Analyse statistique de séquences biologiques

Magistère pluridisciplinaire L2 – Premier semestre 2007-2008
Mathématiques et Biologie

Didier Piau et Christelle Melo de Lima

Didier.Piau@ujf-grenoble.fr

melodelc@ujf-grenoble.fr

http://www-fourier.ujf-grenoble.fr/∼dpiau/

Plan (très) sommaire

0. Motivations

1. Modèles indépendants
   Comment calculer. Limitations

2. Modèles de Markov « simples »
   Comment calculer. Limitations

3. Modèles de Markov cachés
   Apprentissage. Estimation. Algorithmes.

Tout au long de ces parties : Quelques applications
   Hétérogénéités des bactéries. Transferts de gènes.
   Détection de gènes procaryotes

4. Extensions variées et conclusion

Quelques buts possibles de l'analyse de séquences

- Identifier les gènes
- Déterminer la fonction de chaque gène, par exemple en le comparant avec d'autres gènes de fonction connue
- Identifier les protéines impliquées dans la régulation d'un gène
- Identifier les répétitions
- Identifier d'autres régions fonctionnelles : origines de réplication, pseudogènes, séquences rendant possible le repliement compact de l'ADN, etc.

Problème / atout La quantité d'information disponible est gigantesque. Donc nécessité de traitements automatiques.

Modèle Outil pour extraire de l'information.

Un bon modèle doit permettre de révéler des caractéristiques fonctionnelles ou structurelles de la séquence.

Attention : on ne prétend pas donner une description exacte de la séquence, même si le modèle doit refléter le plus possible ses caractéristiques. On ne prétend pas non plus décrire la formation de la séquence ni son évolution au cours du temps (mais : plus sur ce point plus tard).

Modélisation

Séquence génomique de longueur n modélisée par une suite de variables aléatoires X_1, X_2, …, X_n avec $X_i \in \mathcal{A}$, et

$$\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$$

ou bien

$$\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}, \mathtt{G}, \mathtt{H}, \mathtt{I}, \mathtt{K}, \mathtt{L}, \mathtt{M}, \mathtt{N}, \mathtt{P}, \mathtt{Q}, \mathtt{R}, \mathtt{S}, \mathtt{T}, \mathtt{V}, \mathtt{W}, \mathtt{Y}\}.$$

Plus généralement, phénomène aléatoire : X_n est l'observation au temps n.

Qu'est-ce qu'une variable aléatoire ?

On commence par se donner un espace de probabilité $(\Omega, \mathbb{P})$ assez gros pour faire toutes les mesures/expériences qui nous intéressent (et on n'en parle plus–un théorème de mathématiciens nous assure que Ω existe dans les cas qui nous intéressent).

Une variable aléatoire est une fonction $X : \Omega \to \mathcal{A}$. Elle est décrite par les nombres $p(x) = \mathbb{P}(X = x)$ pour tout $x \in \mathcal{A}$.

Donc $p(x) \geqslant 0$ et $\displaystyle\sum_{x\in\mathcal{A}} p(x) = 1$.

La collection $(p(x))_{x\in\mathcal{A}}$ s'appelle la **loi** de X ou la **distribution** de X.

Les mathématiciens notent $\mathbb{P}_X = \displaystyle\sum_{x\in\mathcal{A}} p(x)\,\delta_x$.

En pratique : la loi de X donne $\mathbb{P}(X \in B)$ pour tout $B \subset \mathcal{A}$ et permet de calculer des **moyennes**.

Exemple : Pour calculer un taux de $\mathtt{gc}$, $B = \{\mathtt{g}, \mathtt{c}\}$ et

$$\mathbb{P}(X \in B) = p(\mathtt{g}) + p(\mathtt{c}).$$

Le grand principe :

« Tout se calcule à partir de la loi. »

Donc, si X_1 et X_2 ont séparément la même loi, elles sont indistinguables, considérées séparément, puisque, pour tout $B \subset \mathcal{A}$,

$$\mathbb{P}(X_1 \in B) = \mathbb{P}(X_2 \in B).$$

Par contre, les lois de X_1 et X_2 ne suffisent pas à connaître la loi de la variable aléatoire $Y = (X_1, X_2)$.

Exemple : sur $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, supposons que les 4 variables aléatoires $X_1 : \Omega \to \mathcal{A}$, $X_2 : \Omega \to \mathcal{A}$, $X_1' : \Omega \to \mathcal{A}$ et $X_2' : \Omega \to \mathcal{A}$ ont la distribution uniforme. Donc $p(x) = \frac{1}{4}$ pour tout $x \in \mathcal{A}$ et pour $X = X_1, X_2, X_1'$ ou X_2'.

Supposons que X_1 et X_1' décrivent un même site, donc $X_1 = X_1'$. Par contre, X_2 et X_2' décrivent deux sites complémentaires, donc $X_2 = \mathtt{a}$ si $X_2' = \mathtt{t}$, $X_2 = \mathtt{c}$ si $X_2' = \mathtt{g}$, etc.

Alors $Y_1 = (X_1, X_1')$ et $Y_2 = (X_2, X_2')$ sont deux variables aléatoires à valeurs dans $\mathcal{A} \times \mathcal{A}$ qui n'ont pas la même loi, puisque si $D = \{(x, x') \in \mathcal{A} \times \mathcal{A}\,;\, x = x'\}$,

$$\mathbb{P}(Y_1 \in D) = 1, \quad \mathbb{P}(Y_2 \in D) = 0.$$

Conséquence : une loi « conjointe » donne plus d'informations que toutes les lois « marginales ».

Loi conjointe

Si $X_1, \ldots, X_n : \Omega \to \mathcal{A}$, on se donne $\mathbb{P}(X_{1:n} = x_{1:n})$ pour tout $x_{1:n} \in \mathcal{A}^n$.

Notation : $X_{1:n} = (X_1, X_2, \ldots, X_n)$ et $x_{1:n} = (x_1, x_2, \ldots, x_n)$ donc $X_{1:n} = x_{1:n}$ signifie que $X_k = x_k$ pour tout $1 \leqslant k \leqslant n$.

Conséquence : on se donne $|\mathcal{A}|^n$ nombres $p(x_{1:n})$ positifs ou nuls et de somme 1.

Lois marginales : ce sont les lois de chacune des variables aléatoires X_k prise séparément.

La dépendance la plus simple entre les X_k : aucune !

Le modèle M00

Chaque X_n vaut x avec la même probabilité pour chaque valeur possible de x dans $\mathcal{A}$ et chaque X_n est indépendant des autres X_k pour $k \neq n$. Donc, pour tout $n \geqslant 1$ et tout $x_{1:n}$,

$$\mathbb{P}(X_{1:n} = x_{1:n}) = \frac{1}{|\mathcal{A}|^n}.$$

La propriété d'indépendance signifie que

$$\mathbb{P}(X_{n_1} \in A_1,\, X_{n_2} \in A_2, \ldots, X_{n_k} \in A_k) \;= \\ \mathbb{P}(X_{n_1} \in A_1)\, \mathbb{P}(X_{n_2} \in A_2) \ldots \mathbb{P}(X_{n_k} \in A_k),$$

pour tous k, n_i et A_i.

Avantages : calculs faciles et beaux théorèmes.

Exemple : pour toute partie $B \subset \mathcal{A}^n$,

$$\mathbb{P}(X_{1:n} \in B) = \frac{|B|}{|\mathcal{A}|^n}.$$

Une question récurrente :

> **« Dans une longue séquence $X_{1:n}$ décrite par le modèle M00, que peut-on dire de la proportion de `a` ? »**

Notation : fonction indicatrice $\mathbf{1}(B)$

$$\mathbf{1}(B) = 1 \text{ si } B \text{ est vrai}, \quad \mathbf{1}(B) = 0 \text{ sinon.}$$

Comptage et proportion :

$$N_n(x) = \sum_{k=1}^{n} \mathbf{1}(X_k = x), \quad R_n(x) = N_n(x)/n.$$

Loi exacte (pas intéressante) :

$$\mathbb{P}(N_n(x) = k) = C_n^k \frac{3^{n-k}}{4^n}, \quad 0 \leqslant k \leqslant n.$$

Approximation (plus intéressante) :

$$R_n(x) \to \frac{1}{4} \quad \text{quand } n \text{ devient grand.}$$

(Voir plus tard.)

Donc :

Si les proportions observées sur une longue séquence d'ADN s'éloignent nettement de 25%, 25%, 25% et 25%, problème !

Exemple : le génome d'*Escherichia coli* comporte $4.6 - 5.4$ Mb et

$$\%(\mathtt{a}) = 23.66,\ \%(\mathtt{g}) = 27.89,\ \%(\mathtt{c}) = 25.30,\ \%(\mathtt{t}) = 23.15.$$

On peut montrer que ce sont des écarts trop grands sous M00 (voir plus tard).

Le modèle M0

On garde l'indépendance mais à présent,

$$\mathbb{P}(X_n = x) = p(x)$$

pour des nombres $p(x) \geqslant 0$ avec $\sum_{x \in \mathcal{A}} p(x) = 1$.

Vocabulaire : $(p(x))_{x \in \mathcal{A}}$ s'appelle la loi ou la distribution des X_n.

