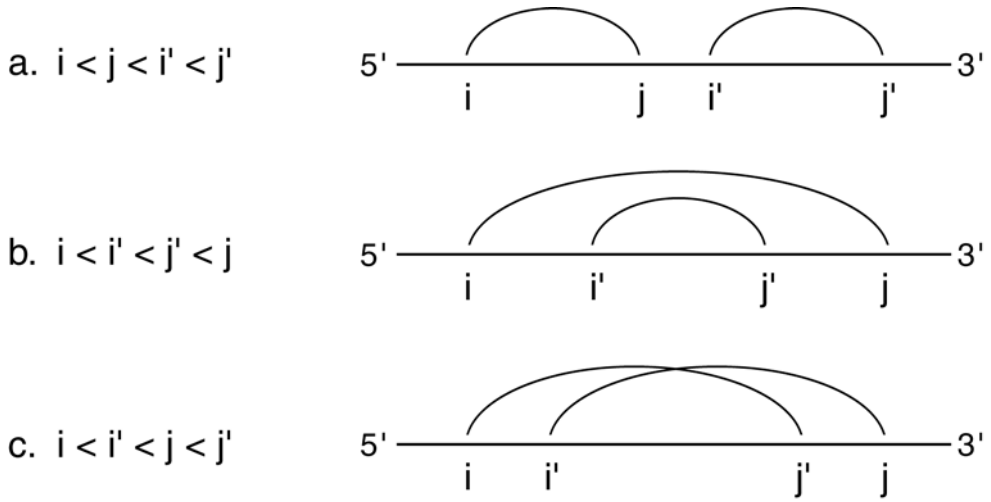
Vendredi 19 Mars 2004

Démo 8 – Prédiction de structures secondaires de l'ARN

Jean-Eudes Duchesne

1. Définitions

Peu importe comment les biologistes définissent les structures secondaires, les définitions biologiques des macromolécules conviennent rarement à un contexte computationnel. Ainsi, il faut redéfinir les structures secondaires en termes significatif d'un point de vu informatique. Ainsi, une hélice peut être définie comme étant une sous séquence commençant à l'indice *i* et se terminant à l'indice *j* où les caractères font des appariements Watson-Crick (et possiblement Wobble) en (i,j), (i+1, j-1), (i+2, j-2) et ainsi de suite, mais avec la condition que (j-i) > 3 (soit la taille minimale de la boucle). Avec cette définition, trois cas peuvent survenir :



Le cas (a.) correspond à deux hélices contiguës, le cas (b.) correspond à une boucle interne et le cas (c.) représente un pseudo nœud.

Le problème à considérer est le suivant : Étant donné une séquence d'ARN S, trouver sa structure secondaire optimale !

2. Approche combinatoire (Pipas & McMahon)

L'approche se fait en deux étapes ; d'abord trouver toutes les hélices possibles et ensuite choisir l'ensemble d'hélices optimal.

① Trouver toutes les hélices possibles dans S.

```
Pour i = 1 à n-p+1
   Pour j = i+p+1 à n
      Si pair(i,j)
         l = 1
         Tant que pair(i+l, j-l), l++
         Si l > lmin ajouter hélice(i,j,l) dans l'ensemble H
Retourner H
```

② Sélectionner le sous-ensemble optimal.

Hic ! À partir d'une séquence donnée, il y a 2^n sous-ensembles possibles. Solution à ce problème : approches probabilistes (Monte Carlo, Algorithmes Génétiques, etc.).

3. Approche récursive (Zuker)

Pour simplifier le problème de prédiction des structures secondaires, il est possible d'assigner des valeurs aux éléments de la structure secondaire. L'idée sous-jacente est que chaque nucléotide possède une énergie (soit sa capacité de faire des liaisons avec d'autres molécules) qui contribue à déstabiliser la molécule et chaque paire réduit l'énergie totale de la molécule puisque ceci empêche les nucléotides qui compose la paire d'interagir avec d'autres éléments. Ainsi, le problème devient donc la minimisation de l'énergie libre pour une séquence S d'ARN donnée.

