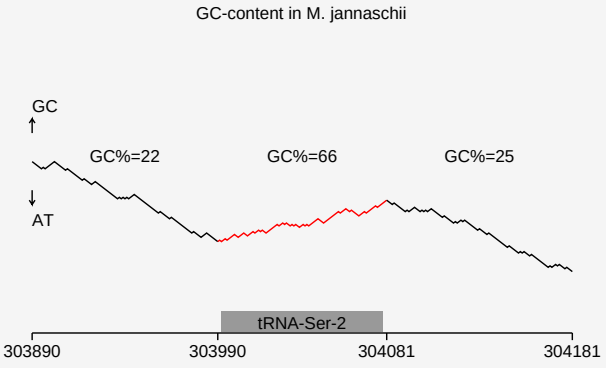
3 Pénalisation de complexité de modèles

3.1 Composition variable et pénalisation de modèles



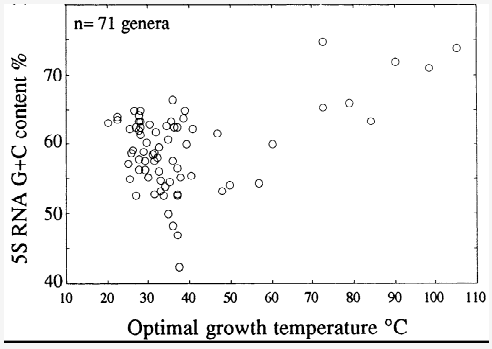
Souvent, on obsèrve une variation de composition au long d'une séquence moléculaire. Par exemple, la fréquence des acides aminés sera différente dans des régions hydrophiles et hydrophobes d'une protéine. Le taux de GC varie aussi au long de génomes. Afin de capturer ce genre de variation, on utilise une notion de *classe* : chaque position appartient à une classe qui détermine la composition.

Modèle probabiliste pour composition variable. On veut un modèle probabiliste pour séquences de longueur n sur un alphabet $\mathcal{A}$. Une séquence $x[0..n-1]$ est l'obsérvation d'une séquence de variables aléatoires $X[0..n-1]$. On assume que la classe $Z[i]$ détermine la distribution de $X[i]$ selon $p_z(x) = \mathbb{P}\Big\{X[i] = x \;\Big|\; Z[i] = z\Big\}$. La distribution p_z est connue dans toute classe z, mais la classification $Z[i]$ est inconnue. Un hypothèse est encodé par une classification particulière $z[0..n-1]$ assumée. Le modèle iid est l'hypothèse null que $z[i] = 0$ à tout i. La vraisemblance d'un hypothèse z s'écrit par

$$L(z) = \mathbb{P}\Big\{X[0..n-1] = x[0..n-1] \;\Big|\; Z[0..n-1] = z[0..n-1]\Big\} = \prod_{k=0}^{n-1} p_{z[i]}\big(x[i]\big). \tag{3.1}$$

La log-vraisemblance se décompose comme

$$\log L(z) = \log \prod_{k=0}^{n-1} p_{z[i]}\big(x[i]\big) = \underbrace{\sum_{i=0}^{n-1} \log p_0(x[i])}_{\text{vrais. de l'hypo null}} + \underbrace{\sum_{i=0}^{n-1} \log \frac{p_{z[i]}(x[i])}{p_0(x[i])}}_{\text{LODS de l'hypo } z} \tag{3.2}$$



Taux de GC chez procaryotes thermophiles. En procaryotes thermophiles et hyperthermophiles (optimum de croissance à 40–100+ °C), les gènes ARN ont un taux GC correlé à la température de croissance. On peut même exploîter cette différence dans la découverte de nouveaux gènes : Klein, Misulovin et Eddy [«Noncoding RNA genes identified in AT-rich hyperthermophiles», *PNAS*, **99** :7542–7547, 2002] ont identifié de nouveaux gènes ARN non-codant par contenu GC, et ont démontré qu'ils sont en fait exprimés.
À la gauche : contenu GC dans gènes 5S et température optimale de croissance [Galtier N et Lobry JR. «Relationships between genomic G+C content, RNA secondary structures, and Optimal Growth Temperature in Prokaryotes», *Journal of Molecuar Evolution*, **44** :632–636, 1997].