Formule :

$$\mathbb{P}(X_{1:n} = x_{1:n}) = p(x_1)\, p(x_2) \ldots p(x_n) = \prod_{x \in \mathcal{A}} p(x)^{N_n(x)}.$$

Théorème (Loi des grands nombres) :
Quand n devient grand, $R_n(x) \to p(x)$ pour chaque $x \in \mathcal{A}$, par exemple au sens où, pour tout $\varepsilon > 0$,

$$\mathbb{P}(|R_n(x) - p(x)| \geqslant \varepsilon) \to 0.$$

Preuve : (assez) facile et utilise des notions que l'on retrouvera plus tard. On va calculer l'espérance et la variance de $R_n(x)$.

Rappel : si Y prend la valeur réelle y avec probabilité $\pi(y)$, on note $\mathbb{E}(Y)$ l'espérance (la moyenne) de Y, c'est-à-dire

$$\mathbb{E}(Y) = \sum_y \pi(y)\, y.$$

À savoir :

(1) Linéarité

$$\mathbb{E}(a_1 Y_1 + a_2 Y_2) = a_1 \mathbb{E}(Y_1) + a_2 \mathbb{E}(Y_2).$$

(2) Comparaisons : si $Y_1 \geqslant Y_2$, alors $\mathbb{E}(Y_1) \geqslant \mathbb{E}(Y_2)$.

(3) Espérance et indépendance : si Y_1 et Y_2 sont indépendantes,

$$\mathbb{E}(Y_1 Y_2) = \mathbb{E}(Y_1)\, \mathbb{E}(Y_2).$$

(4) Variance :

$$\mathrm{var}(Y) = \mathbb{E}([Y - \mathbb{E}(Y)]^2).$$

On a aussi : $\mathrm{var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$.

Remarque : la variance de Y mesure la dispersion de Y autour de sa moyenne $\mathbb{E}(Y)$. Une petite variance signifie une petite dispersion. D'ailleurs :

Inégalité de Bienaymé-Tchebychev :

$$\mathbb{P}(|Y - \mathbb{E}(Y)| \geqslant \varepsilon) \leqslant \frac{\mathrm{var}(Y)}{\varepsilon^2}.$$

Retour à la loi des grands nombres : on va calculer l'espérance et la variance de $R_n(x)$.

1) Espérance :

$$\mathbb{E}(N_n(x)) = \mathbb{E}\left(\sum_k \mathbf{1}(X_k = x)\right) = \sum_k \mathbb{E}(\mathbf{1}(X_k = x)) = n\, p(x).$$

2) Variance :

$$\begin{aligned}
\mathbb{E}(N_n(x)^2) &= \mathbb{E}\left(\sum_k \mathbf{1}(X_k = x) + \sum_{k \neq \ell} \mathbf{1}(X_k = X_\ell = x)\right) \\
&= \sum_k \mathbb{E}(\mathbf{1}(X_k = x)) + \sum_{k \neq \ell} \mathbb{E}(\mathbf{1}(X_k = X_\ell = x)) \\
&= n\, p(x) + n\,(n-1)\, p(x)^2.
\end{aligned}$$

Donc $\mathrm{var}(N_n(x)) = n\, p(x)\,(1 - p(x))$.

3) Conclusion :

Donc $\mathbb{E}(R_n(x)) = p(x)$ et $\mathrm{var}(R_n(x)) = p(x)\,(1 - p(x))/n$.

Il reste à appliquer Bienaymé-Tchebychev :

$$\mathbb{P}(|R_n(x) - p(x)| \geqslant \varepsilon) \leqslant \frac{p(x)\,(1 - p(x))}{n\, \varepsilon^2} \leqslant \frac{1}{4n\, \varepsilon^2} \to 0.$$

Conclusion de la présentation théorique du modèle :

Le modèle M0 tient compte de la composition en chacun des nucléotides (par exemple), c'est-à-dire en chacune des lettres de $\mathcal{A}$. (Et c'est tout, voir plus bas !)

Conséquences :

1) Estimateur du maximum de vraisemblance.

2) Fréquences des mots

Estimateur du maximum de vraisemblance

Rappels : EMV

Dans les situations concrètes, on dispose d'observations : c'est la séquence $x_{1:n}$; et on cherche le meilleur modèle dans une classe donnée, pour rendre compte de cette séquence.

Si la classe de modèles consiste en les lois $\mathbb{P}^\vartheta$ pour les paramètres $\vartheta \in \Theta$, une option est de recourir à l'estimateur du maximum de vraisemblance.

La vraisemblance de la suite d'observations $x_{1:n}$ sous le modèle $\mathbb{P}^\vartheta$ est

$$V(\vartheta) = \mathbb{P}^\vartheta(X_{1:n} = x_{1:n}).$$

L'estimateur du maximum de vraisemblance consiste à choisir la valeur de ϑ qui maximise $V(\vartheta)$, soit

$$\widehat{\vartheta} \leftarrow \max_\vartheta V(\vartheta).$$

Si la classe est M0, ϑ correspond aux poids $p = (p(x))_{x\in\mathcal{A}}$ et on peut tout calculer !

On veut maximiser

$$\log V(p) = \sum_{x\in\mathcal{A}} N_n(x)\, \log p(x),$$

sous la contrainte

$$\sum_x p(x) = 1.$$

Rappel : le principe des extrema liés

Si on veut trouver les points y dans $\mathbb{R}$ où la fonction $\varphi(y)$ est extrémale, on résoud $\varphi'(y) = 0$. Si on veut trouver les points y dans $\mathbb{R}^n$ où la fonction $\Phi(y)$ est extrémale, on résoud $\operatorname{grad}\Phi = 0$, où $\operatorname{grad}\Phi(y)$ est le vecteur gradient de Φ au point y, soit

$$\operatorname{grad}\Phi(y) = \left(\frac{\partial\Phi}{\partial y_i}\right)_{1\leqslant i\leqslant n}.$$

Si maintenant y est soumis à la contrainte $C(y) = 0$, le principe des extrema liés affirme que, si y est un point où $\Phi(y)$ soumise à la contrainte $C(y) = 0$ est extrémale, alors les gradients de Φ et de C en y sont proportionnels. Donc, il existe un nombre réel λ indépendant de $1 \leqslant i \leqslant n$, tel que

$$\frac{\partial\Phi}{\partial y_i} = \lambda\,\frac{\partial C}{\partial y_i}, \quad 1 \leqslant i \leqslant n.$$

Si on est soumis à plusieurs contraintes $C_1(y) = \cdots = C_k(y) = 0$, il existe k nombres réels λ_1, …, λ_k tels que

$$\frac{\partial\Phi}{\partial y_i} = \lambda_1\,\frac{\partial C_1}{\partial y_i} + \cdots + \lambda_k\,\frac{\partial C_k}{\partial y_i}, \quad 1 \leqslant i \leqslant n.$$

Le nombre réel λ, ou les nombres réels λ_1, …, λ_k, s'appellent les multiplicateurs de Lagrange du problème d'extrema liés.

Retour au cas M0

Ici, $\Phi(p) = \log V(p)$ et $C(p) = \sum_x p(x) - 1$, on calcule

$$\frac{\partial \log V}{\partial p(x)} = \frac{N_n(x)}{p(x)}, \quad \frac{\partial C}{\partial p(x)} = 1.$$

Les extrema liés signifient que $N_n(x)/p(x)$ ne dépend pas de x, donc $p(x) = N_n(x)/\lambda$. Comme la somme des $p(x)$ vaut 1, on obtient le résultat suivant, assez logique somme toute.

L'estimateur du maximum de vraisemblance de $(p(x))_{x\in\mathcal{A}}$ dans le modèle M0 pour la séquence $x_{1:n}$ est donné par

$$\widehat{p}(x) = \frac{N_n(x)}{n}, \quad x \in \mathcal{A}.$$

Une partie du cours va être consacrée à des généralisations de ce résultat.

Fréquences des mots

L'ensemble de tous les mots $\mathcal{A}^*$ est la réunion des $\mathcal{A}^n$ pour tout $n \geqslant 1$. Pour des raisons techniques, on se donne aussi un mot vide noté $*$.

La longueur d'un mot $\mathtt{w} \in \mathcal{A}^*$ est $|\mathtt{w}| = n$ si $\mathtt{w} \in \mathcal{A}^n$. La longueur du mot vide est $|*| = 0$.

Estimation pour le modèle M0

Estimateur du maximum de vraisemblance (EMV)

La vraisemblance de la suite d'observations $x_{1:n}$ sous le modèle $\mathbb{P}^\vartheta$ est

$$V(\vartheta) = \mathbb{P}^\vartheta(X_{1:n} = x_{1:n}).$$

L'estimateur du maximum de vraisemblance consiste à choisir la valeur de ϑ qui maximise $V(\vartheta)$, soit

$$\widehat{\vartheta} \leftarrow \max_\vartheta V(\vartheta).$$

Si la classe est M0, ϑ correspond aux poids $p = (p(x))_{x\in\mathcal{A}}$ et on peut tout calculer.

