

Modèles de Markov parcimonieux : sélection de modèle et estimation

P.-Y. Bourguignon & D. Robelin
{bourguignon,robelin}@genopole.cnrs.fr
Laboratoire Statistique et Génome
523, place des terrasses de l'Agora – 91000 Evry – France

15 juin 2004

Résumé

La modélisation de séquences biologiques par les chaînes de Markov présente un intérêt certain. Cependant, elle se heurte à divers écueils, dont le principal est la croissance exponentielle de la dimension du paramètre avec l'ordre du modèle. Depuis une vingtaine d'années, des modèles de Markov contraints ont été introduits, en particulier sous la forme des chaînes de Markov à longueur variable ou des *sparse Markov transducers*. Nous proposons une généralisation de ces modèles contraints, ainsi qu'un algorithme permettant le calcul direct du modèle présentant la probabilité a posteriori la plus grande parmi l'ensemble des modèles proposés. Deux exemples illustrant l'intérêt de cette modélisation et de l'algorithme de sélection de modèles sont proposés.

1 Introduction

1.1 Chaînes de Markov pour la modélisation de séquences biologiques

Les chaînes de Markov, comme modèle à mémoire courte universel, sont largement utilisées pour la modélisation de séquences biologiques [Wat95]. Bien que les séquences de nucléotides (pour ce qui est de l'ADN) ou d'acides aminés (pour ce qui est des protéines) ne soient pas des réalisations de chaînes de Markov à proprement parler, leur modélisation en tant que telles permet d'en extraire une information pertinente pour le biologiste.

Cependant, la déviation des séquences biologiques par rapport au modèle de Markov incite à les modéliser avec des modèles dont l'ordre est le plus grand possible, afin de capturer la plus grande quantité d'information possible. Mais la croissance exponentielle du nombre de paramètres avec l'ordre du modèle pose des limitations pratiques sur ce dernier, sans quoi les estimateurs des paramètres présentent une variance importante.

Un compromis entre ordre du modèle et dimension du paramètre peut être trouvé en recourant à une étape de sélection d'un modèle inclus dans le modèle de Markov d'ordre k avant l'estimation. Dans l'approche développée par RISSANEN [Ris83], la sélection de modèle est conduite dans la classe des *chaînes de Markov à longueur variable* par l'algorithme *context*. Les chaînes de Markov à longueur variable, dont nous rappelons la définition ci-dessous, permettent en effet d'imposer l'égalité entre certaines lignes de la matrice de transition, et par conséquent d'économiser les dimensions du paramètre là où la séquence le permet. Ce faisant, on introduit une modélisation de la structure de dépendance de la séquence vis-à-vis de son passé, qui dans certains peut refléter une réalité biologique. C'est ainsi que ESKIN, GRUNDY et SINGER [EGS00] introduisent une manière plus générale d'imposer les égalités entre lignes de la matrice de transition, les *Sparse Markov Transducers*, qui permet de construire des modèles spécifiques d'une classe de protéine, et par conséquent de réaliser des classifications de ces dernières.

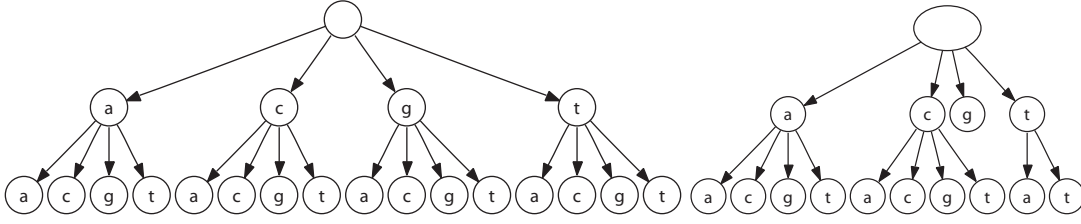


FIG. 1 – **A gauche** : exemple de représentation arborescente d'un modèle de Markov d'ordre 2.
A droite : exemple de représentation arborescente d'un modèle VLMC

Suivant leur mouvement généralisateur, nous introduisons une nouvelle classe de modèles inclus dans les modèles markoviens : les modèles de markov parcimonieux. Ceux-ci permettent d'imposer l'égalité entre lignes de la matrice de transition de manière plus générale que dans les chaînes de Markov à longueur variable et dans les *sparse Markov transducers*.

1.2 Les chaînes de Markov à longueur variable

On s'intéresse à un processus à temps discret $(X_t)_{t \in \mathbb{Z}} \in \mathcal{X}^{\mathbb{Z}}$ sur $(\Omega, \mathcal{A}, \mathcal{P})$, avec $|\mathcal{X}| < \infty$. Une réalisation de ce processus est notée $(x_t)_{t \in \mathbb{Z}}$. Que l'on s'intéresse au processus ou à l'une de ses réalisations, on note x_i^j la séquence extraite $x_i \dots x_j$, et on note $|x_i^j| = j - i + 1$ la longueur de cette séquence. On définit une chaîne de Markov à longueur variable comme suit :

Définition 1.1 *Le processus stationnaire $(X_t)_{t \in \mathbb{Z}}$ est une chaîne de Markov de longueur variable s'il existe une fonction c telle que :*

$$c : x_{-\infty}^0 \rightarrow x_{-l(x_{-\infty}^0)+1}^0, \text{ où } l \text{ est définie par}$$

$$l(x_{-\infty}^0) = \min \{k | \forall x_1 \in \mathcal{X}, \mathbb{P}(X_1 = x_1 | X_{-\infty}^0 = x_{-\infty}^0) = \mathbb{P}(X_1 = x_1 | X_{-k+1}^0 = x_{-k+1}^0)\} \quad (1)$$

Les séquences appartenant à l'image de la fonction c sont appelés contexte, et la fonction c est par conséquent appelée fonction contexte.

