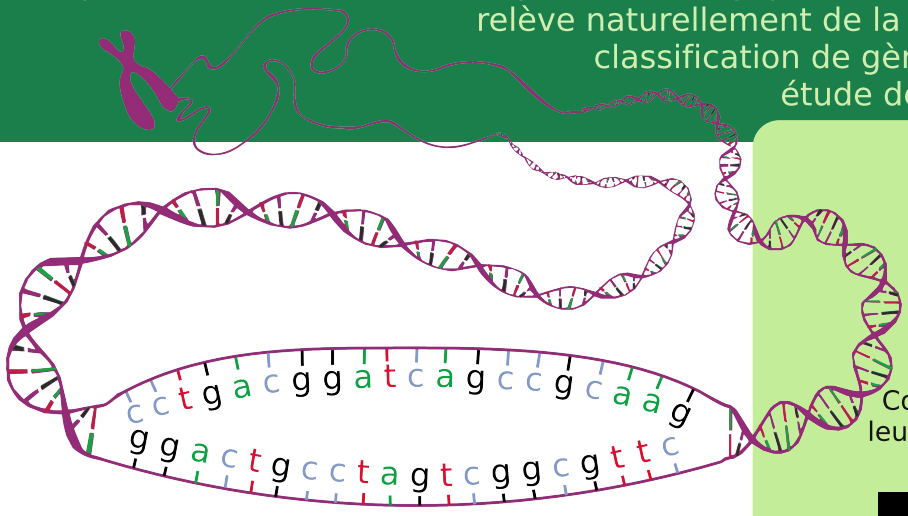
Statistique et Génomique

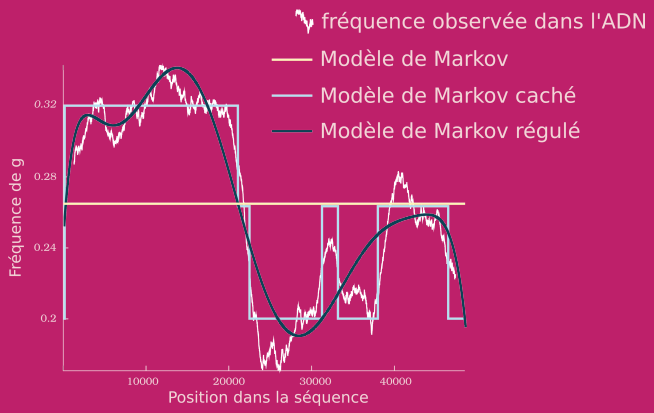
La ***génomique*** s'intéresse à l'information génétique propre à chaque organisme vivant (son génome). Grâce aux progrès technologiques, la ***bioinformatique*** rend désormais disponibles des milliers de séquences biologiques (ADN, ARN, protéines). Leur analyse relève naturellement de la statistique : modélisation, classification de gènes, analyse de leur expression, étude de leurs interactions...



Modélisation des séquences biologiques

Une séquence d'ADN est vue comme un texte long de centaines de milliers de pages, écrit dans un alphabet à 4 lettres $\{a,t,c,g\}$. Certains mots jouent des rôles biologiques importants (coder le début d'un gène, garantir la stabilité du génome...). Ils ont des répartitions particulières.
Un ***modèle statistique*** permet de comparer ce qui est observé sur le génome à une séquence aléatoire, pour détecter des mots significativement rares ou fréquents.

Modèles de Markov



Choisir un modèle consiste ici à définir un processus aléatoire $(X_i)_i$ ressemblant à la séquence d'ADN donnée : X_i représente la lettre à la position i. Les ***modèles markoviens*** permettent de prendre en compte la dépendance entre des lettres successives. Pour tenir compte de l'hétérogénéité de la séquence (succession de régions dites «codantes» et «non codantes»), on utilise une ***chaîne de Markov cachée*** : le texte est divisé en régions homogènes, et la façon dont se succèdent les lettres varie selon la région. Si l'on considère une variation continue de la loi de succession des lettres, on parle de ***chaîne de Markov régulée***.

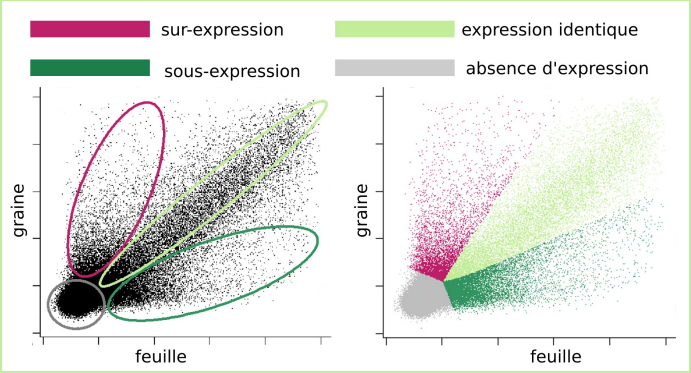
$$\mathbb{P}(X_t \in A \mid X_u, u<t) = \mathbb{P}(X_t \in A \mid X_{t-1}, \dots, X_{t-k})$$

Analyse différentielle de gènes : un problème de classification

Bien que toutes les cellules d'un organisme vivant contiennent le même génome, celui-ci est utilisé différemment selon le rôle de la cellule : c'est l'***expression des gènes***.
Comment classer les parties de l'ADN selon leur expression dans différentes conditions ?

On s'intéresse à l'ADN de la plante *Arabidopsis thaliana*, et à son expression au niveau de la graine et de la feuille.
Un cadre statistique adapté est le ***modèle de mélange gaussien*** à 4 classes : sur ou sous-expression au niveau de la graine par rapport à la feuille, expression identique, ou absence d'expression.
On estime les paramètres (moyennes et matrices de variance), et les probabilités d'appartenance à chaque classe.
La ***règle du maximum a posteriori*** assigne chaque partie de l'ADN à la classe pour laquelle la probabilité d'appartenance est la plus grande.

$$\sum_{j=1}^{4} \pi_j \mathcal{N}\left(\mu_j, \Sigma_j^2\right)$$



LMRS Laboratoire de Mathématiques Raphaël Salem
cnrs
UNIVERSITÉ DE ROUEN NORMANDIE
litis