L'estimateur du maximum de vraisemblance de $(p(x))_{x\in\mathcal{A}}$ dans le modèle M0 pour la séquence $x_{1:n}$ est donné par

$$\widehat{p}_n(x) = \frac{N_n(x)}{n}, \quad x \in \mathcal{A}.$$

Conséquence : l'EMV est consistant. Quand $n \to \infty$,

$$\widehat{p}_n(x) \to p(x).$$

Deux prolongements :

- Mesurer la taille de l'erreur $|\widehat{p}_n(x) - p(x)|$.
- Généraliser ce résultat aux mots.

Taille de l'erreur pour le modèle M0

Résultat théorique : le théorème central limite
Outil : la variance (déjà vue)

Rappel :

$$\mathbb{E}(N_n(x)) = np(x), \quad \mathrm{var}(N_n(x)) = np(x)(1-p(x))$$

Donc

$$\mathbb{E}(R_n(x)) = p(x), \quad \mathrm{var}(R_n(x)) = p(x)\,(1-p(x))/n$$

Théorème central limite Quand n est grand, $R_n(x)$ ressemble à une variable aléatoire gaussienne de même moyenne et de même variance :

$$R_n(x) \approx \mathcal{N}(m_x, \sigma_x^2/n), \quad m_x := p(x),\ \sigma_x^2 := p(x)\,(1-p(x)).$$

Par exemple :

$$\mathbb{P}(R_n(x) \geqslant t) \approx \mathbb{P}(\mathcal{N}(m_x, \sigma_x^2/n) \geqslant t).$$

Rappel : $\mathcal{N}(m, \sigma^2) = m + \sigma\mathcal{N}(0,1)$. Et

$$\mathbb{P}(a \leqslant \mathcal{N}(0,1) \leqslant b) = \int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x$$

Donc

$$\boxed{\mathbb{P}\left(R_n(x) \geqslant m_x + t\frac{\sigma_x}{\sqrt{n}}\right) \approx \mathbb{P}(\mathcal{N}(0,1) \geqslant t)}$$

Idem pour « inférieur ou égal »

Table de quantiles de la loi normale centrée réduite

$$\mathbb{P}(m - \sigma \leqslant \mathcal{N}(m,\sigma) \leqslant m + \sigma) = 68.26\%$$

$$\mathbb{P}(m - 2\sigma \leqslant \mathcal{N}(m,\sigma) \leqslant m + 2\sigma) = 95.44\%$$

$$\mathbb{P}(m - 3\sigma \leqslant \mathcal{N}(m,\sigma) \leqslant m + 3\sigma) = 99.74\%$$

À retenir :

Plus de ± 2 fois l'écart-type : 5%
Plus de ± 3 fois l'écart-type : 1%
Et pour M0, l'écart-type est en $1/\sqrt{n}$

Approximation : pour t positif pas trop petit,

$$\mathbb{P}(\mathcal{N}(0,1) \geqslant t) \approx \exp(-t^2/2)$$

Donc, pour $t > p(x)$ et t pas trop proche de $p(x)$,

$$\mathbb{P}(R_n(x) \geqslant t) \approx \exp\left[-\frac{n(t-p(x))^2}{2p(x)(1-p(x))}\right]$$

Idem pour $t < p(x)$

Fréquences des mots

L'ensemble de tous les mots $\mathcal{A}^*$ est la réunion des $\mathcal{A}^n$ pour tout $n \geqslant 1$. Pour des raisons techniques, on se donne aussi un mot vide noté $*$.

La longueur d'un mot $\mathtt{w} \in \mathcal{A}^*$ est $|\mathtt{w}| = n$ si $\mathtt{w} \in \mathcal{A}^n$. La longueur du mot vide est $|*| = 0$.

La loi des grands nombres ci-dessus affirme que, dans le modèle M0, $R_n(\mathtt{w}) \to p(\mathtt{w})$ pour tout mot $\mathtt{w}$ de longueur 1. En fait :

> Pour toute longueur $L \geqslant 1$ et tout mot $\mathtt{w} \in \mathcal{A}^*$ de longueur L, dans le modèle M0, quand n devient grand,
> $$R_n(\mathtt{w}) \to p(\mathtt{w}) \text{ avec } p(\mathtt{w}) = p(w_1) \cdots p(w_L).$$

La preuve de ce résultat est omise : sur le principe, on reprend la preuve du cas d'une lettre, donc on montre que

1) $\mathbb{E}(N_n(\mathtt{w})) = n\, p(\mathtt{w})$,
2) $\mathrm{var}(N_n(\mathtt{w}))$ se comporte comme un multiple de n,
3) on conclut par Bienaymé-Tchebychev.

Comme pour les lettres, on peut quantifier :

> $\mathrm{var}(N_n(\mathtt{w})) \sim n\sigma_{\mathtt{w}}^2$ et $\sigma_{\mathtt{w}}^2$ est calculable
> $R_n(\mathtt{w}) \approx p(\mathtt{w}) + \mathcal{N}(0,1)\sigma_{\mathtt{w}}/\sqrt{n}$

$$\mathbb{P}(N_n(\mathtt{w}) \geqslant np(\mathtt{w}) + 2\sigma_{\mathtt{w}}\sqrt{n}) \approx 2.28\%$$

Exemples : $\mathtt{w} = AA$, $\mathtt{w} = AT$. Calculer $\sigma_{\mathtt{w}}^2$.

Validation ou rejet de M0

On compte les lettres, on en déduit des valeurs plausibles de $p(x)$: on choisit le modèle M0 du maximum de vraisemblance. Mais on voudrait rejeter une séquence

$$AACCGGTTAACCTTGGCCAAAATTGG\ldots$$

Sou M0, le fréquence d'un mot $\mathtt{w} = w_1w_2$ doit vérifier

$$R_n(\mathtt{w}) \approx p(\mathtt{w}) = p(w_1)\,p(w_2),\ R_n(w_1) \approx p(w_1),\ R_n(w_2) \approx p(w_2),$$

donc

$$\frac{N_n(\mathtt{w})}{n} = R_n(\mathtt{w}) \approx R_n(w_1)\,R_n(w_2) = \frac{N_n(w_1)\,N_n(w_2)}{n^2}.$$

> $R_n(w_1w_2) \approx R_n(w_1)\,R_n(w_2)$
> avec une erreur de l'ordre de $1/\sqrt{n}$

A contrario, si $n\,N_n(w_1w_2)$ est très différent de $N_n(w_1)\,N_n(w_2)$, le modèle M0 n'est pas pertinent !

Exemple : ADN d'*E. coli*. (Voir transparent.)

Que faire ? Réponse : les chaînes de Markov (à suivre).

Le modèle M1

À présent, les positions successives X_n ne sont plus indépendantes. On commence par le cas le plus simple : la distribution de X_n est influencée par la valeur de X_{n-1}. Les biologistes parlent de modèle M1, les mathématiciens de chaînes de Markov.

Définition Si $\mathbb{P}(B_2) \neq 0$, la probabilité conditionnelle de B_1 sachant B_2 est

$$\mathbb{P}(B_1|B_2) = \frac{\mathbb{P}(B_1 \cap B_2)}{\mathbb{P}(B_2)}.$$

Commentaire Cela correspond à l'intuition : considérons le cas où B_1 = je suis en retard en cours, B_2 = il neige. On peut penser que, si B_2, la circulation dans l'agglomération grenobloise devient plus difficile, donc B_1 a plus de chances d'être réalisé. Il vaudrait mieux évaluer B_1 par une probabilité éventuellement différente de $\mathbb{P}(B_1)$, qui rende compte du fait que B_2 est réalisé (il neige) : cette nouvelle valeur, c'est $\mathbb{P}(B_1|B_2)$.

Définition

La suite $X_{1:n}$ est une chaîne de Markov si, pour tout $1 \leqslant k \leqslant n-1$ et tout $x_{1:k+1}$,
$\mathbb{P}(X_{k+1} = x_{k+1}|X_{1:k} = x_{1:k}) = \mathbb{P}(X_{k+1} = x_{k+1}|X_k = x_k)$.

On parle aussi de mémoire à distance 1 : si on s'intéresse à la position $k+1$, on peut oublier les valeurs aux positions $1 : k-1$ et ne garder que la position k.

Un calcul facile montre alors que

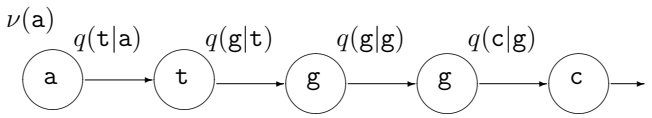
$$\mathbb{P}(X_{1:k} = x_{1:k}) = \mathbb{P}(X_1 = x_1)\times$$
$$\times\, \mathbb{P}(X_2 = x_2|X_1 = x_1) \times \cdots \times \mathbb{P}(X_k = x_k|X_{k-1} = x_{k-1}).$$

On voit que la loi d'une chaîne de Markov (stationnaire, si on veut préciser) est décrite complètement dès qu'on connaît $\mathbb{P}(X_1 = x)$ pour tout $x \in \mathcal{A}$ et $\mathbb{P}(X_k = x'|X_{k-1} = x)$ pour tout $(x, x') \in \mathcal{A} \times \mathcal{A}$. On note

$$\nu(x) = \mathbb{P}(X_1 = x), \quad q(x'|x) = \mathbb{P}(X_k = x'|X_{k-1} = x).$$

On note aussi $q(x, x') = q(x'|x)$ (attention : l'ordre de x et x' change !).