Remarque 1.1 $l(X_{-\infty}^0)$ est la longueur du contexte de X_1 , et on a la relation : $l(X_{-\infty}^0) = |c(X_{-\infty}^0)|$.

Notons \mathcal{C} l'ensemble $\{w \in \mathcal{A}^{|w|}, \exists x_{-\infty}^0, w = l(x_{-\infty}^0)\}$. Il s'agit de l'ensemble des contextes pris en compte par la chaîne de Markov à longueur variable. Ceux-ci peuvent être représentés comme les feuilles d'un arbre appelé *arbre de contexte* (cf. figure 1). Cet arbre est formé de nœuds labélisés par des lettres de \mathcal{X} . On le construit à partir d'un nœud racine, qui représente la lettre à prédire, et on l'étend vers le bas de telle manière que chaque nœud interne à l'arbre a au plus $|\mathcal{X}|$ fils. L'extension s'arrête lorsque la séquence formée depuis la racine jusqu'à la feuille est un contexte. Le modèle VLMC est entièrement spécifié dès lors que l'on associe à chaque feuille son vecteur de probabilités de transition. Un tel arbre est classiquement dénommée arbre à suffixes probabiliste [RST96].

2 Modèles de Markov parcimonieux

On peut voir une VLMC comme une chaîne de Markov classique, dont la matrice de transitions est soumise à des contraintes d'égalités entre certaines lignes. Cependant, les ensembles de lignes dont on peut imposer l'égalité est limité par la structure d'arbre. Cette limitation présente un coût en termes de dimension du paramètre, que l'on peut encore abaisser en permettant que deux sous-arbres issus de nœuds partageant le même père soient identiques. Par identiques, on entend identité des sous-arbres ainsi que des probabilités associées à leur feuille. Dans un tel modèle, on introduit une modélisation de la structure de dépendance de la séquence vis-à-vis de son passé plus fine que dans le cadre des VLMC.

2.1 Notations

Dans un but de généralisation, nous allons nous intéresser à l'ensemble des manières de partitionner l'alphabet de la chaîne de Markov. Pour cela, nous devons introduire quelques notations.

Définition 2.1 On appelle alphabet généralisé l'ensemble des parties de \mathcal{X} , $\bar{\mathcal{X}} = \mathcal{P}(\mathcal{X}) \setminus \{\emptyset\}$. On a $|\bar{\mathcal{X}}| = 2^{|\mathcal{X}|} - 1$. Un élément de cet alphabet est appelé un symbole.

Exemple 2.1 Pour l'étude de séquences d'ADN, on utilise un alphabet de nucléotides $\mathcal{X} = \{a, c, g, t\}$. Sur cet alphabet, l'alphabet généralisé est

$$\bar{\mathcal{X}} = \{a, c, g, t, [ac], [ag], [at], [cg], [ct], [gt], [acg], [act], [agt], [cgt], [acgt]\}$$

de cardinal 15. La notation $[ag]$ désigne "a ou g". Le codon start est alors décrit par le motif $[ag]tg$.

Définition 2.2 On appelle motif \bar{w} un mot formé sur l'alphabet généralisé $\bar{\mathcal{X}}$: $\bar{w} \in \bar{\mathcal{X}}^{|\bar{w}|}$. Lorsque la longueur d'un motif n'est pas fixée par le contexte, on appellera h -motif un motif de longueur h .

Dans toute la suite, on distinguera le terme *contexte* qui désigne un élément de \mathcal{X}^h , du terme *motif* désignant un élément de $\bar{\mathcal{X}}^h$.

Remarque 2.1 Avec la définition précédente, un motif peut également être vu comme un ensemble de contextes, tous de longueurs identiques. Par exemple, le motif $[ac]gt$ est décrit sans ambiguïté par l'ensemble de mots $\{agt, cgt\}$. Cependant, tout ensemble de mots ne définit pas nécessairement un motif; par exemple l'ensemble des codons "stop" $\{taa, tag, tga\}$ ne peut être résumé par un motif. Dans la suite, nous noterons $w \in \bar{w}$ lorsque w est un mot décrit par le motif \bar{w} .

Définition 2.3 Dans une séquence de longueur n , pour un motif $\bar{w} \in \bar{\mathcal{X}}^h$ avec $h < n$, on note $N(\bar{w})$ le comptage du motif \bar{w} dans la séquence, défini par la relation : $N(\bar{w}) = \sum_{w \in \bar{w}} N(w)$.

Un motif permet donc de décrire un ensemble de mots qui ne diffèrent que par quelques lettres comme un seul élément. En substituant des motifs aux contextes dans les chaînes de Markov à longueur variable, on permet donc à plusieurs contextes de partager la même loi de probabilité sur la lettre à suivre. Cependant, cette approche demande des précautions. Il faut en effet s'assurer qu'un contexte donné correspond à un et un seul motif, sans quoi le modèle comportera une indétermination sur la loi de la lettre à suivre. La définition suivante précise les ensembles de motifs qui n'induisent pas de telle indétermination.