Le modèle assumé comprend deux classes : taux GC normale (classe 0) et taux GC élevée (classe 1). On assume aussi les règles de Chargaff : $\pi_{\mathtt{A}} = \pi_{\mathtt{T}}$ et $\pi_{\mathtt{G}} = \pi_{\mathtt{C}}$. En conséquence, il suffit d'encoder la séquence x en binaire par les codes ambigus $\mathtt{W} = \{\mathtt{A}, \mathtt{T}\}$ et $\mathtt{S} = \{\mathtt{C}, \mathtt{G}\}$. Ainsi, un seul paramètre $\phi_i = 2p_i(\mathtt{A})$ définit chaque classe $i = 0, 1$ car $p_i(\mathtt{T}) = \phi_i/2$ et $p_i(\mathtt{C}) = p_i(\mathtt{G}) = (1 - \phi_i)/2$, ou bien $p_i(\mathtt{W}) = \phi_i$ et $p_i(\mathtt{S}) = 1 - \phi_i$. W(fr)

3.2 Pénalisation de modèles.

Selon équation (3.2), l'hypothèse $z[i] = \{x[i] = \mathtt{S}\}$ explique la séquence le mieux, parce qu'il maximise la vraisemblance ($p_0(\mathtt{S}) > p_1(\mathtt{S})$ et $p_0(\mathtt{W}) < p_1(\mathtt{W})$). Clairement, une telle partition est complètement inutile (classification change à trop de positions) quand on cherche des *régions* de taux GC élevée. On devrait choisir le meilleur hypothèse en incluant une mesure de **complexité**, et pénaliser des segmentation avec trop de points de changement.

Déscription minimale. Le principe de déscription minimale (*minimum description length*—MDL), proposé par Jorma Rissannen [«A Universal prior for integers and estimation by minimum description length», *Annals of Statistics*, **11** :416–431, 1983], est qu'on devrait choisir le modèle qui fournit la plus courte déscription jointe. Dans notre exemple, on doit encoder le modèle z et les données x : W(en)

$$\mathsf{encodage}(z, x) = \mathsf{encodage}(x|z) \ \#\ \mathsf{encodage}(z).$$

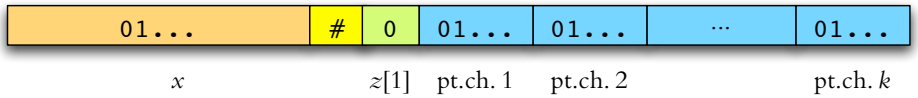
Le meilleur modèle z minimise la longueur de l'encodage complète :

$$\mathsf{MDL}(z) = \big|\mathsf{encodage}(z, x)\big| = \big|\mathsf{encodage}(x|z)\big| + 1 + \big|\mathsf{encodage}(z)\big|.$$

Ici, $\mathsf{encodage}(x|z)$ est l'encodage de la séquence $x[0..n-1]$ étant donnée la partition $z[0..n-1]$. Avec l'encodage Huffman , on encode la valeur $x[i]$ sur $-\lg p_{z[i]}(x[i])$ bits. On voit que la longueur de l'encodage des données est proportionnelle au logarithme de la vraisemblance. W(fr)

$$\big|\mathsf{encodage}(x|z)\big| = -\sum_{k=0}^{n-1} \lg p_{z[i]}\big(x[i]\big) = -\lg L(z)$$

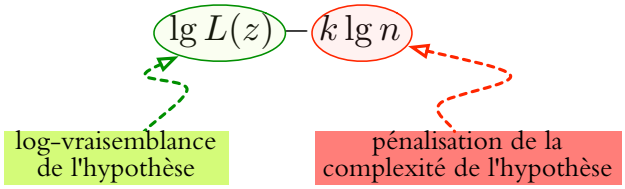
Encodage d'une segmentation. Pour encoder notre hypothèse z, on note que le but est d'identifier des intervalles de taux GC élevé : on cherche une segmentation ou partition avec peu de segments. Pour encoder, il suffit de donner les points de changement $z[i] \neq z[i-1]$. On utilise un bit additionnel pour encoder $z[0]$. Chaque point de changement prend $\lg n$ bits (entier entre 1 et $n-1$). Au total, l'encodage d'une segmentation à k points de changement prend $(1 + k \lg n)$ bits.