Le modèle M1



Par exemple,

$$\mathbb{P}(X_{1:5} = \mathtt{atggc}) = \nu(\mathtt{a})\, q(\mathtt{t}|\mathtt{a})\, q(\mathtt{g}|\mathtt{t})\, q(\mathtt{g}|\mathtt{g})\, q(\mathtt{c}|\mathtt{g}).$$

Paramètres de la chaîne de Markov :

- Loi initiale ν : $\nu(x) \geqslant 0$ pour tout $x \in \mathcal{A}$ et $\sum_{x\in\mathcal{A}} \nu(x) = 1$.
- Matrice de transition q : $q(x, x') \geqslant 0$ pour tous x et $x' \in \mathcal{A}$ et, pour tout $x \in \mathcal{A}$,

$$\sum_{x'\in\mathcal{A}} q(x, x') = 1.$$

Donc $0 \leqslant \nu(x) \leqslant 1$ et $0 \leqslant q(x, x') \leqslant 1$.

$$\nu = (\nu(\mathtt{a}), \nu(\mathtt{c}), \nu(\mathtt{g}), \nu(\mathtt{t})),$$

et, dans l'ordre $\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}$,

$$q = \begin{pmatrix} q(\mathtt{a},\mathtt{a}) & q(\mathtt{a},\mathtt{c}) & q(\mathtt{a},\mathtt{g}) & q(\mathtt{a},\mathtt{t}) \\ q(\mathtt{c},\mathtt{a}) & q(\mathtt{c},\mathtt{c}) & q(\mathtt{c},\mathtt{g}) & q(\mathtt{c},\mathtt{t}) \\ q(\mathtt{g},\mathtt{a}) & q(\mathtt{g},\mathtt{c}) & q(\mathtt{g},\mathtt{g}) & q(\mathtt{g},\mathtt{t}) \\ q(\mathtt{t},\mathtt{a}) & q(\mathtt{t},\mathtt{c}) & q(\mathtt{t},\mathtt{g}) & q(\mathtt{t},\mathtt{t}) \end{pmatrix}.$$

Une chaîne de Markov (un modèle M1) est un processus sans mémoire (autre que celle de sa valeur actuelle). Rappel :

$$\mathbb{P}(X_{n+1} = x | X_{1:n} = x_{1:n}) = \mathbb{P}(X_{n+1} = x | X_n = x_n).$$

Cas ADN : M1 donne une meilleure approximation de la réalité que le modèle indépendant (M0). Mais en fait, bien sûr, les dépendances sont encore plus complexes.

On utilisera très vite des dépendances à m pas (modèles Mm). Le principe reste le même donc on va décrire M1.

Quelques remarques

• On peut utiliser les modèles pour une séquence génomique donnée. Alors $q(x, x')$ donne la probabilité que le site $n + 1$ soit occupé par un x' sachant que le site n soit occupé par un x, i.e. n est un indice **spatial**.

On peut aussi utiliser n comme un indice **temporel**.

Donc X_n est le nucléotide en un site donné après n réplications de la molécule d'ADN et (par exemple) les sites évoluent indépendamment les uns des autres.

On peut penser qu'il y a eu beaucoup de réplications donc on s'intéressera à la distribution quand n devient grand.

Deux exemples classiques de modèles M1 d'évolution :

On s'intéresse à un site fixé et on suppose qu'il évolue indépendamment du reste de la séquence (ce qui est tout à fait faux, biologiquement ! !).

Jukes-Cantor Pour tous $x \neq x'$, $q_{\mathrm{JC}}(x, x') = p$ avec $0 \leqslant p \leqslant \frac{1}{3}$.

$$q_{\mathrm{JC}} = \begin{pmatrix} 1-3p & p & p & p \\ p & 1-3p & p & p \\ p & p & 1-3p & p \\ p & p & p & 1-3p \end{pmatrix}.$$

Le paramètre p dépend de l'échelle de temps considérée (voir plus loin).

Kimura Purines $\mathtt{a}, \mathtt{g}$ vs. pyrimidines $\mathtt{c}, \mathtt{t}$.

Pour chaque transition, probabilité u. Pour chaque transversion, probabilité v. Donc $0 \leqslant u + 2v \leqslant 1$. Dans l'ordre $\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}$,

$$q_{\mathrm{K}} = \begin{pmatrix} 1-u-2v & v & u & v \\ v & 1-u-2v & v & u \\ u & v & 1-u-2v & v \\ v & u & v & 1-u-2v \end{pmatrix}.$$

Même remarque que pour J-C.

Hasegawa, Kishino, Yano, autres...

Description du modèle M1

Lois = calcul matriciel !

Théorème
Dans un modèle M1, la distribution après n pas vaut νq^n.

Preuve :

$$\begin{aligned} \mathbb{P}(X_{n+1} = x) &= \sum_{x_{1:n} \in \mathcal{A}^n} \mathbb{P}(X_{1:n+1} = x_{1:n} x) \\ &= \sum_{x_{1:n} \in \mathcal{A}^n} \nu(x_1) \prod_{i=2}^{n} q(x_{i-1}, x_i)\, q(x_n, x) \\ &= (\nu q^n)(x). \end{aligned}$$

Rappel : $(MN)(x, y) = \sum_z M(x, z) N(z, y)$.

Par exemple,

$$\mathbb{P}(X_4 = \mathtt{g} | X_1 = \mathtt{a}) = \sum_{x=\mathtt{a}}^{\mathtt{t}} \sum_{z=\mathtt{a}}^{\mathtt{t}} q(\mathtt{a}, x)\, q(x, z)\, q(z, \mathtt{g}).$$

Problème : comment calculer la distribution de X_{101} ?
Additionner $4^{100} = 2^{200} \approx 10^{60}$ termes ? ? ?

Premier principe des processus M1
Convergence vers un équilibre (stochastique).

C'est-à-dire :

(1) Les νq^n varient avec n, on sent l'effet de l'âge.

(2) Chaque νq^n dépend de ν, on se souvient de son état initial.

(3) Mais tout ceci disparaît quand n devient grand, on finit par tout oublier. Convergence vers l'équilibre : si n est grand,

$$\mathbb{P}_\nu(X_n = x) \approx \pi(x).$$

Remarque : $\pi(x)$ est indépendant de n et de ν.

Que vaut π ? C'est un exemple de distribution stationnaire.

Distribution stationnaire : $\mu q = \mu$.

Si on converge, c'est vers une distribution stationnaire.

Le résultat
• Hypothèses techniques : chaîne de Markov finie, irréductible, apériodique.
• Alors la distribution stationnaire π existe et est unique et pour toute loi initiale ν,

$$\mathbb{P}_\nu(X_n = x) \to_{n\to\infty} \pi(x).$$

L'hypothèse d'apériodicité est satisfaite dès que $q(x,x) \neq 0$ pour au moins un $x \in \mathcal{A}$.

Remarque : Si on part de π, la loi de X_n est π pour tout n, donc calculs faciles.

Mais attention ! (X_n) n'est pas i.i.d. de loi π (voir ci-dessous).

Exemple : Jukes-Cantor

Loi après n pas ? Réponse : $(q_{\mathrm{JC}})^n$ est de type Jukes-Cantor, pour le paramètre

$$p_n = \tfrac{1}{4}\,(1 - (1-4p)^n). \qquad (*)$$

Comme $p_n \to \frac{1}{4}$, on sait que $(q_{\mathrm{JC}})^n \approx \begin{pmatrix} .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \end{pmatrix}$ et

$$\nu(q_{\mathrm{JC}})^n \approx \begin{pmatrix} .25 & .25 & .25 & .25 \end{pmatrix} = \pi.$$

Remarque : en fait, on n'avait pas besoin de faire le calcul de q^n, il suffisait de résoudre $\pi\, q = \pi$, ce qui, ici, est trivial.

En tous cas : si n est grand, $\mathbb{P}_\nu(X_n = x) \approx \pi(x) = .25$ pour tout $x \in \mathcal{A}$ et tout ν.

Caveat !

Sous le modèle M0, on aurait

$$\mathbb{P}(Y_n = x,\, Y_{n+1} = x') = \pi(x)\,\pi(x').$$

Ici, même pour n grand,

$$\mathbb{P}(X_n = x,\, X_{n+1} = x') \approx \pi(x)\, q(x,x').$$

Par exemple $\mathbb{P}(X_n = \mathtt{a},\, X_{n+1} = \mathtt{a}) \approx \frac{1}{4}\,(1-3p) \neq \frac{1}{4}\,\frac{1}{4}$.
Alors que $\mathbb{P}(X_n = \mathtt{a}) \approx \frac{1}{4}$ et $\mathbb{P}(X_{n+1} = \mathtt{a}) \approx \frac{1}{4}$.

Exemple : Kimura

Loi après n pas : encore de type Kimura, un peu compliqué.

Par contre, il est facile de savoir que

$$(q_{\mathrm{K}})^n \approx \begin{pmatrix} .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \end{pmatrix} = \pi,$$

et $\nu(q_{\mathrm{K}})^n \approx \begin{pmatrix} .25 & .25 & .25 & .25 \end{pmatrix}$.