Définition 2.4 On appelle ensemble de symboles admissible tout ensemble $\mathcal{S} \subset \bar{\mathcal{X}}$ qui forme une partition de \mathcal{X} . La classe des ensembles de symboles admissibles est notée \mathcal{P} .

Il y a donc autant d'ensembles de symboles admissibles que de partitions de \mathcal{X} , soit $e^{-1} \sum_{k=1}^{\infty} k^{|\mathcal{X}|} / k!$. Dans le cas particulier d'un alphabet à quatre lettres, on a 15 manières de construire un ensemble de symboles admissibles. A titre d'exemple, l'ensemble $\{a, [cg], t\}$ constitue un ensemble admissible de symboles. En revanche, $\{a, [ag], t\}$ ne l'est pas car a apparaît deux fois, et c n'apparaît pas.

2.2 Définition des modèles de Markov parcimonieux

Nous introduisons à présent la définition des modèles de Markov parcimonieux, par extension de la définition des chaînes de Markov à longueur variable. On envisage maintenant un arbre τ de profondeur h , dans lequel chaque nœud a pour ensemble de fils un ensemble de symboles admissibles. Dans cet arbre, un nœud terminal est naturellement associé à un chemin de la racine jusqu'au nœud, et l'ensemble des nœuds terminaux de l'arbre est donc indexé par un ensemble de h -motifs. Lus en sens inverse (du nœud vers la racine), ces motifs sont les *contextes* du modèle de Markov parcimonieux.

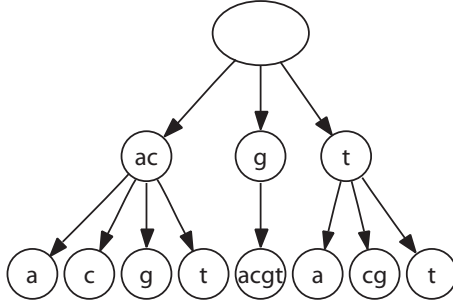


FIG. 2 – Exemple d’arbre décrivant un modèle de Markov parcimonieux

Définition 2.5 On note \mathcal{T}_h l’ensemble des arbres dont toutes les branches sont de longueur h et tel que l’ensemble des fils de chaque nœud interne est indexé par un ensemble de symboles admissible (voir figure 2).

Définition 2.6 Pour un arbre $\tau \in \mathcal{T}_h$, on note $M(\tau)$ l’ensemble des h -motifs qui, lus de droite à gauche, définissent le chemin de la racine au nœud terminal d’une branche de τ . Un ensemble de motifs \mathcal{W} qui est l’image par M d’un arbre $\tau \in \mathcal{T}_h$ est appelé ensemble de motifs admissible.

Définition 2.7 On appelle processus de Markov parcimonieux d’ordre h un processus $(X_t)_{0 \leq t \leq n}$ à valeurs dans \mathcal{X} , tel qu’il existe un entier h et un ensemble de h -motifs admissibles \mathcal{W}_0 vérifiant :

$$\mathcal{W}_0 = \operatorname{argmin}\{|\mathcal{W}|, \mathcal{W} \text{ admissible}, \forall w \in \mathcal{W}, \forall x_{-\infty}^1 \text{ vérifiant } x_{-h+1}^0 \in w, \forall t \text{ vérifiant } h \leq t \leq n, \\ \mathbb{P}(X_t = x_t | X_0^{t-1} = x_0^{t-1}) = \mathbb{P}(X_t = x_t | X_{t-h}^{t-1} = w)\} \quad (2)$$

L’ensemble \mathcal{W}_0 est appelé ensemble de motifs minimaux du processus de Markov parcimonieux. Dans la suite, on notera $\theta_{w,x_1} = \mathbb{P}(X_1 = x_1 | X_{-h+1}^0 = w)$ et $(\mu(w))_{w \in \mathcal{X}^h}$ la loi des h premières lettres.

Cette définition formelle appelle une explicitation. Il est immédiat de remarquer qu’un processus de Markov parcimonieux (noté *PMM*) est une chaîne de Markov d’ordre h , dont il est classique de considérer que les lignes de la matrice de transition sont indexées par l’ensemble des mots de longueur h . Un PMM permet d’imposer l’égalité de lignes dans la matrice de transitions, dans la mesure où chaque ensemble d’indices de lignes égales entre elles définit bien un motif, et que l’ensemble de ces motifs est admissible.

Ainsi formalisé, un processus PMM se définit par la donnée de l’ensemble \mathcal{W} des h -motifs qui réalise le minimum dans (1), ainsi que du vecteur $\theta_{\mathcal{W}}$ des probabilités d’émission de chacune des lettres après chacun de ces motifs. La donnée du seul ensemble \mathcal{W} spécifie un modèle PMM.

Définition 2.8 Un modèle de Markov parcimonieux est l’ensemble des processus de Markov parcimonieux qui partagent le même ensemble de motifs minimaux. Un tel modèle est donc naturellement associé à un unique ensemble de motifs admissible.

