BGPG

M1 MABS
Contrôle terminal de Bioinformatique pour la génomique et postgénomique (EM8BBSCM) - Mai 2013

Problème 1

Un ami qui habite dans une autre région que vous, vous fait part de l'état du sol de son jardin sur une durée de quatre jours. La suite d'observations est la suivante (sec, humide, humide, détrempé).
Vous posséder les informations suivantes :
On considère que le temps peut être ensoleillé, nuageux ou pluvieux.
Le vecteur de probabilité initial est :
ensoleillé : 0.5; nuageux : 0.2; pluvieux : 0.3

Connaissant le temps qu'il a fait hier, nous pouvons prédire le temps qu'il fera aujourd'hui grâce à la matrice des probabilités de transitions suivante :

	Temps aujourd'hui			
		ensoleillé	nuageux	pluvieux
Temps hier	ensoleillé	0.5	0.4	0.1
	nuageux	0.3	0.1	0.6
	pluvieux	0.3	0.4	0.3

On possède trois possibilités pour décrire l'état du sol : sec, humide et détrempé.
Nous possédons la matrice des probabilités d'émissions suivantes :

	observations			
		sec	humide	détrempé
Etats cachés	ensoleillé	0.8	0.15	0.05
	nuageux	0.3	0.5	0.2
	pluvieux	0.05	0.35	0.6

Question 1
Représenter le HMM correspondant en indiquant les différentes probabilités de transitions et d'émissions.

Question 2

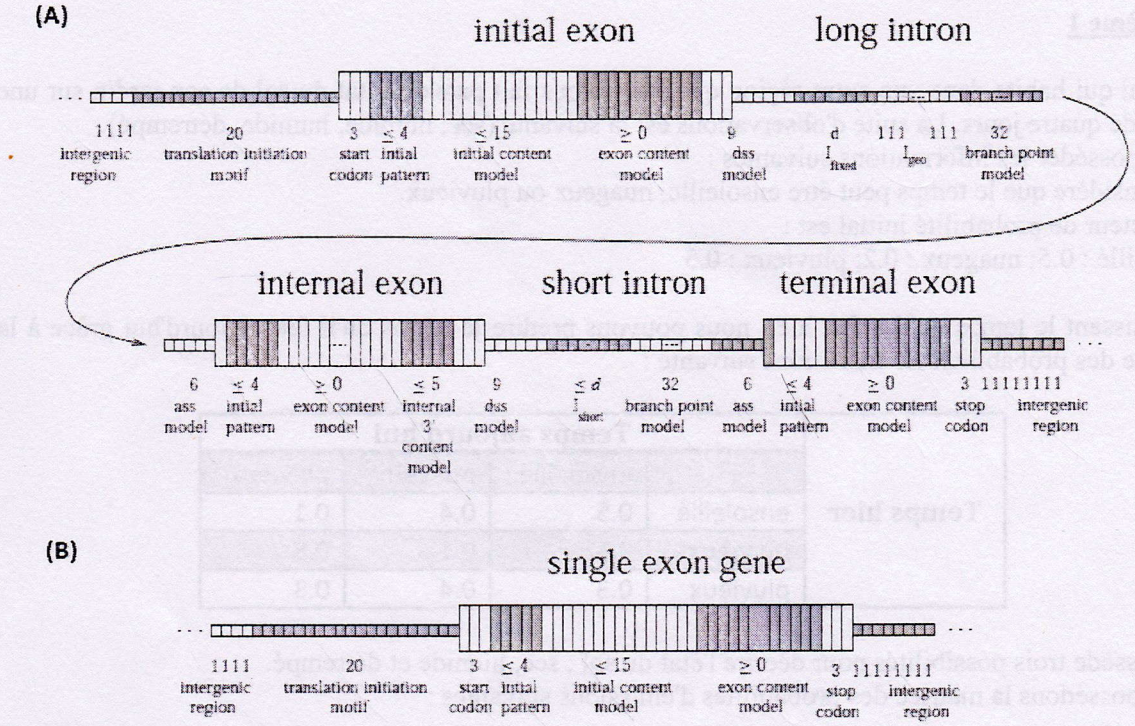
a) Déterminer la séquence d'états cachés la plus probable correspondant à la séquence d'observation (sec, humide, humide, détrempé), *i. e.,* trouver le chemin le plus probable dans le HMM de cette séquence d'observation et donner sa probabilité. Représenter le résultat sous la forme d'un treillis.
b) Calculer, étant donné le HMM, la probabilité de cette séquence d'observation.

Dans les deux cas, vous indiquerez l'algorithme qui doit être utilisé.

Problème 2

Augustus est une autre méthode dédiée à la prédiction *ab initio* de gènes codant pour des protéines dans les génomes eucaryotes. Elle est basée sur un modèle de Markov caché et repose sur le développement d'un nouveau modèle pour la prédiction des introns.
La figure ci-dessous résume les différentes structures qui peuvent être rencontrées dans un gène codant eucaryote avec (A) ou sans intron (B). Certaines parties de l'ADN sont modélisées par des sous-modèles dont le nom et les contraintes sur les longueurs (pour la version chez l'homme) sont indiqués en dessous de ces régions ADN. Attention, cette figure représente un fragment génomique synthétisant

l'ensemble des structures pouvant être rencontrées et non pas un fragment génomique réel, c'est-à-dire que par exemple, après un exon initial ou avant un exon terminal, vous pouvez avoir soit un intron long, soit un intron court. De même, un gène peut être composés de plusieurs exons séparés par l'un ou l'autre type d'intron.



(Figure extraite de Stanke and Waack (2003), *Bioinformatics*, **19**, ii215-ii225).

Les sous-modèles :

- **dss model** - donor splice site model : prend en compte les 3 derniers nucléotides de l'exon, le dinucléotide GT consensus et 4 nucléotides en plus dans l'intron.
- **ass model** - acceptor splice site model : prend en compte 3 nucléotides dans l'intron avant le dinucléotide consensus AG, le dinucléotide AG et le premier nucléotide de l'exon.
- **branch point model** : modélise le site du lasso et prend en compte une région de 32 nucléotides.
- $\mathbf{I}_{short}$: modélise les petits introns dont la taille est $\leq d$.
- $\mathbf{I}_{fixed}$: pour les grands introns, modélise une longueur fixe de d nucléotides.
- $\mathbf{I}_{geo}$: pour les grands introns, modèle qui émet un seul nucléotide.
- **initial pattern** : modélise au plus 4 nucléotides après le codon start ou le modèle de la jonction 3' d'épissage (petite erreur sur le dessin, 5 positions représentées au lieu de 4).
- **intergenic region** : modèle qui émet un seul nucléotide.

Les autres sous-modèles sont suffisamment explicites et ne seront pas détaillés.

1) En utilisant la figure ci-dessus, réaliser le schéma du HMM qui modélisera les différents états et les transitions possibles entre eux sur le brin direct d'une part et indirect d'autre part. Pour la clarté du schéma, les différents sous-modèles des états exons ne seront pas développés dans le schéma général, *i.e.*, par exemple, l'état single exon sera simplement représenté par E_{single}.

2) Pour chacun des états exons, vous détaillerez par la suite le HMM modélisant les différents sous-modèles et ceci pour l'état sur le brin direct et celui sur le brin complémentaire.

3) Parmi les sous-modèles, quels sont ceux qui émettent une sous-séquence de longueur variable ?