Pénalité de complexité de hypothèse. Le meilleur hypothèse donc minimise la longueur de l'encodage

$$\mathsf{MDL}(z) = -\lg L(z) + k \lg n + O(1).$$

Dans d'autres, on cherche à maximiser la log-vraisemblance pénalisée



AIC. Hirotugu Akaike [«A new look at the statistical model identification», *IEEE Transactions on Automatic Control*, **19** :716–723, 1974] a proposé la pénalisation simple par le nombre de paramètres dans le modèle. Dans le cas d'un modèles à k paramètres (ici : k points de changement), le critère d'Akaike (*Akaike's "an Information Criterion"* — AIC) se définit par W(en)

$$\mathsf{AIC}(z) = -2 \ln L(z) + 2k. \tag{3.3}$$

Le meilleur modèle devrait minimiser AIC.

BIC. Selon le critère BIC (*Bayesian Information Criterion* — BIC), suggéré par Gideon Schwarz [«Estimating the dimension of a model», *Annals of Statistics*, **6** :461–464, 1978], on cherche à minimiser W(en)

$$\mathsf{BIC}(z) = -2 \ln L(z) + k \ln n \tag{3.4}$$

où k est le nombre de paramètres et n est la taille des données.

MDL, AIC et BIC proposent la pénalisation d'un modèle à k paramètres par un terme proportionnel à k. On voit qu'en général, $\mathsf{AIC}(z) < \mathsf{BIC}(z) < \mathsf{MDL}(z)$:

log-vraisemblance	pénalité de modèle		
	AIC	BIC	MDL
$\log L(z)$	k	$\frac{k}{2} \log n$	$k \log n$

En pratique, AIC tend à favoriser des modèles trop complèxes (*overfitting*), mais BIC donne une pénalité raisonnable. En tout cas, on cherche à maximiser $\log L(z) - k\alpha$, où α détermine la politique de pénalisation de complexité.

Segmentation optimale. Soit $V_i(k)$ le maximum de la vraisemblance pénalisée pour $z[0..k]$ finissant par $z[k] = i$:

$$V_i(k) = \max_{z[0..k];z[k]=i} \left\{ \sum_{j=0}^{k} \log p_{z[j]}(x[j]) - \alpha \cdot \underbrace{\sum_{j=1}^{k} \{z[j-1] \neq z[j]\}}_{\text{points de changement}} \right\} \tag{3.5}$$

Théorème 3.1. *Soit $V_i(k)$ la meilleure segmentation de préfixe $x[0..k]$ comme défini en* (3.5)*. On a les récurrences suivantes.*

$$V_0(0) = \log p_0\big(x[0]\big) \tag{3.6a}$$
$$V_1(0) = \log p_1\big(x[0]\big) \tag{3.6b}$$
$$V_0(k) = \log p_0\big(x[k]\big) + \max\big\{V_0(k-1), V_1(k-1) - \alpha\big\} \qquad \{k > 0\} \tag{3.6c}$$
$$V_1(k) = \log p_1\big(x[k]\big) + \max\big\{V_1(k-1), V_0(k-1) - \alpha\big\} \qquad \{k > 0\} \tag{3.6d}$$

On peut calculer la meilleure segmentation en utilisant les récurrences de (3.6), dans un algorithme à **programmation dynamique** : calculer et stocker $V_i(k)$ dans l'ordre $k = 0, 1, \ldots$. En évaluant les max en (3.6c) et (3.6d), il faut stocker un bit pour chaque $V_i(k)$ qui montre si le 1er ou 2e terme a été utilisé dans la récurrence. Cela permet de retracer la meilleure segmentation (procédure de *backtracking*) à partir de $\max_i V_i(n-1)$.

W(en)