Remarque 2.2 Par souci de simplicité, nous prendrons comme loi initiale μ la loi uniforme sur les mots de longueur h . Dans toute la suite, cette quantité sera omise du vecteur des paramètres du modèle.

Le paramètre d’un modèle de Markov parcimonieux \mathcal{W}_0 a pour ensemble de définition :

$$\Theta_{\mathcal{W}_0} = \left\{ \left((\mu(w))_{w \in \mathcal{X}^h}, (\theta_{w,u})_{\substack{w \in \mathcal{W}_0 \\ u \in \mathcal{X}}} \right), \sum_{w \in \mathcal{X}^h} \mu(w) = 1 \text{ et } \forall w \in \mathcal{W}_0, \sum_{u \in \mathcal{X}} \theta_{w,u} = 1 \right\}$$

3 Approche bayésienne et estimation

Les PMM sont en nombre très important (de l'ordre de 10^{82} pour un alphabet à 4 lettres et un ordre $h = 5$). Il est donc vain d'aborder le problème de la sélection de modèles dans cette classe de manière gloutonne, par exemple en évaluant un critère du type AIC ou BIC sur l'ensemble des modèles afin de pouvoir les comparer. Par ailleurs, les théories fréquentistes de la sélection de modèles ne permettent pas la comparaison de modèles qui ne sont pas emboîtés. Ces deux raisons incitent à sortir des approches classiques pour adopter un point de vue bayésien. Nous verrons dans la partie suivante que cette approche conduit à un algorithme exact de sélection du modèle possédant la probabilité a posteriori maximale.

3.1 Modèle bayésien

Nous adoptons la modélisation classique de la sélection de modèle bayésienne, dans laquelle on tire le modèle selon sa loi a priori, puis le paramètre selon sa loi a priori conditionnellement au modèle choisi, puis enfin la séquence selon la loi ainsi obtenue. Un modèle PMM est représenté par son arbre τ , et on note θ_τ le paramètre courant dans le modèle τ .

La loi sur les modèles, $\pi(\tau)$, est choisie dans un premier temps uniforme. Ce choix est volontaire. Il permet de ne pas introduire de pénalisation arbitraire, et donc d'évaluer la pénalisation propre à l'approche bayésienne de la sélection de modèle. Nous discuterons plus loin du choix de la loi a priori sur les modèles.

Conditionnellement à l'arbre τ tiré, on munit l'ensemble Θ_τ des paramètres d'une loi a priori $\pi(\theta|\tau)$. Le choix d'un produit tensoriel sur les motifs de τ de lois de Dirichlet présente l'avantage de permettre le calcul explicite des lois a posteriori, tout en restant peu informatif. On a donc :

$$\pi((\theta_{w,u})|\tau) = \prod_{w \in M(\tau)} \prod_{u \in \mathcal{X}} \theta_{w,u}^{\alpha_u} \mathbb{1}_{\sum_{u \in \mathcal{X}} \theta_{w,u} = 1}$$

Le paramètre θ étant tiré selon cette loi, une séquence $(X_t)_{0 \leq t \leq n}$ est générée selon la vraisemblance :

$$l(X, \theta, \tau) = \mu(X_0, \dots, X_{h-1}) \prod_{w \in M(\tau)} \prod_{u \in \mathcal{X}} \theta_{w,u}^{N(wu)}$$

où $N(wu)$ désigne le comptage du motif wu dans la séquence. Avec cette modélisation, la marginale de la séquence s'écrit :

$$\mathbb{P}(X) = \sum_{\tau \in \mathcal{T}_h} \pi(\tau) \int_{\theta_\tau} l(X, \theta_\tau, \tau) \pi(\theta_\tau|\tau) d\theta_\tau$$

3.2 Estimation dans un modèle PMM

Nous traitons à présent de l'estimation dans un modèle PMM τ donné. Nous proposons deux approches de l'estimation :

- **Maximum de vraisemblance** Bien que fréquentiste, cette méthode d'estimation classique est inévitable.
- **Maximum a posteriori** Compte-tenu du choix d'une loi de Dirichlet comme a priori sur le paramètre, le maximum a posteriori présente la même forme que l'estimateur du maximum de vraisemblance. Son calcul est donc immédiat et ne nécessite pas le recours à une méthode d'échantillonnage.

Chacun de ces estimateurs présente un intérêt spécifique. Le maximum de vraisemblance, en plus de la simplicité de son calcul, est proposé par l'ensemble des logiciels de modélisation, contrairement au maximum a posteriori. En revanche, le maximum a posteriori permet de travailler sur des séquences courtes en utilisant une information a priori, qui en l'occurrence se traduit par le recours aux pseudo-comptages $N(wu) + \alpha_{w,u}$.

