Mélanges de Chaînes de Markov lissées pour la détection de domaines dans les protéines

Christopher KERMORVANT[†] Pierre DUPONT[‡]

[†] EURISE, Université Jean Monnet, Saint-Etienne, France
[‡] Dept. INGI, Université Catholique de Louvain, Louvain la Neuve, Belgique

Courriel : kermorva@univ-st-etienne.fr, pdupont@info.ucl.ac.be

Résumé

Nous proposons l'utilisation d'un modèle basé sur un mélange de chaînes de Markov lissées pour la détection de domaines dans les protéines. Ce modèle est une amélioration d'un modèle basé sur un mélange de chaînes de Markov élaguées récemment proposé. Ses performances pour la détection de domaines sont proches de celles des HMM alors que son coût calculatoire aussi bien lors de l'estimation que de la détection est plus faible. Sur la base de données SWISSPROT, nous montrons que l'utilisation du lissage entraîne une augmentation du taux de détection correcte avec un nombre de paramètres inférieur par rapport au modèle élagué.

Mots clés : *Protéines, détection de motifs, mélange de chaînes de Markov, lissage*

1 Introduction

De nombreuses bases de données rassemblent des informations concernant les protéines (séquences, fonctions, descriptions) et les organisent en familles. Elles servent entre autre à l'analyse des nouvelles protéines, par exemple, pour la recherche de sous-séquences (domaines, motifs) potentiellement associée à une structure ou une fonction. Différents types de modèles ont été proposés dans ce but, allant de modèles probabilistes complexes comme les modèles de Markov cachés [4] à la description de simples motifs caractéristiques [1]. Cependant, afin de tirer profit des mises à jour constantes de ces bases et de la quantité croissante de données disponibles, les outils d'analyse de protéines doivent être faciles à construire et de faible complexité calculatoire.

Les chaînes de Markov permettent de modéliser une distribution de probabilité sur un ensemble d'événements discrets. La probabilité d'un événement ne dépends alors que des K événements précédents, K étant appelé l'ordre de la chaîne. Une chaîne de Markov peut ainsi modéliser une famille de protéines par une distribution sur de courtes séquences, supposées caractéristiques de la famille considérée. Cependant ces modèles sont limités par le fait que leur nombre de paramètres est exponentiel en l'ordre de la chaîne. Un algorithme d'apprentissage de mélange de chaînes de Markov d'ordre variable a été proposé [11] permettant un compromis entre l'ordre maximal des chaînes et le nombre de paramètres du modèle. Récemment, un outil de détection de domaines basé sur ce modèle a été proposé [3]. Il atteint des performances de détection de domaines proches de celles des HMM pour un coût calculatoire de la détection inférieure. De plus, ce modèle ne nécessite pas pour son apprentissage de séquences alignées, réduisant ainsi le coût calculatoire de la phase d'apprentissage et le risque d'erreur.

Cependant, ce modèle néglige un aspect souvent démontré comme important pour les modèles basés sur des chaînes de Markov : le lissage des probabilités. En effet, même si les chaînes de Markov sont estimées sur de grandes bases de données, et même pour des ordres faibles, la taille de l'espace des événements est telle que de nombreux événements possibles ne sont pas observés durant la phase d'estimation et se voient alors attribuer à tord une probabilité nulle. Le lissage consiste alors à estimer la probabilité de ces événements non vus.

Dans cet article, nous étudions l'utilisation de mélanges de chaînes de Markov pour la détection de domaines dans les protéines. Nous comparons les performances de différents modèles sur la base SWISSPROT et montrons l'importance du lissage pour ces modèles.

2 Chaînes de Markov

2.1 Définition

Soit Σ un ensemble fini d'événements. Une séquence $X = (x_n)_{n\in\mathbb{N}}$ d'événements sur Σ définit une chaîne de Markov si pour tout $n \in \mathbb{N}$

$$P(x_n = s_n | x_0 = s_0, \cdots, x_{n-1} = s_{n-1}) \quad = \quad P(x_n = s_n | x_{n-1} = s_{n-1}) \tag{1}$$

Cette égalité définit l'ordre de la chaîne de Markov. Dans ce cas, la chaîne de Markov est d'ordre 1 car la probabilité de l'événement x_n ne dépend que de x_{n-1}. On peut noter qu'une chaîne de Markov d'ordre quelconque K peut

