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 gc, $B = \{\mathsf{g}, \mathsf{c}\}$ et

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

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 $\mathcal{A} = \{\mathsf{a}, \mathsf{c}, \mathsf{g}, \mathsf{t}\}$, 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 $X_2 = \mathsf{a}$ si $X_2' = \mathsf{t}$, $X_2 = \mathsf{c}$ si $X_2' = \mathsf{g}$, 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 $|*| = 0$.

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.

Plus intéressant est le cas suivant : on part d'une séquence $x_{1:n}$ et on choisit le modèle M0 du maximum de vraisemblance. Alors, le nombre d'occurrences d'un mot $\mathtt{w} = w_1 w_2$ doit vérifier

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

donc

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

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

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

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