3.3 Maximum de vraisemblance

Le calcul de l'estimateur du maximum de vraisemblance dans un modèle PMM est semblable au calcul du même estimateur dans le modèle de Markov. En effet, le problème de maximisation à résoudre se formule ainsi :

$$\max \prod_{w \in \mathcal{W}} \prod_{u \in \mathcal{X}} \theta_{w,u}^{N(wu)}, \text{ sous les contraintes } \begin{cases} \forall w \in \mathcal{W}, \sum_{u \in \mathcal{X}} \theta_{w,u} = 1 \\ \forall w \in \mathcal{W}, \forall u \in \mathcal{X}, \theta_{w,u} \geq 0 \end{cases}$$

Dans le produit, les termes associés à des w différents peuvent être optimisés individuellement car les variables sont séparées et les termes tous positifs. Pour un terme w donné, dont nous omettons l'indice pour plus de simplicité, le lagrangien du problème d'optimisation s'écrit :

$$\mathcal{L}((\theta_u)_{u \in \mathcal{X}}, \lambda, (\lambda_u)_{u \in \mathcal{X}}, (\alpha_u)_{u \in \mathcal{X}}) = \prod_{u \in \mathcal{X}} \theta_u^{N(wu)} + \lambda \left(\sum_{u \in \mathcal{X}} \theta_u - 1 \right) + \sum_{u \in \mathcal{X}} \lambda_u (\theta_u - \alpha_u^2)$$

Si l'un des $N(wu)$ est nul, la vraisemblance possède une dérivée nulle par rapport $\theta_{w,u}$. Réciproquement, si $N(wu)$ est non-nul (ce que nous supposons vérifié pour au moins l'un des u pour tout w), prendre $\theta_{w,u}$ donne la valeur 0 à la vraisemblance, valeur qui ne peut être l'optimum. On a donc équivalence entre la nullité de $N(wu)$ et celle de $\theta_{w,u}$ à l'optimum. Le système d'optimalité associé s'écrit donc pour tout $u \in \mathcal{X}$:

$$\frac{\partial \mathcal{L}}{\partial \theta_u} = \mathbb{1}_{N(wu) > 0} \frac{N(wu)}{\theta_u} \prod_{v \in \mathcal{X}} \theta_v^{N(wv)} + \lambda + \lambda_u \quad (1) \qquad \frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{u \in \mathcal{X}} \theta_u - 1 \quad (2)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_u} = \theta_u - \alpha_u^2 \quad (3) \qquad \frac{\partial \mathcal{L}}{\partial \alpha_u} = -2\lambda_u \alpha_u \quad (4)$$

chacun de ces termes étant nul à l'optimum. Pour une transition $\theta_{w,u}$ qui n'est jamais observée dans la séquence, (1) conduit à la relation $\lambda + \lambda_u = 0$. Au contraire, si $N(wu) > 0$, on a $\theta_u > 0$ d'après (1), et donc $\alpha_u > 0$ d'après (3), ce qui implique que λ_u est nul d'après (4). Si bien que la relation suivante est toujours vérifiée :

$$\hat{\theta}_u = - \frac{N(wu) \prod_{v \in \mathcal{X}} \theta_v^{N(wv)}}{\lambda}$$

Sommant ces termes sur l'alphabet, il vient :

$$\hat{\lambda} = - \prod_{v \in \mathcal{X}} \theta_v^{N(wv)} \times \sum_{v \in \mathcal{X}} N(wv)$$

soit finalement :

$$\forall w \in \mathcal{W}, \forall u \in \mathcal{X}, \quad \hat{\theta}_u = \frac{N(wu)}{\sum_{v \in \mathcal{X}} N(wv)}$$

3.4 Maximum a posteriori

Fort du calcul précédent, le calcul du maximum a posteriori est immédiat. Cela provient du choix d'une loi a priori de Dirichlet, qui présente la particularité d'être conjuguée avec la vraisemblance markovienne. En effet, si p désigne la densité a posteriori du paramètre, on a :

$$p(\theta | X, \tau) = \mu(X_0, \dots, X_{h-1}) \prod_{w \in M(\tau)} \prod_{u \in \mathcal{X}} \theta_{w,u}^{N(wu) + \alpha_{w,u}}$$

et le calcul précédent montre que, dans ce cas, le maximum a posteriori $\bar{\theta}_{w,u}$ s'écrit :

$$\forall w \in \mathcal{W}, \forall u \in \mathcal{X}, \quad \bar{\theta}_{w,u} = \frac{N(wu) + \alpha_{w,u}}{\sum_{u \in \mathcal{X}} N(wu) + \alpha_{w,u}}$$

4 Sélection de modèle

Nous nous intéressons à présent au problème de la sélection de modèle dans la classe des PMM. Cet aspect de la théorie des PMM est essentiel, car il ouvre vers quantité d'applications de cette classe de modèles, en particulier à la classification. Nous proposons une approche bayésienne car elle conduit à un algorithme d'une complexité raisonnable compte-tenu de la combinatoire du problème.

4.1 Calcul de la probabilité a posteriori d'un modèle

Compte-tenu du formalisme proposé précédemment, le calcul de la probabilité a posteriori d'un modèle τ ne pose pas de problème particulier. En effet, on a :

$$\mathbb{P}(\tau|X) = \frac{\pi(\tau)}{\mathbb{P}(X)} \int_{\theta_\tau} l(X, \theta_\tau, \tau) \pi(d\theta_\tau|\tau)$$

Compte-tenu que la loi de Dirichlet utilisée comme a priori est conjuguée de la vraisemblance du modèle de Markov, le calcul précédent est immédiat :

$$\mathbb{P}(\tau|X) = \frac{\pi(\tau)}{\mathbb{P}(X)} \prod_{w \in M(\tau)} \frac{\prod_{u \in \mathcal{X}} (N(wu) + \alpha_{w,u} - 1)!}{(\sum_{u \in \mathcal{X}} (N(wu) + \alpha_{w,u} - 1))!}$$

