,j) / f(j))$ où M est une matrice PAM 1, $M^n(i,j)$ où est la probabilité que A_i se transforme en A_j en n unités PAM.

Le modèle implique deux suppositions :

1) Les mutations sont indépendantes. C'est à dire que chaque substitution dépend seulement de la position à laquelle elle se produit et à la probabilité correspondante.
2) Les séquences comparées ont une composition « moyenne normale » en acides aminés.

Sources d'erreurs :

1) Certaines séquences n'ont pas une composition « moyenne normale » en acides aminés.
2) La méthode n'est pas assez précise pour donner des probabilités fiables pour des cas de substitutions rares. Par exemple 32 paires différentes ont une probabilité 0 dans PAM 1.
3) Les erreurs d'approximation ou d'expérimentation dans la matrice PAM 1 sont amplifiées dans l'extrapolation de la matrice PAM 250.
4) La représentation temporelle de l'évolution est imparfaite. On observe souvent des blocks de nucléotides conservés ou des espèces qui évoluent à des vitesses différentes. Ceci implique que le remplacement d'un nucléotide n'est pas nécessairement équiprobable sur la longueur de la séquence.