être interprétée comme une chaîne d'ordre 1 sur l'espace Σ^K. Étant donné une séquence $X = (x_n)_{n\in\mathbb{N}}$, on peut estimer les paramètres de la chaîne de Markov d'ordre 1 sous-jacente de la manière suivante :

$$\forall (s_{n-1}, s_n) \in \Sigma^2,\ P(x_n = s_n | x_{n-1} = s_{n-1}) = \frac{c(s_{n-1}, s_n)}{\sum_{s\in\Sigma} c(s_{n-1}, s)}$$

où $c(s_{n-1}, s_n)$ est le nombre de fois que la sous-séquence (s_{n-1}, s_n) est apparue dans la séquence $X = (x_n)_{n\in\mathbb{N}}$. Dans le cas d'une chaîne d'ordre K, les événements s sont des éléments de Σ^K de la forme $s = (s_1, \cdots, s_K)$. Les paramètres de la chaîne de Markov sont alors estimés par :

$$P(x_n = (s_{n-K+1}, \cdots, s_n) | x_{n-1} = (s_{n-K}, \cdots, s_{n-1})) = \frac{c(s_{n-K}, \cdots, s_n)}{\sum_{s\in\Sigma} c(s_{n-K}, \cdots, s_{n-1}, s)}$$

où $c(s_{n-K}, \cdots, s_n)$ est le nombre de fois que la sous-séquence $s_{n-K}, \cdots, s_n$ est apparue dans la séquence $X = (x_n)_{n\in\mathbb{N}}$. Une fois les paramètres de la chaîne de Markov $\mathcal{M}$ estimés, on peut calculer la vraisemblance d'une séquence $Y = (y_n)_{n=1..N}$ étant donné cette chaîne comme suit (dans le cas d'ordre 1) :

$$P(y_0 = s_0, \cdots, y_N = s_N | \mathcal{M}) = p(y_0 = s_0) \prod_{i=1}^{N} P(y_i = s_i | y_{i-1} = s_{i-1}) \tag{2}$$

2.2 Lissage

Dans les applications réelles, la taille de la séquence $X = (x_n)_{n=1..N}$ est finie. Il en résulte que plus l'ordre K de la chaîne de Markov estimée est grand, plus le nombre sous-séquences de longueur K observées par rapport au nombre de sous-séquences de longueur K possibles est petit. De nombreuses probabilités sont alors estimées avec un faible nombre d'observations et sont donc peu fiables. En particulier, de nombreux événements se voient attribuer à tord une probabilité nulle. De nombreuses méthodes de lissage ont été proposées pour améliorer l'estimation des événements de faible probabilité. On peut citer entre autres les lois de succession [9], le décomptage [10], l'interpolation linéaire [6], la formule de Turing-Good [7]. Une des meilleures méthodes de lissage a été proposée par Kneser et Ney [8]. Pour une chaîne d'ordre K, cette méthode consiste à décompter une certaine quantité d_C de la masse de probabilité des événements observés, et à la redistribuer à tous les événements possibles selon β une chaîne de Markov d'ordre $K-1$. Formellement, on a :

$$P(x_n = (s_{n-K+1}, \cdots, s_n) | x_{n-1} = (s_{n-K}, \cdots, s_{n-1})) = \begin{cases} \frac{c(s_{n-K}, \cdots, s_n)}{\sum_{s\in\Sigma} c(s_{n-K}, \cdots, s)} \gamma(s_{n-K}, \cdots, s_{n-1}) \beta(s_{n-K}, \cdots, s_{n-1}, s_n) & \text{si } c(c(s_{n-K}, \cdots, s_n) > 0 \\ \gamma(s_{n-K}, \cdots, s_{n-1}) \beta(s_{n-K}, \cdots, s_n) & \text{sinon} \end{cases} \tag{3}$$

où $\gamma(s_{n-K}, \cdots, s_{n-1})$ est un facteur de normalisation. La distribution β n'est pas estimée comme en 2.1, mais de la manière suivante :

$$\beta(s_{n-K}, \cdots, s_n) = \frac{c(\bullet, s_{n-K+1}, \cdots, s_n)}{\sum_{s\in\Sigma} c(\bullet, s_{n-K+1}, \cdots, s_{n-1}, \bullet))}$$