En passant au logarithme, et en rejetant la probabilité des données ainsi que l'a priori sur le modèle (puisqu'il est choisi uniforme) dans une constante C , il vient :

$$\mathbb{P}(\tau|X) = \sum_{w \in M(\tau)} \left(\sum_{u \in \mathcal{X}} \ln(\Gamma(N(wu) + \alpha_{w,u})) - \ln(\Gamma(\sum_{u \in \mathcal{X}} N(wu) + \alpha_{w,u})) \right) + C$$

L'ensemble des termes de cette somme est en bijection avec l'ensemble \mathcal{W} des motifs du modèle. En d'autres termes, la probabilité a posteriori d'un arbre est la somme sur ses feuilles de la quantité $\sum_{u \in \mathcal{X}} \ln(\Gamma(N(wu) + \alpha_{w,u})) - \ln(\Gamma(\sum_{u \in \mathcal{X}} N(wu) + \alpha_{w,u}))$

4.2 Sélection par maximum a posteriori

Les approches classiques de la sélection de modèle ne permettent que la comparaison de modèles deux-à-deux, dans la mesure où l'un est inclus dans l'autre. Dans le cas général de la sélection d'un modèle au sein d'une classe de modèles vaste, dans laquelle les relations d'inclusion sont complexes, il est rarement possible de construire un algorithme qui ne soit pas glouton pour exhiber le modèle préféré. Une manière de contourner ce travers des tests de rapport de vraisemblance est de recourir aux vraisemblances pénalisées. Dans cette approche, le modèle préféré est l'argument maximum d'un critère dont le calcul est relativement simple. Cependant, sauf à disposer de propriétés de monotonie du critère à minimiser le long de certains chemins dans la classe de modèles, cette approche ne permet pas toujours une réduction de la complexité des algorithmes de sélection de modèle.

Dans l'approche bayésienne, nous montrons que le critère à maximiser (la probabilité a posteriori du modèle) présente une forme qui permet le calcul direct du modèle possédant la plus grande probabilité a posteriori par programmation dynamique. Ainsi, le nombre de modèles à comparer est restreint, ce qui permet d'atteindre une complexité algorithmique raisonnable. Le critère que nous proposons consiste donc à préférer le modèle dont la probabilité a posteriori est la plus grande.

De plus, nous détaillons l'algorithme de programmation dynamique pour le calcul du modèle sélectionné. En particulier, nous montrons que l'implémentation du sélecteur de modèles pour l'ensemble des PMM d'ordre 5 sur un alphabet de taille 4 est possible avec une consommation mémoire et processeur accessible à tout ordinateur grand public.

5 Algorithme de calcul exact du maximum a posteriori par programmation dynamique

Nous présentons à présent un algorithme de programmation dynamique qui conduit à la sélection du modèle possédant la plus grande probabilité a posteriori. Il ne nécessite qu'un parcours d'un arbre augmenté, construit sur l'alphabet généralisé $\bar{\mathcal{X}}$, et sa complexité est donc directement calculable en fonction de la taille de l'alphabet \mathcal{X} et de la profondeur h de la classe de PMM envisagée.

5.1 Récursivité

Afin d'établir la compatibilité du problème d'optimisation avec la théorie de la programmation dynamique, nous établissons dans un premier temps un résultat de récursivité du critère, qui permet l'élimination d'un grand nombre de modèles dominés à chaque itération du calcul.

Nous introduisons auparavant quelques notations utiles à l'exposé. Un nœud de profondeur k dans un arbre τ est identifié par le k -motif $w = w_1 \dots w_k$, où w_1 est le label du nœud, w_2 celui de son père, et ainsi de suite jusqu'à la racine. Pour un nœud w donné, on note $T(w)$ l'ensemble des sous-arbres possibles sous le nœud w .

Etant donné un motif, on introduit la *fonction de score* $S : \bigcup_{k=0}^h \bar{\mathcal{X}}^k \rightarrow \mathbb{R}$, définie par la relation :

$$\forall w \in \bigcup_{k=0}^h \bar{\mathcal{X}}^k, \quad S(w) = \max_{\tau \in T(w)} \sum_{w' \in \mathcal{M}(\tau)} \left(\sum_{u \in \mathcal{X}} \ln(\Gamma(N(w'u) + \alpha_{w',u})) - \ln(\Gamma(\sum_{u \in \mathcal{X}} N(w'u) + a_{w',u})) \right)$$

La somme précédente peut être réécrite sous la forme :

$$\forall w \in \bigcup_{k=0}^h \bar{\mathcal{X}}^k, \quad S(w) = \max_{S \in \mathcal{P}} \sum_{s \in S} S(sw)$$

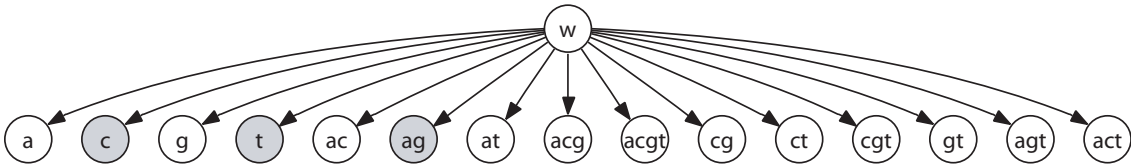
où sw désigne le motif obtenu en accolant le symbole s au début du motif w . Autrement dit, la valeur de S pour un nœud donné peut-être déduit des valeurs de S pour l'ensemble de ses $|\bar{\mathcal{X}}|$ fils possibles, en comparant les sommes des scores de chacun des fils pour chacun des ensembles de symboles admissibles. Dans le cas particulier de $w = \emptyset$, on a $S(w) = \max_{\tau \in \mathcal{T}_h} S(w) = \max_{\tau \in \mathcal{T}_h} \mathbb{P}(\tau|X)$.

