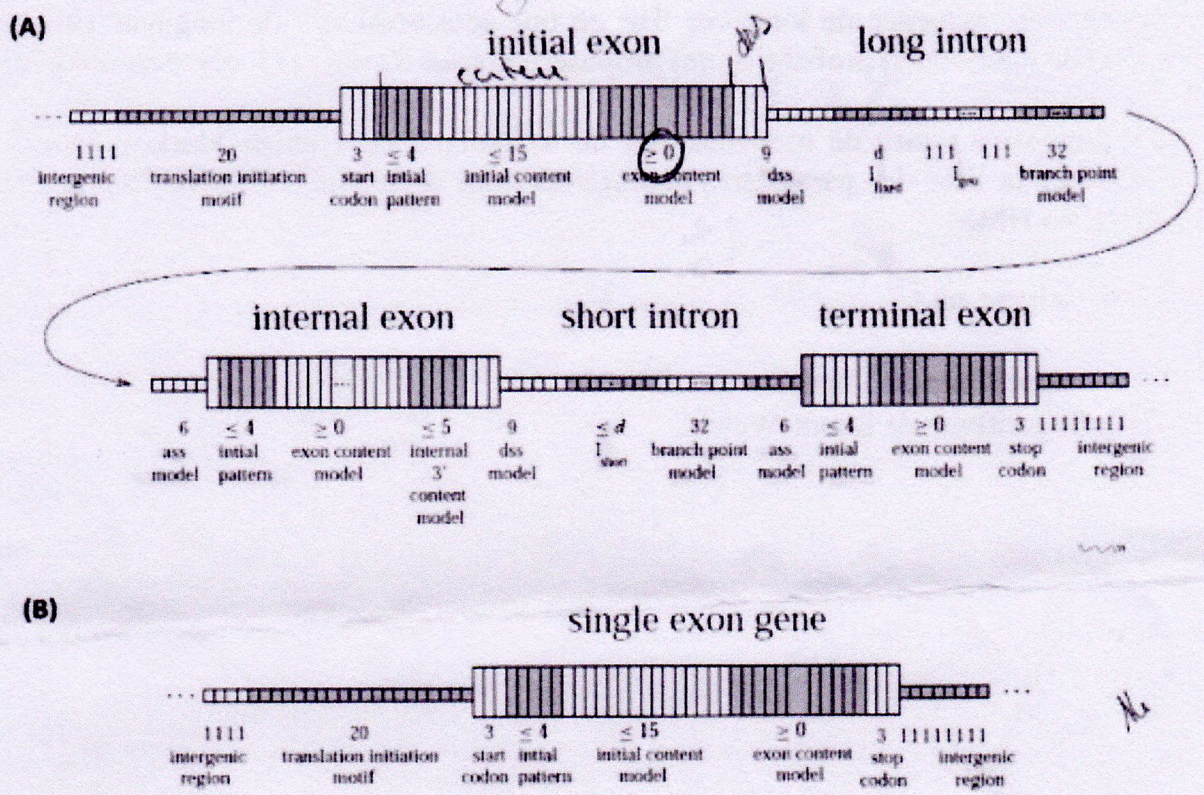
M1 MABS
Contrôle terminal de Bioinformatique pour la génomique et postgénomique (durée 2h)
(EM8BBSCM) - Mai 2012

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.



(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) Certains sous-modèles modélisent un contenu. Ces sous-modèles émettent-ils un seul nucléotide, une sous-séquence de longueur fixe ou une sous-séquence de longueur variable ? Quel(s) type(s) de méthode(s) probabiliste(s) proposeriez-vous d'associer à ces sous-modèles ?

4) Le HMM que vous venez de modéliser est un « Generalized Hidden Markov Model » (GHMM). Donner la liste des paramètres nécessaires pour définir un GHMM. Expliquer la différence avec un HMM.

5) A quelle fin utilise-t-on :

- l'algorithme de Viterbi
- l'algorithme de Baum-Welch