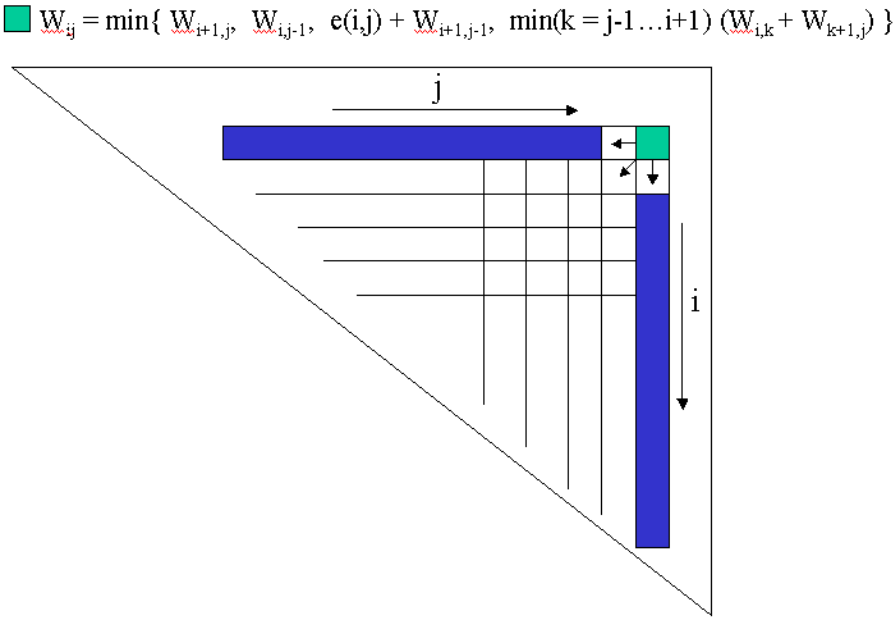
Un modèle typique simple assigne d'énergie –3, -2 et –1 au paires CG, AU et GU (Wobble), respectivement. L'énergie de la structure entière est la somme de toutes les énergies :

$$E(S) = \sum_{(i,j)\in S} e(i,j)$$

Pour calculer l'énergie libre minimale, il suffit de définir le problème de façon récursive. Posons W = min E(S), où S comprend toutes les structures secondaires. L'énergie associée l'appariement de s_i à s_j est donné par e(i,j). W(i,j) est calculé pour tous les fragments i..j de la molécule.

```
W(i,j) = 0 pour j-i < 4
W(i,j) = min {W(i+1,j),
              W(i,j-1),
              e(i,j)+W(i+1,j-1),
              min_{k=i+1..j-1}{W(i,k)+W(k+1,j)}}
```

Matrice de programmation dynamique :



4. Exemples

Considérons le modèle simple suivant : les paires A-T et C-G contribuent négativement à la structure globale (-1) et tous les autres éléments structurels n'ont aucun effet sur l'énergie de la structure globale. Sous le modèle de la minimisation de l'énergie libre, prédisez les structures associées aux séquences suivantes :

i. acagattctg

W(i,j)	a	c	a	g	a	t	t	c	t	g
a	0	0	0	0	0	-1	-1	-1	-2	-3
c	0	0	0	0	0	0	-1	-1	-2	-3
a	0	0	0	0	0	0	-1	-1	-1	-2
g	0	0	0	0	0	0	0	-1	-1	-1
a	0	0	0	0	0	0	0	0	-1	-1
t	0	0	0	0	0	0	0	0	0	0
t	0	0	0	0	0	0	0	0	0	0
c	0	0	0	0	0	0	0	0	0	0
t	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0

Structure :

ii. acagattctggctttaagca

W(i,j)	a	c	a	g	a	t	t	c	t	g	g	c	t	t	t	a	a	g	c	a
a	0	0	0	0	0	-1	-1	-1	-2	-3	-3	-3	-4	-4	-4	-4	-4	-5	-6	-6
c	0	0	0	0	0	0	-1	-1	-2	-3	-3	-3	-3	-3	-3	-3	-4	-5	-6	-6
a	0	0	0	0	0	0	-1	-1	-2	-2	-2	-2	-3	-3	-3	-3	-3	-4	-5	-5
g	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-2	-2	-2	-2	-3	-3	-4	-5
a	0	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-2	-3	-4	-4
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-2	-2	-3	-4
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-2	-2	-3	-4
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	-2	-3	-4
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	-2	-3	-4
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-2	-3	-3
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-2	-3	-3
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-2	-2	-2
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	-1	-1
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Structure :