avec $c(\bullet, s_{n-K+1}, \cdots, s_n) = \left|\{s \mid c(s, s_{n-K+1}, \cdots, s_n) > 0\}\right|$. Cette expression est basée non pas sur une estimation de la fréquence d'occurrence de la séquence $s_{n-K}, \cdots, s_n$ mais sur le rapport du nombre de contextes différents dans lesquels s_n a été observé après la séquence $s_{n-K+1}, \cdots, s_{n-1}$ sur le nombre de contextes différents dans lesquels on a observé la séquence $s_{n-K+1}, \cdots, s_{n-1}$. Ce rapport pouvant lui-même être nul et donc lissé par une chaîne d'ordre inférieur, on voit alors que le modèle sous-jacent n'est plus une chaîne de Markov d'ordre K fixé, mais un mélange de chaînes de Markov d'ordres allant de 1 à K.

2.3 Élagage

Ron et *al* [11] ont proposé un algorithme qui permet d'augmenter l'ordre maximal d'un mélange de chaînes de Markov tout en gardant un nombre raisonnable de paramètres à estimer. Dans leur approche, étant donné un ordre maximal K, et pour toute chaînes de Markov d'ordre $i \in [1, \cdots, K]$, une probabilité n'est conservée dans le modèle que si elle est supérieure à une seuil et si elle est suffisamment différente de la probabilité correspondante dans le modèle d'ordre $i-1$. Formellement, ces conditions s'énoncent :

$$P(x_n = (s_{n-i+1}, \cdots, s_n) | x_{n-1} = (s_{n-i}, \cdots, s_{n-1})) > P_{min}$$

et

$$\frac{P(x_n = (s_{n-i+1}, \cdots, s_n) | x_{n-1} = (s_{n-i}, \cdots, s_{n-1}))}{P(x_n = (s_{n-i+2}, \cdots, s_n) | x_{n-1} = (s_{n-i+1}, \cdots, s_{n-1}))} \begin{cases} \geq r \\ \text{ou bien} \\ \leq \frac{1}{r} \end{cases}$$

où P_{min} et r sont des paramètres du modèle. Dans le cas contraire, cette probabilité est considérée comme nulle.

2.4 Application à la détection de domaines

Les mélanges de chaînes de Markov décrits dans la section précédente peuvent être utilisés pour la détection de domaines dans les protéines [3]. En effet, on peut associer à chaque domaine un mélange de chaînes de Markov et estimer les paramètres du mélange sur un ensemble de séquences de ce domaine. Étant donné la séquence d'une protéine inconnue, la vraisemblance de cette séquence sachant le modèle d'un domaine peut-être considérée comme un indice de la présence ou non de ce domaine dans la protéine. Une valeur de vraisemblance élevée indique alors une présence probable du domaine dans la protéine.

3 Expériences

3.1 Protocole expérimental

Nous avons utilisé deux bases de données pour tester les modèles : la base SWISSPROT [2] qui contient des séquences de protéines de plusieurs êtres vivants et la base PFAM, qui contient des domaines fonctionnels extraits des séquences de SWISSPROT par une procédure semi-automatique et classés par famille. Nous avons étiqueté les séquences de SWISSPROT par le nom des domaines qu'elles contiennent, comme défini dans les familles PFAM. Afin de nous comparer à des résultats similaires récemment publiés [3, 5], nous avons utilisé la version 1.0 de PFAM, qui contient 22307 domaines regroupés en 175 familles.

3.2 Entraînement des modèles

Pour chaque famille, nous avons entraîné les modèles sur 80% des domaines de PFAM. Les mélanges de chaînes de Markov avec élagage (notés PST pour *probabilistic suffix trees*) ont été entraînés avec le logiciel et les paramètres donnés par Bejerano [3]. Nous avons d'autre part entraîné des mélange de chaînes de Markov avec lissage de Kneser-Ney pour des valeur d'ordre maximum de 0 à 4, noté MCM0 à MCM4.

3.3 Test des modèles

Tous les modèles ont été testés pour la détection de domaine dans les séquences complètes des protéines de SWISSPROT correspondant à la totalité de la base PFAM. Afin de déterminer le taux de détection correcte, nous avons utilisé le critère d'iso-point [3, 5]. Pour chaque modèle de famille, un iso-point est calculé sur l'ensemble de test. L'iso-point est la valeur pour laquelle il y a autant de séquences ne contenant pas le domaine dont la vraisemblance est supérieure que de séquences contenant le domaine dont la vraisemblance est inférieure. Les séquences contenant le domaine et dont la vraisemblance est supérieure à l'iso-point sont considérées comme correctement détectées.