Donc $(q_{\mathrm{JC}})^n \approx (q_{\mathrm{K}})^n$ mais (encore une fois), par exemple,

$$\mathbb{P}_{\mathrm{JC}}(X_n = \mathtt{a},\, X_{n+1} = \mathtt{a}) \approx \tfrac{1}{4}\,(1 - 3p)$$

et

$$\mathbb{P}_{\mathrm{K}}(X_n = \mathtt{a},\, X_{n+1} = \mathtt{a}) \approx \tfrac{1}{4}\,(1 - u - 2v)$$

donc les deux sont a priori différents.

Plus important : les deux modèles donnent des prévisions « macroscopiques » différentes.

Deuxième principe des processus M1
Convergence des fréquences empiriques.

Rappel : comptages $N_n(\mathtt{w}) = \sum_{i=1}^{n} \mathbf{1}(X_{i:i+\ell-1} = \mathtt{w})$ avec $\ell = |\mathtt{w}|$.

Et $R_n(\mathtt{w}) = N_n(\mathtt{w})/n$. Donc $R_n(\mathtt{w})$ est une variable aléatoire.

Le résultat (Loi des grands nombres)

• Hypothèses techniques : chaîne de Markov finie, irréductible, apériodique, π distribution stationnaire.
• Alors, il existe une fonction Π telle que, pour tout mot $\mathtt{w}$ et toute loi initiale ν,

$$R_n(\mathtt{w}) \to_{n\to\infty} \Pi(\mathtt{w}),$$

au moins aux sens où $\mathbb{E}_\nu(R_n(\mathtt{w})) \to \Pi(\mathtt{w})$ et, pour tout $\varepsilon > 0$,

$$\mathbb{P}_\nu(|R_n(\mathtt{w}) - \Pi(\mathtt{w})| \geqslant \varepsilon) \to_{n\to\infty} 0.$$

Ici,

$$\Pi(\mathtt{w}) = \mathbb{P}_\pi(X_{1:\ell} = \mathtt{w}) = \pi(w_1)\, q(w_1, w_2) \dots q(w_{\ell-1}, w_\ell).$$

Conséquence : si n est grand, $R_n(w_1) \approx \pi(w_1)$, mais aussi

$$R_n(w_1 w_2) \approx \pi(w_1)\, q(w_1, w_2),$$

et encore

$$R_n(w_1 w_2 w_3) \approx \pi(w_1)\, q(w_1, w_2)\, q(w_2, w_3) \dots$$

Estimation statistique dans le modèle M1

Comment estimer les paramètres $q(x,y)$ à partir d'une trajectoire $x_{1:n}$? Réponse : EMV !

$$v(q) = \nu(x_1) \prod_{i=2}^{n} q(x_{i-1}, x_i) = \nu(x_1) \prod_{x,y \in \mathcal{A}} q(x,y)^{N_{n-1}(x,y)}.$$

Donc

$$\log v(q) = \mathrm{C}^{\mathrm{te}} + \sum_{x,y \in \mathcal{A}} N_{n-1}(x,y) \, \log q(x,y).$$

Contraintes :

$$\sum_{y \in \mathcal{A}} q(x,y) = 1, \quad x \in \mathcal{A}.$$

Le résultat :

$$\widehat{q}(x,y) = N_{n-1}(x,y)/N_{n-1}(x).$$

On connaît une distribution $\widehat{\pi}$ stationnaire pour $\widehat{q}$, sans calcul :

$$\widehat{\pi}(x) = N_{n-1}(x)/(n-1).$$

En réalité, $\widehat{\pi}$ est seulement « presque » stationnaire car il y a un problème de comptage ± 1 au temps n.

Mais on fait comme si.

En pratique :

$$\widehat{q}(x,y) = N_n(x,y)/N_n(x), \quad \widehat{\pi}(x) = N_n(x)/n.$$

Avantage : estimation facile, on compte des mots de longueur 1 et 2.

Désavantage ou avantage ? Les fréquences des mots de longueur $\geqslant 3$ sont prédites par le modèle M1. Par exemple,

$$\begin{aligned} N_n(uvw) &= n \, R_n(uvw) \\ &\approx n \, \Pi(uvw) = n \, \pi(u) \, q(u,v) \, q(v,w) \\ &= n \, \Pi(uv) \, \Pi(vw)/\pi(v) \\ &\approx N_n(uv) \, N_n(vw)/N_n(v). \end{aligned}$$

Conséquence : si $N_n(uvw)$ est vraiment différent de

$$N_n(uv) \, N_n(vw)/N_n(v),$$

le modèle M1 est douteux.

Renvoi à la littérature : la procédure d'Arndt et al. pour résoudre le modèle d'évolution consiste à imposer l'égalité des fréquences

$$R_n(uvw) = R_n(uv) \, R_n(vw)/R_n(v).$$

Modèles Mm

Tout se passe comme pour M1. Pour tous $x \in \mathcal{A}$ et $x_{1:n} \in \mathcal{A}^n$, on demande

$$\mathbb{P}(X_n = x | X_{1:n-1} = x_{1:n-1}) = \mathbb{P}(X_n = x | X_{n-m:n-1} = x_{n-m:n-1}).$$

Paramètres :

- Loi initiale ν sur $\mathcal{A}^m$:

$$\nu(x_{1:m}) = \mathbb{P}(X_{1:m} = x_{1:m}).$$

- Transitions q de $\mathcal{A}^m$ vers $\mathcal{A}$:

$$q(x_{1:m}, x) = q(x | x_{1:m}) = \mathbb{P}(X_{n+m} = x | X_{n:n+m-1} = x_{1:m}).$$

Loi d'une séquence ($n \geqslant m$)

$$\mathbb{P}_\nu(X_{1:n} = x_{1:n}) = \nu(x_{1:m})\, q(x_{1:m}, x_{m+1}) \dots q(x_{n-m:n-1}, x_n).$$

Log-vraisemblance ($n \gg m$)

$$\log \mathbb{P}_\nu(X_{1:n} = x_{1:n}) \approx \sum_{x,\mathtt{w}} N_n(\mathtt{w}x) \log q(x|\mathtt{w}),$$

somme sur les lettres x et les mots $\mathtt{w}$ de longueur m, et $N_n(\mathtt{w}x)$ le nombre d'occurrences du mot $\mathtt{w}x$ dans la séquence $x_{1:n}$.

Remarque fondamentale $(X_n)_n$ suit un modèle Mm si et seulement si $(Y_n)_n$ suit un modèle M1, avec

$$Y_n = X_{n:n+m-1} = (X_n, X_{n+1}, \dots, X_{n-m+1}).$$

Conséquence : tous les résultats démontrés pour les processus M1 fonctionnent aussi pour $(X_n)_n$ mais il faut passer par $(Y_n)_n$.

- Convergence de $(Y_n)_n$ vers un équilibre stochastique π_m unique et indépendant de la distribution de départ. Et π_m est l'unique distribution sur $\mathcal{A}^m$ solution du système suivant : pour toute lettre x de $\mathcal{A}$ et tout mot $\mathtt{w}$ de longueur $m-1$,

$$\pi_m(\mathtt{w}x) = \sum_{y \in \mathcal{A}} \pi_m(y\mathtt{w})\, q(y\mathtt{w}, x).$$

- Conséquence pour $(X_n)_n$: pour toute lettre x,

$$\lim_{n\to\infty} \mathbb{P}_\nu(X_n = x) = \rho(x).$$

Pour toute lettre x dans $\mathcal{A}$ et tout $0 \leqslant i \leqslant m-1$,

$$\rho(x) = \sum_{\mathtt{w}} \pi_m(x\mathtt{w}) = \sum_{\mathtt{w}} \pi_m(\mathtt{w}x) = \sum_{\mathtt{w}',\mathtt{w}''} \pi_m(\mathtt{w}'x\mathtt{w}''),$$

où les deux premières sommes portent sur les mots $\mathtt{w}$ de longueur $m-1$ et la dernière somme porte sur les mots $\mathtt{w}'$ et $\mathtt{w}''$ de longueurs respectives i et $m-1-i$.

- Convergence des fréquences empiriques de $(X_n)_n$ (qui sont aussi des fréquences empiriques de $(Y_n)_n$) : pour tout mot $\mathtt{w}$,

$$\lim_{n\to\infty} R_n(\mathtt{w}) = \Pi(\mathtt{w}).$$

Si $|\mathtt{w}| = \ell < m$,

$$\Pi(\mathtt{w}) = \pi_m(\mathtt{w} \times \mathcal{A}^{m-\ell}) = \sum_y \pi_m(\mathtt{w}y),$$

où la somme porte sur tous les mots y de longueur $m-\ell$.