Cette relation de récursivité permet donc d'appliquer la théorie de la programmation dynamique pour exhiber le modèle possédant la plus grande probabilité a posteriori. Cela conduit à l'algorithme décrit ci-après.

5.2 Algorithme

Nous décrivons à présent l'algorithme permettant la sélection de modèle. Il s'agit de déterminer l'arbre qui maximise la probabilité a posteriori, soit le critère S pris en sa racine.

Pour cela, on considère l'arbre de profondeur h dont l'ensemble des noeuds internes présentent $|\bar{\mathcal{X}}|$ fils, comme sur la figure suivante dans le cas de l'alphabet de nucléotides $\{a, c, g, t\}$ (où w désigne le noeud interne courant) :



La première étape consiste à calculer le score $S(w)$ pour chacune des feuilles de cet arbre. Puis on itère sur cet arbre le processus suivant, en commençant par les noeuds internes les plus profonds :

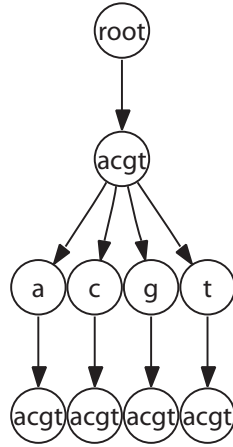


FIG. 3 – Exemple de PMM sélectionné sur une séquence simulée

1. pour chacun des ensembles de symboles admissible \mathcal{S} , calculer la somme des S pris en chacun des fils dans \mathcal{S} ,
2. retenir le sous-ensemble de symboles admissible \mathcal{S}_0 qui réalise le maximum de cette somme (l'ensemble des nœuds grisés ci-dessus, par exemple),
3. éliminer les fils associés à des symboles qui ne sont pas dans \mathcal{S}_0 .

Une fois que l'ensemble des nœuds d'un niveau donné ont été traités ainsi, on passe au niveau supérieur. Une fois arrivé à la racine, l'arbre obtenu décrit un PMM qui possède la propriété de maximiser la probabilité a posteriori parmi les arbres de même profondeur h .

6 Applications

Nous présentons à présent deux exemples illustrant l'intérêt des modèles de Markov parcimonieux pour l'économie de dimensions du paramètre. Le premier est fondé sur des simulations, et montre que l'algorithme est capable de retrouver la structure de dépendance de la séquence sur un échantillon relativement court. Le second traite de séquences biologiques réelles.

6.1 Séquence simulée

A titre d'exemple, nous avons simulé une séquence d'ordre 2 sur l'alphabet des nucléotides, dont la matrice de transition présente la structure suivante :

$$\Pi = \begin{pmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \end{pmatrix}$$

où les B_1, \dots, B_4 désignent des matrices de taille 4×4 dont toutes les lignes sont égales. Autrement dit, la distribution de probabilité sur X_t dépend de X_{t-2} , mais pas de X_{t-1} .

L'arbre sélectionné, sur une séquence simulée de longueur 2000, est représenté en figure 3.

On remarque que l'algorithme a sélectionné l'arbre sous-jacent à la structure de la matrice Π sans erreur.

6.2 Comparaison au modèle markovien classique

Afin de réaliser une comparaison tangible entre PMM et chaînes de Markov classiques, nous avons calculé la valeur du critère BIC atteinte par chacune des deux modélisations sur des séquences

Ordre	CDS			Génome complet		
	f_{PMM}	f_{10}	Δ	f_{PMM}	f_{10}	Δ
3	94,2%	66,5%	53,9	100%	44,3%	1996,3
4	77,2%	70,9%	682,7	100%	32%	2319,4
5	80,8%	60,3%	1332,37	100%	35%	1017,4

TAB. 1 – Comparaison du modèle PMM au modèle markovien classique

biologiques. Ces séquences représentent l'ensemble des génomes bactériens disponibles au NCBI, soit 224 à ce jour. Nous avons conduit une extraction des parties codantes (CDS) (ou du moins annotées comme telles dans les fichiers GenBank) sur chacune des séquences. L'ensemble des CDS extraits d'une même séquence sont conservés dans un unique fichier, ce qui permet la sélection de modèle ainsi que l'estimation sur l'ensemble des CDS d'un même organisme.

Pour toutes les séquences ainsi obtenues, nous avons estimé un modèle markovien d'ordre k et le modèle PMM d'ordre maximal identique, et avons calculé le BIC associé à chacun de ces modèles, selon la formule :

$$BIC = -2 \ln l(x, \theta) + k \ln n$$

où k désigne le nombre de paramètres du modèle, l désigne la vraisemblance, et n la longueur de la séquence. Un traitement particulier a été appliqué aux séquences codantes : les modèles sont estimés de manière phasée, i.e. un modèle différent est utilisé pour prédire les lettres selon qu'elles apparaissent en phase 1, 2 ou 3. Il est en effet courant de considérer que ces modèles s'ajustent beaucoup mieux sur des séquences codantes.