3.4 Résultats

La figure 1 présente les taux de détection correcte sur l'ensemble des séquences de SWISSPROT et le nombre de paramètres pour les PST et des mélanges de chaînes de Markov (MCM) d'ordre maximal allant de 0 à 4 (MCM0 à MCM4) lissées par la méthode de Kneser-Ney. Les performances des mélanges de chaînes de Markov lissées sont supérieures à celles des PST dès que l'ordre maximal du mélange est au moins égal à 3. De plus le nombre de paramètres des MCM est inférieur à celui des PST.

	MCM0	MCM1	MCM2	MCM3	MCM4	PST
% détection correcte	13.9	53.0	81.3	89.5	90.0	85.8
Nbre de paramètres	$2,2\ 10^1$	$4,0\ 10^2$	$3,3\ 10^3$	$9,9\ 10^3$	$1,8\ 10^4$	$5,1\ 10^4$

FIG. 1 – Taux de détection correcte sur l'ensemble des séquences de SWISSPROT et nombre de paramètres pour les MCM d'ordre 0 à 4 avec lissage Kneser-Ney et les PST

4 Conclusion

Nous avons montré que l'utilisation d'un modèle basé sur des mélanges de chaînes de Markov lissées permet la détection de domaines dans les séquences de protéines, avec un taux de détection supérieur au modèle basé sur des

mélanges de chaînes de Markov élaguées récemment proposé. L'intérêt de ces modèles est un faible coût calculatoire, à la fois pour l'estimation et la détection, et des performances proches de celles de HMM [3]. Une comparaison avec les HMM sur un critère de détection absolument identique (par exemple l'iso-point de la vraisemblance des séquences) serait nécessaire. De plus, il serait intéressant de tester le comportement des MCM lissées sur des bases contenant des protéines plus éloignées, situations dans lesquelles les HMM sont moins performants.

Références

[1] BAIROCH (A.), « PROSITE : A dictionary of sites and patterns in proteins », *Nucleic Acids Research*, *19*, "1991, p. 2241–2245.

[2] BAIROCH (A.) et APWEILER (R.), « The SWISS-PROT protein sequence data bank and its new supplement TrEMBL », *Nucleic Acids Research*, *24*, 1996, p. 21–25.

[3] BEJERANO (G.) et YONA (G.), « Variations on probabilistic suffix trees : statistical modeling and prediction of protein families », *Bioinformatics*, *17*, n° 1, Jan 2001, p. 23–43.

[4] EDDY (S.), *HMMER user's guide : biological analysis using profile hidden Markov models*, Department of Genetics, Washington University School of Medecine, 1998, http ://hmmer.wustl.edu/.

[5] ESKIN (E.), GRUNDY (W.) et SINGER (Y.), « Protein family classification using sparse markov transducers », dans *Proc. Int. Conf. on Intelligent Systems for Molecular Biology*, August 2000.

[6] JELINEK (F.) et MERCER (R.), « Interpolated estimation of markov source parameters from sparse data », dans GELSEMA (E.) et KANAL (L.), *Pattern recognition in practice*, Amsterdam, 1980, North-Holland, p. 381–397.

[7] KATZ (S. M.), « Estimation of probabilities from sparse data for the language model component of a speech recognizer », *IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-35*, n° 3, mars 1987, p. 400–401.

[8] KNESER (R.) et NEY (H.), « Improved backing-off for M-gram language modeling », dans *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Detroit, MI, mai 1995, p. 181–184.

[9] LIDSTONE (G.), « Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities », *Trans. Fac. Actuar.*, *8*, 1920, p. 182–192.

[10] NEY (H.), ESSEN (U.) et KNESER (R.), « On structuring probabilistic dependencies in stochastic language modelling », *Computer Speech and Language*, *8*, 1994, p. 1–38.

[11] RON (D.), SINGER (Y.) et TISHBY (N.), « The power of amnesia : Learning probabilistic automata with variable memory length », *Machine Learning*, n° 25, 1996, p. 117–150.