Si $|\mathtt{w}| = \ell \geqslant m$,

$$\Pi(\mathtt{w}) = \pi_m(w_{1:m})\, q(w_{1:m}, w_{m+1}) \dots q(w_{\ell-m:\ell-1}, w_\ell).$$

Estimation statistique du modèle Mm

Pour toute lettre x et tout mot $\mathtt{w}$ de longueur m,

$$\widehat{q}(\mathtt{w}, x) = \frac{N_n(\mathtt{w}x)}{N_n(\mathtt{w})},$$

et

$$\widehat{\pi}_m(\mathtt{w}) = \frac{N_n(\mathtt{w})}{n}, \qquad \widehat{\rho}(x) = \frac{N_n(x)}{n}.$$

• **Remarque** Les fréquences des mots de longueur $\geqslant m+2$ sont toutes prédites par le modèle : par exemple, pour toutes lettres x et y et tout mot $\mathtt{w}$ de longueur m, $x\mathtt{w}y$ est un mot de longueur $m+2$ et il faut avoir

$$N_n(x\mathtt{w}y) \approx \frac{N_n(x\mathtt{w})\, N_n(\mathtt{w}y)}{N_n(\mathtt{w})}.$$

• **Remarque** Équilibre à trouver entre ordre du modèle et longueur de la séquence observée. Le modèle Mm comporte $|\mathcal{A}|^{m+1}$ paramètres (la matrice q) avec $|\mathcal{A}|^m$ contraintes puisque la somme de chaque ligne vaut 1, donc

$$|\mathcal{A}|^m\,(|\mathcal{A}|-1) \quad \text{paramètres.}$$

• **Un additif : l'état « Fin »**

Pas fait : si on veut modéliser des séquences de longueurs finies, on ajoute un état « Fin » tel que $q(\text{Fin}|\text{Fin}) = 1$ (un cimetière, pour les mathématiciens).

En résumé

• **Apprentissage dans un modèle Mm**

Comptage des mots jusqu'à la longueur $m+1$ incluse.

• **Vraisemblance dans un modèle Mm**

Les comptages des mots de longueur $m+1$ (et la loi initiale) suffisent à calculer $\mathbb{P}(\mathbf{x})$.

• **Discrimination entre modèles Mm**

On utilise la vraisemblance pour déterminer si une nouvelle séquence $\mathbf{x}$ est plutôt décrite par un modèle $+$ ou $-$, donc on calcule

$$\ell(\mathbf{x}) = \log\left(\frac{\mathbb{P}_+(\mathbf{x})}{\mathbb{P}_-(\mathbf{x})}\right) = \sum_{x,\mathtt{w}} N(\mathtt{w}x)\log\left(\frac{\mathbb{P}_+(x|\mathtt{w})}{\mathbb{P}_-(x|\mathtt{w})}\right).$$

Première partie des données : estimation de q_+ et q_-. Deuxième partie des données : loi empirique de $\ell(\mathbf{x})$ quand $\mathbf{x}$ suit le modèle $+$ puis quand $\mathbf{x}$ suit le modèle $-$.

Si les deux lois empiriques diffèrent nettement, on peut tester de nouvelles séquences, sinon, c'est raté.

Problème et suite du cours

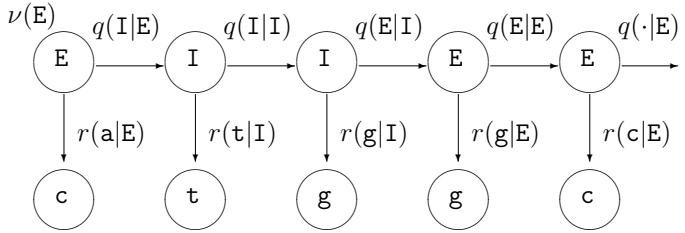
Dans tous les modèles Mm, la séquence est statistiquement homogène.

Pour l'ADN : gènes/régions intergéniques, introns/exons, etc.

Idée : décrire chaque type de région par un modèle Mm spécifique, puis recoller ces différents modèles.

Les chaînes de Markov cachées (HMM)

Exemple : recherche de structures dans un gène.
E = exon, I = intron.



Deux composantes :

• États $(S_n)_{n\geqslant 1}$: chaîne de Markov avec $S_n \in \mathcal{S}$, $\mathcal{S}$ fini. La loi initiale est ν avec $\nu(s) = \mathbb{P}(S_1 = s)$ et les transitions sont

$$\mathbb{P}(S_{n+1} = s' | S_n = s) = q(s, s') = q(s'|s).$$

• Observations $(X_n)_{n\geqslant 1}$: $X_n \in \mathcal{A}$, $\mathcal{A}$ fini et chaque état S_n émet l'observation X_n selon une loi qui dépend de S_n, donc

$$\mathbb{P}(X_n = x | S_n = s) = r(x|s) = r_s(x).$$

Dans l'exemple : $\mathcal{S} = \{\mathtt{E}, \mathtt{I}\}$ et $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

Dans les vrais modèles, $\mathcal{S}$ est beaucoup plus gros : phases des codons, exons initiaux, intermédiaires, finaux, etc.

On a décrit le modèle « M1M0 » (on utilise en fait des modèles MmMk, voir plus tard).

Historique Reconnaissance de la parole (89). Génomique dès 1989. Mais aussi : modélisation de la croissance des plantes, courbes de consommation électrique, fiabilité de logiciels, etc.

Modèle complet M

$$\begin{cases} \text{Distribution de l'état initial :} \\ \qquad \nu(s) \text{ pour } s \in \mathcal{S} \\ \text{Probabilités de transition entre états :} \\ \qquad q(s'|s) \text{ pour } (s, s') \text{ dans } \mathcal{S}^2 \\ \text{Probabilités d'émission des observations :} \\ \qquad r(x|s) \text{ pour } (x, s) \text{ dans } \mathcal{A} \times \mathcal{S} \end{cases}$$

Trois calculs

On fixe un modèle **M** (c'est-à-dire ν, q et r) et une longueur T de séquences.
On note $\mathbf{x} = x_{1:T}$ une séquence d'observations dans $\mathcal{A}^T$.
On note $\mathbf{s} = s_{1:T}$ une séquence d'états dans $\mathcal{S}^T$.

• Loi des états seuls : $\mathbb{P}(S_{1:T} = \mathbf{s}) = \nu(s_1)\, q(s_2|s_1) \dots q(s_T|s_{T-1})$.

• Loi globale : $\mathbb{P}(X_{1:T} = \mathbf{x},\, S_{1:T} = \mathbf{s})$ vaut

$$\nu(s_1)\, r(x_1|s_1)\, q(s_2|s_1)\, r(x_2|s_2) \dots q(s_T|s_{T-1})\, r(x_T|s_T).$$

• Loi des observations seules : $\mathbb{P}(X_{1:T} = \mathbf{x})$ vaut

$$\sum_{\mathbf{s}} \nu(s_1)\, r(x_1|s_1)\, q(s_2|s_1)\, r(x_2|s_2) \dots q(s_T|s_{T-1})\, r(x_T|s_T),$$

où la somme porte sur toutes les séquences $\mathbf{s}$ dans $\mathcal{S}^T$.

Trois objectifs

• Le modèle $\mathbf{M}$ et la séquence $\mathbf{x}$ sont donnés. Calculer la probabilité $\mathbb{P}_{\mathbf{M}}(\mathbf{x})$ d'observer $\mathbf{x}$ sous le modèle $\mathbf{M}$.

[Évaluation : algorithmes avant/arrière]

• Le modèle $\mathbf{M}$ et la séquence $\mathbf{x}$ sont donnés. Déterminer la séquence d'états $\mathbf{s}$ qui donne la plus grande chance $\mathbb{P}_{\mathbf{M}}(\mathbf{x}|\mathbf{s})$ d'émettre $\mathbf{x}$ sous le modèle $\mathbf{M}$.

[Estimation : algorithme de Viterbi]

• La séquence $\mathbf{x}$ est donnée. Déterminer le modèle $\mathbf{M}$ qui donne la plus grande chance $\mathbb{P}_{\mathbf{M}}(\mathbf{x})$ d'émettre $\mathbf{x}$.

[Identification/apprentissage : algorithme de Baum-Welch]

Premier objectif : émission

On fixe $\mathbf{M}$. On veut calculer $\mathbb{P}(\mathbf{x})$. Approche directe :

$$\mathbb{P}(\mathbf{x}) = \sum_{\mathbf{s}\in\mathcal{S}^T} \mathbb{P}(\mathbf{x}|\mathbf{s})\,\mathbb{P}(\mathbf{s}).$$

Pour chaque suite d'états $\mathbf{s}$,

$$\mathbb{P}(\mathbf{x}|\mathbf{s}) = r(x_1|s_1)\cdots r(x_T|s_T),$$
$$\mathbb{P}(\mathbf{s}) = \nu(s_1)\,q(s_2|s_1)\cdots q(s_T|s_{T-1}).$$

Pour chaque suite d'états $\mathbf{s}$, $2T$ multiplications, **mais $|\mathcal{S}|^T$ valeurs de s possibles**, donc $T\,|\mathcal{S}|^T$ opérations : boum !

La procédure « forward »

Pour chaque état s et chaque temps $1 \leqslant t \leqslant T$, on calcule la probabilité « forward »

$$\mathbf{f}_t(s) = \mathbb{P}(S_t = s,\, X_{1:t} = x_{1:t}).$$

• Début : au temps $t = 1$, pour chaque état s,

$$\mathbf{f}_1(s) = \nu(s)\,r(x_1|s).$$

• Récurrence $t \to t+1$ avec $1 \leqslant t \leqslant T-1$: pour chaque état s,

$$\mathbf{f}_{t+1}(s) = r(x_{t+1}|s)\sum_{s'\in\mathcal{S}} \mathbf{f}_t(s')\,q(s|s').$$

• Fin : $\mathbb{P}(\mathbf{x}) = \sum_{s\in\mathcal{S}} \mathbf{f}_T(s)$.