Nous avons conduit ce travail pour des ordres compris entre 3 et 5. Pour chacun des ordres envisagés, nous donnons les valeurs des statistiques suivantes :

- f_{PMM} , proportion des séquences favorables au modèle parcimonieux,
- f_{10} , proportion des séquences dont le BIC est au moins 10 fois plus grand dans le modèle classique que dans le modèle PMM,
- $\bar{\Delta}$, différence moyenne entre le BIC du modèle Markov et celui du modèle PMM, le BIC étant normalisé par la longueur de la séquence.

Il est donc manifeste que les modèles de markov parcimonieux permettent une économie substantielle de paramètres dont le coût en termes de qualité d'ajustement est acceptable. Ce phénomène est plus net lorsque l'on s'intéresse à des modèles non phasés, ce qui suggère une plus grande robustesse du modèle parcimonieux.

Enfin, nous proposons deux graphiques représentant les logarithmes en base 10 des valeurs des BIC obtenus pour des modèles d'ordre 5, dans le cas de séquences codantes avec modèle phasé, puis dans le cas de génomes complets.

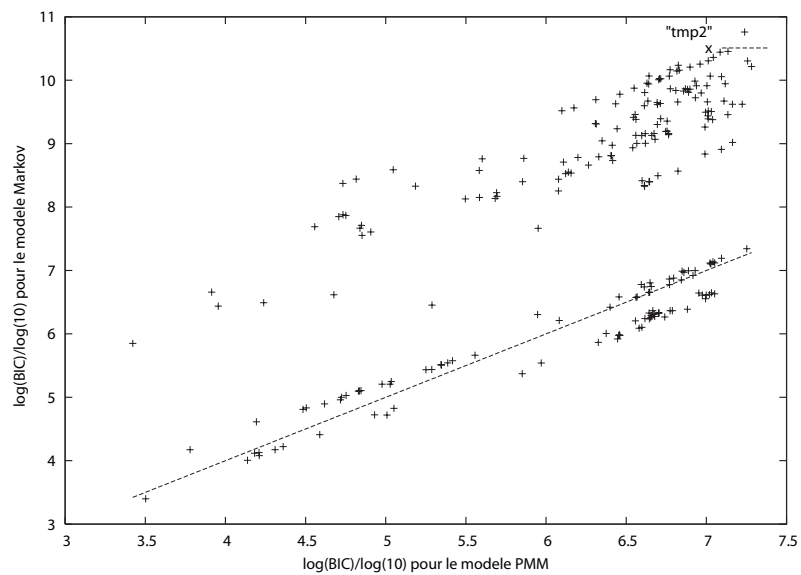


FIG. 4 – Logarithmes en base 10 des BIC obtenus par le modèle parcimonieux phasé (en abscisse) et par le modèle markovien phasé classique (en ordonnées), sur des séquences codantes.

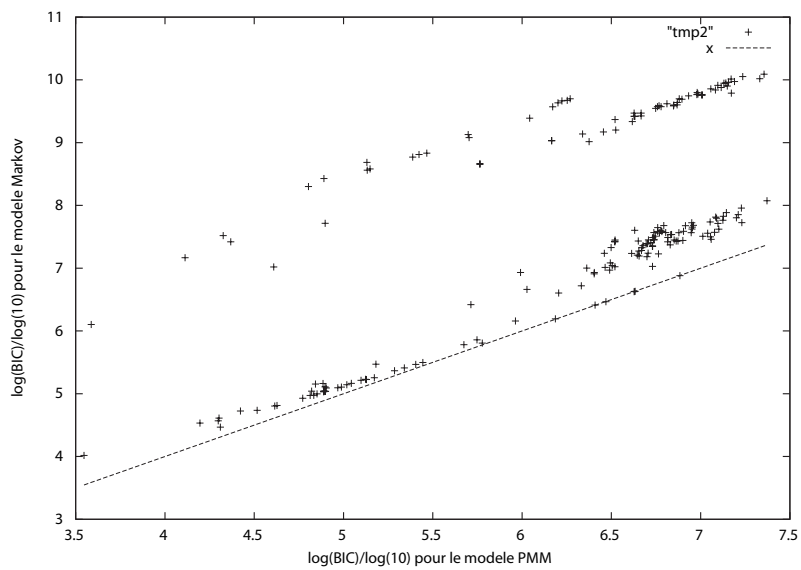


FIG. 5 – Logarithmes en base 10 des BIC obtenus par le modèle parcimonieux (en abscisse) et par le modèle markovien classique (en ordonnées), sur des génomes bactériens complets.

Références

- [EGS00] E. Eskin, W. N. Grundy, and Y. Singer. Protein family classification using sparse markov transducers. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 134–135, August 20-23 2000.
- [Ris83] J. Rissanen. A universal data compression system. *IEEE Trans. Inform. Theory*, 1983.
- [RST96] D. Ron, Y. Singer, and N. Tishby. The power of amnesia : Learning probabilistic automata with variable memory length. *Machine learning*, 1996.
- [Wat95] M. S. Waterman. *Introduction to computational biology*. Chapman & Hall, 1995.