Nombre d'opérations : $T\,|\mathcal{S}|\,(|\mathcal{S}|+1)$. Taille mémoire : $T\,|\mathcal{S}|$ (car on veut souvent garder tous les $\mathbf{f}_t(s)$, voir plus bas).

Exemple : $|\mathcal{S}| = 5$, $T = 100$, on est passé de 10^{72} à 3000 opérations.

La procédure « backward »

Pour chaque état s et chaque temps $1 \leqslant t \leqslant T$, on calcule la probabilité « backward »

$$\mathbf{b}_t(s) = \mathbb{P}(X_{t:T} = x_{t:T} | S_t = s).$$

- Début : au temps $t = T$, $\mathbf{b}_T(s) = r(x_T|s)$.
- Récurrence $t \to t-1$ avec $2 \leqslant t \leqslant T$: pour chaque état s,

$$\mathbf{b}_{t-1}(s) = \sum_{s' \in \mathcal{S}} q(s'|s)\, r(x_{t-1}|s')\, \mathbf{b}_t(s').$$

- Fin : $\mathbb{P}(\mathbf{x}) = \sum_{s \in \mathcal{S}} \mathbf{b}_1(s)\, \nu(s)$.

Nombre d'opérations : $T\,|\mathcal{S}|\,(2|\mathcal{S}|)$. Taille mémoire : $|\mathcal{S}|$.

Conséquences des procédures avant/arrière

1) Loi a posteriori des états

La loi de l'état S_t sachant $\mathbf{x}$ se déduit de

$$\mathbb{P}(S_t = s,\, \mathbf{x}) = \mathbf{f}_t(s)\, \mathbf{b}_t(s).$$

Donc la loi conditionnelle de S_t vaut

$$\mathbb{P}(S_t = s |\, \mathbf{x}) = \frac{\mathbf{f}_t(s)\, \mathbf{b}_t(s)}{\mathbb{P}(\mathbf{x})}.$$

Pas besoin de calculer $\mathbb{P}(\mathbf{x})$ puisque, par exemple,

$$\mathbb{P}(\mathbf{x}) = \sum_{s' \in \mathcal{S}} \mathbf{b}_1(s')\, \nu(s') = \sum_{s' \in \mathcal{S}} \mathbf{f}_t(s')\, \mathbf{b}_t(s').$$

Rappel : les sommes sur $\mathcal{S}$ sont accessibles, pas celles sur $\mathcal{S}^T$.

2) Chemin uniformément optimal

Chemin $\mathbf{s}^{**} = s^{**}_{1:T}$ avec, pour chaque $1 \leqslant t \leqslant T$,

$$s_t^{**} = \operatorname{argmax}\, \mathbb{P}(S_t = s|\mathbf{x}),\ \ s \in \mathcal{S}.$$

Attention : le chemin $\mathbf{s}^{**}$ dans l'espace d'états peut être illégal.

3) Évaluation d'une fonctionnelle des chemins

Soit g une fonction sur l'espace d'états $\mathcal{S}$. On peut calculer

$$G(t|\mathbf{x}) = \mathbb{E}(g(S_t)|\mathbf{x}) = \sum_s g(s)\, \mathbb{P}(S_t = s|\mathbf{x}).$$

Deuxième objectif : décodage

Le modèle $\mathbf{M}$ est fixé. La séquence $\mathbf{x} = x_{1:T}$ est fixée.

Objectif : « décoder » la séquence d'observations $\mathbf{x}$, c'est-à-dire trouver le chemin $\mathbf{s} = s_{1:T}$ dans l'espace $\mathcal{S}$ des états qui a engendré ce chemin $\mathbf{x} = x_{1:T}$ dans l'espace $\mathcal{A}$ des observations.

Chemin le plus probable :

$$\mathbf{s}^* = \operatorname{argmax} \mathbb{P}_{\mathbf{M}}(\mathbf{s}|\mathbf{x}).$$

Exemple typique : détection de gènes eucaryotes. Version (très) simplifiée : trois états cachés introns/exons/intergénique.

Concrètement, étant donnée la séquence

`ccgtactagctgtagctgtgac`...`atcgggggctctggatctgcagactgg`

où sont les exons ?

Tester tous les chemins est impossible. Donc algorithme de programmation dynamique.

L'algorithme de Viterbi

Pour chaque état s et chaque temps $1 \leqslant t \leqslant T$, on calcule la vraisemblance « partielle »

$$\mathbf{v}_t(s) = \max_{\mathbf{u}} \mathbb{P}_{\mathbf{M}}(S_{1:t-1} = \mathbf{u},\, S_t = s, X_{1:t} = x_{1:t}).$$

• Début : si $t = 1$, pour chaque état s,

$$\mathbf{v}_1(s) = \nu(s)\, r(x_1|s).$$

• Récurrence $t \to t+1$ avec $1 \leqslant t \leqslant T-1$: pour chaque état s,

$$\mathbf{v}_{t+1}(s) \;=\; r(x_{t+1}|s) \max_{s' \in \mathcal{S}} \Big(\mathbf{v}_t(s')\, q(s|s') \Big).$$

On garde en mémoire les états

$$\mathbf{m}_t(s) \;=\; \operatorname{argmax}_{s' \in \mathcal{S}} \Big(\mathbf{v}_t(s')\, q(s|s') \Big).$$

• Fin et rétro-propagation : $s_T^* = \operatorname{argmax}_{s' \in \mathcal{S}} \mathbf{v}_T(s')$.

Pour chaque temps $1 \leqslant t \leqslant T-1$,

$$s_t^* = \mathbf{m}_t(s_{t+1}^*).$$

Nombre d'opérations : $T\,|\mathcal{S}|\,(|\mathcal{S}|+1)$. Taille mémoire : $T\,|\mathcal{S}|$.

Problèmes d'underflow

On multiplie des petites probabilités. Par exemple, pour des séquences génomiques de 100′000 bases, probabilités de l'ordre de $10^{-100'000}$.

Solution : le logarithme de $\prod_i$ vaut $\sum_i$ log donc on manipule

$$\mathbf{w}_t(s) = \log \mathbf{v}_t(s).$$

L'étape de récurrence $t \to t+1$ devient : pour chaque s dans $\mathcal{S}$,

$$\mathbf{w}_{t+1}(s) \;=\; \log r(x_{t+1}|s) + \max_{s' \in \mathcal{S}} \Big(\mathbf{w}_t(s') + \log q(s|s') \Big).$$

On a toujours

$$\mathbf{m}_t(s) \;=\; \mathrm{argmax}_{s' \in \mathcal{S}} \Big(\mathbf{w}_t(s') + \log q(s|s') \Big).$$

La rétropropagation est similaire :

$$s_T^* = \mathrm{argmax}_{s' \in \mathcal{S}} \mathbf{w}_T(s'), \quad s_t^* = \mathbf{m}_t(s_{t+1}^*).$$

Stabilité numérique de l'algorithme « avant »

Même problème mais on ne peut pas passer au logarithme.

Une solution : renormaliser par a_t au temps t et calculer

$$\widetilde{\mathbf{f}}_t(s) = \mathbf{f}_t(s) \prod_{i \leqslant t} a_i.$$

Nouvelle récurrence ? (Exercice.)

Un exemple « historique » : les îlots CpG

Attention : CpG désigne `c` puis `g` sur un même brin, et non pas une paire complémentaire `c-g` en un locus donné des deux brins.

Principe biologique : la cytosine `c` des CpG a tendance à être méthylée, souvent en thymine `t`. Donc les dinucléotides `cg` sont plus rares que le produit des fréquences de `c` et de `g`...

...Sauf autour des promoteurs de certains gènes, où la méthylation est réprimée !

Fait d'expérience : plus de `cg` et de `c` et de `g` autour des régions promotrices qu'ailleurs ; on parle d'îlots CpG.

Objectif : trouver les îlots CpG.

Remarque : problème de dinucléotides donc M1 naturel.

Référence : Durbin, Eddy, Krogh, Mitchison (1998).

Ensemble d'entraînement de 60 kb, 48 îlots CpG.

Deux modèles M1 par EMV (comptages), notés + pour les îlots CpG et − pour le reste.

$$q_+ = \begin{pmatrix} .180 & .274 & .426 & .120 \\ .171 & .368 & .274 & .188 \\ .161 & .339 & .375 & .125 \\ .079 & .355 & .384 & .182 \end{pmatrix}.$$

$$q_- = \begin{pmatrix} .300 & .205 & .285 & .210 \\ .322 & .298 & .078 & .302 \\ .248 & .246 & .298 & .208 \\ .177 & .239 & .292 & .293 \end{pmatrix}.$$

Premier problème Identifier une séquence $\mathbf{x}$ comme étant un îlot CpG ou non.

Calculs de vraisemblance : le (log)score de $\mathbf{x}$ est

$$\log\left(\frac{\mathbb{P}_+(\mathbf{x})}{\mathbb{P}_-(\mathbf{x})}\right) = \sum_{x,x'\in\mathcal{A}} N_{\mathbf{x}}(x,x') \log\left(\frac{q_+(x,x')}{q_-(x,x')}\right).$$

Deuxième problème Trouver la place des îlots CpG dans une séquence donnée.

Approche naïve : utiliser des fenêtres glissantes et calculer le (log)score de chaque fenêtre. Inconvénient : quelle(s) longueur(s) de fenêtre choisir ?

En fait : HMM.

Option de Durbin et al. un peu dégénérée : en passant de + à − ou vice versa, on saute vers une des 4 lettres choisies avec la même probabilité.

Chemin +/− le plus probable estimé par Viterbi.

Donc prédiction des îlots CpG d'une nouvelle séquence.

Exemple :

```
a c g a t c g c g c c a c g g t t t a t a t a a g c a a
−−−−−+++++++++−−−−−−−−−−−−−−−
```

La suite de + est une île prédite.

Troisième objectif : estimation

La séquence $\mathbf{x} = x_{1:T}$ est fixée.

Objectif : trouver le modèle $\mathbf{M}$ qui rende le mieux compte de la séquence d'observations $\mathbf{x}$. Modèle le plus probable :

$$\mathbf{M}^* = \operatorname{argmax} \mathbb{P}_{\mathbf{M}}(\mathbf{x}).$$

On utilise un cas particulier de l'algorithme EM (pour expectation/maximisation) : ré-estimation itérative et convergence vers un optimum local.

L'algorithme de Baum-Welch

0) Principe

On part d'un modèle $\mathbf{M}$; on ré-estime les valeurs des paramètres du modèle, ce qui donne $\widehat{\mathbf{M}}$ avec

$$\mathbb{P}(\mathbf{x}|\mathbf{M}) \leqslant \mathbb{P}(\mathbf{x}|\widehat{\mathbf{M}}).$$

Puis on recommence avec $\widehat{\mathbf{M}}$ en lieu et place de $\mathbf{M}$.

1) Notations

On fixe un modèle $\mathbf{M}$. Pour des états s et s',

$$\mathbf{c}_t(s,s') = \mathbb{P}_{\mathbf{M}}(S_t = s,\ S_{t+1} = s'|\mathbf{x}),$$

et

$$\mathbf{c}_t(s) = \mathbb{P}_{\mathbf{M}}(S_t = s|\mathbf{x}) = \sum_{s'\in\mathcal{S}} \mathbf{c}_t(s,s').$$

Si on somme $\mathbf{c}_t(s,s')$ et $\mathbf{c}_t(s)$ le long de la séquence, on obtient les quantités

$$\begin{aligned}\mathbf{C}(s,s') &= \sum_{t=1}^{T} \mathbf{c}_t(s,s') = \mathbb{E}_{\mathbf{M}}(N_T(s,s')|\mathbf{x}),\\ \mathbf{C}(s) &= \sum_{t=1}^{T} \mathbf{c}_t(s) = \mathbb{E}_{\mathbf{M}}(N_T(s)|\mathbf{x}).\end{aligned}$$

Enfin, on peut sommer $\mathbf{c}_t(s)$ le long de la séquence en ne gardant que les sites t où l'observation x_t vaut x, soit

$$\mathbf{C}_x(s) = \sum_{t=1}^{T} \mathbf{c}_t(s)\,\mathbf{1}(x_t = x).$$

2) Rappel sur avant/arrière

Pour un modèle $\mathbf{M}$ donné,

$$\mathbf{c}_t(s,s') = \frac{\mathbb{P}_{\mathbf{M}}(S_t = s,\, S_{t+1} = s',\, \mathbf{x})}{\mathbb{P}_{\mathbf{M}}(\mathbf{x})},$$

soit

$$\mathbf{c}_t(s,s') = \frac{\mathbf{f}_t(s)\, q(s'|s)\, r(x_{t+1}|s')\, \mathbf{b}_{t+1}(s')}{\mathbb{P}_{\mathbf{M}}(\mathbf{x})}.$$

Donc on peut calculer $\mathbf{c}_t(s)$, $\mathbf{C}(s,s')$, $\mathbf{C}(s)$ et $\mathbf{C}_x(s)$ comme des sommes (au moins à un facteur près).

3) L'étape $\mathbf{M} \to \widehat{\mathbf{M}}$

On est prêt à estimer les transitions des états par une nouvelle matrice $\widehat{q}$ et les émissions par une nouvelle matrice $\widehat{r}$. On utilise les estimateurs du maximum de vraisemblance dans le modèle $\mathbf{M}$, donc

$$\widehat{q}(s'|s) = \frac{\mathbf{C}(s,s')}{\mathbf{C}(s)}, \quad \widehat{r}(x|s) = \frac{\mathbf{C}_x(s)}{\mathbf{C}(s)}.$$

Le nouveau modèle $\widehat{\mathbf{M}}$ utilise les paramètres $\widehat{q}$ et $\widehat{r}$.

4) L'algorithme de Baum-Welch

- Initialiser la valeur de $\mathbf{M}$.
- Appliquer l'étape $\mathbf{M} \to \widehat{\mathbf{M}}$.
- Comparer les vraisemblances $\mathbb{P}(\mathbf{x}|\mathbf{M})$ et $\mathbb{P}(\mathbf{x}|\widehat{\mathbf{M}})$.
- Retourner à l'étape $\mathbf{M} \to \widehat{\mathbf{M}}$ jusqu'à stabilisation de la vraisemblance.

Remarque À chaque étape, $\mathbb{P}(\mathbf{x}|\mathbf{M}) \leqslant \mathbb{P}(\mathbf{x}|\widehat{\mathbf{M}})$.

Remarque Nécessité de définir un « critère d'arrêt » : différence des vraisemblances inférieure à un seuil absolu ; gain proportionnel inférieur à un seuil absolu ; idem sur un nombre d'itérations fixé à l'avance ; etc.

5) Extension au cas de plusieurs séquences observées

On suppose (souvent abusivement) que les I séquences observées $\mathbf{x}^i$ sont indépendantes les unes des autres et issues d'un même modèle $\mathbf{M}$, donc leur vraisemblance jointe sous $\mathbf{M}$ vaut

$$\mathbb{P}(\mathbf{x}^1|\mathbf{M})\,\mathbb{P}(\mathbf{x}^2|\mathbf{M})\cdots\mathbb{P}(\mathbf{x}^I|\mathbf{M}).$$

Pour chaque séquence $\mathbf{x}^i$ observée, on peut calculer $\mathbf{C}(s,s'|\mathbf{x}^i)$, $\mathbf{C}(s|\mathbf{x}^i)$ et $\mathbf{C}_x(s|\mathbf{x}^i)$ comme expliqué ci-dessus dans le cas d'une séquence.

Ensuite, on utilise $\widehat{q}(s'|s)=\dfrac{\mathbf{C}(s,s')}{\mathbf{C}(s)}$ et $\widehat{r}(x|s)=\dfrac{\mathbf{C}_x(s)}{\mathbf{C}(s)}$, en ayant posé $\mathbf{C}(s,s')=\displaystyle\sum_{i=1}^{I}\mathbf{C}(s,s'|\mathbf{x}^i)$, $\mathbf{C}_x(s)=\displaystyle\sum_{i=1}^{I}\mathbf{C}_x(s|\mathbf{x}^i)$, etc.

Tout se passe comme si on avait concaténé toutes les séquences en une seule.

6) Stabilité

En itérant Baum-Welch, on augmente la vraisemblance de la collection de séquences observées. Donc la vraisemblance converge.

Mais il n'y a pas (forcément) convergence dans l'espace des modèles. En pratique, la suite $\mathbf{M}_n$ des modèles obtenus après n itérations peut osciller violemment, même si le score de $\mathbf{M}_n$ converge.

Par ailleurs : problème des maxima locaux. Solution : utiliser plusieurs valeurs initiales différentes, faire tourner l'algorithme pour chacune de ces valeurs initiales, et espérer.

Ou partir de valeurs de $\mathbf{M}$ significatives biologiquement.

En conclusion

Procédure : 1. Choisir un ensemble d'états : codant/non codant, introns/exons/intergénique, prendre en compte les phases, lessignaux peptidiques, etc. 2. Choisir les transitions licites.
En pratique, permettre toutes les transitions donne de mauvais modèles (problèmes d'estimation). Donc utiliser les connaissances biologiques.
Une fois que la classe du modèle est définie, utiliser un ensemble d'entraînement pour estimer q et r. Ensuite, on peut analyser de nouvelles séquences par Viterbi, avant/arrière, etc.

Critiques Attention aux nouvelles séquences trop éloignées (au sens biologique) des séquences d'entraînement.

De réels problèmes Comment choisir l'ordre du modèle : critères BIC, AIC, etc. L'ordre k du modèle markovien engendrant les observations sous l'état s peut dépendre de s : VLHMM (Variable Length HMM). Pour éviter de faire exploser le nombre total de paramètres, on se permet des longueurs plus grandes dans certains états seulement. Modèles voisins des HMM : semi-chaînes de Markov cachées, etc.

Quelques applications des HMM à la génomique Hétérogénéité des séquences sans renseignements a priori. Transferts horizontaux de gènes. Recherche de motifs. Prédiction et annotation de gènes. Alignements de séquences. Reconstruction d'arbres phylogénétiques. Prédiction de structures secondaires. Etc.

Moralité On cherche à détecter une structure composée de modules élémentaires, chacun des modules est puisé dans une collection finie, on veut connecter les modules entre eux mais la mosaïque est inconnue.

– Fin –