Détection de séquences atypiques basée sur un modèle de Markov d'ordre variable

Cécile Low-Kam*, Anne Laurent**
Maguelonne Teisseire**

*I3M, Univ. Montpellier 2 - CNRS, Pl. Eugène Bataillon, Montpellier, France
cecile.lowkam@math.univ-montp2.fr
**LIRMM, Univ. Montpellier 2 - CNRS, 161, rue Ada, Montpellier, France
{laurent,teisseire}@lirmm.fr

Résumé. Récemment, le nombre et le volume des bases de données séquentielles biologiques ont augmenté de manière considérable. Dans ce contexte, l'identification des anomalies est essentielle. La plupart des approches pour les extraire se fondent sur une base d'apprentissage ne contenant pas d'outlier. Or, dans de très nombreuses applications, les experts ne disposent pas d'une telle base. De plus, les méthodes existantes demeurent exigeantes en mémoire, ce qui les rend souvent impossibles à utiliser. Nous présentons dans cet article une nouvelle approche, basée sur un modèle de Markov d'ordre variable et sur une mesure de similarité entre objets séquentiels. Nous ajoutons aux méthodes existantes un critère d'élagage pour contrôler la taille de l'espace de recherche et sa qualité, ainsi qu'une inégalité de concentration précise pour la mesure de similarité, conduisant à une meilleure détection des outliers. Nous démontrons expérimentalement la validité de notre approche.

1 Introduction

Un outlier est défini dans (Hawkins (1980)) comme "*une observation qui s'écarte tellement des autres qu'elle est susceptible d'avoir été générée par un mécanisme différent*". Ces dernières années, la détection d'outliers a été étudiée pour des types de données très divers.

En effet, les applications associées à la découverte d'anomalies sont très nombreuses, dans des domaines aussi variés que la détection de fraudes ou l'analyse de séquences biologiques. Parmi elles, les bases d'ADN et de protéines ont fait l'objet de nombreuses études pour une meilleure compréhension des phénomènes biologiques, par exemple par l'extraction de motifs (Ferreira et Azevedo (2007)). La perspective d'identifier des anomalies peut alors compléter les propositions actuelles.

Mais effectuer cette recherche demeure problématique, puisque les outliers sont rares par définition. De plus, ils ne doivent pas être confondus avec le bruit inhérent à tout jeu de données. Néanmoins, certaines propositions existent et nous pouvons citer celles fondées par exemple sur des tests de discordance, sous l'hypothèse d'une distribution de probabilités des observations donnée, dans le cadre univarié ou multivarié (Barnett et Lewis (1994)). D'autres

se basent également sur la notion de distance, afin de détecter les anomalies dans le cas de données multidimensionnelles (Knorr et Ng (1998)).

Toutefois, ces méthodes ne sont pas adaptées à certaines bases biologiques qui sont particulières par leur structure séquentielle et leur taille importante. L'enjeu est alors de sélectionner un modèle approprié. Une approche très efficace pour la découverte d'anomalies dans de telles bases a été proposée dans (Sun et al. (2006)). Elle se fonde sur un modèle d'arbre de suffixes et l'introduction d'une mesure de similarité. Mais cette méthode peut être très exigeante en mémoire, et nécessite une base de séquences typiques pour construire un modèle. Nous proposons donc dans cet article d'étendre cette approche afin de pallier ces inconvénients. En particulier, nous utilisons un critère d'information qui sélectionne un modèle adéquat et parcimonieux. En ce qui concerne la mesure de similarité, nous obtenons des bornes plus précises au-delà desquelles les séquences sont considérées comme atypiques. Ceci permet une meilleure détection des anomalies dans la base considérée.

2 Extraction d'anomalies dans les bases séquentielles

Dans cette section, nous décrivons une approche pour l'extraction d'anomalies dans les bases séquentielles. Nous considérons des bases de données de séquences de la forme $s = s_1...s_\ell$ de ℓ lettres. On note $P^T(s)$ la probabilité de s dans la base. Pour $2 \leq i \leq \ell$, $P^T(s_i|s_1...s_{i-1}) = \frac{P^T(s_1...s_i)}{P^T(s_1...s_{i-1})}$ est la probabilité que s_i suive $s_1...s_{i-1}$. L'hypothèse sur laquelle se base l'approche de Sun *et al.* est que les séquences d'ADN ou de protéines possèdent une propriété de "*mémoire courte*" (Ron et al. (1996)) : il existe un entier $1 \leq K \leq i-1$ tel que

$$P^T(s_i|s_1...s_{i-1}) = P^T(s_i|s_{i-K}...s_{i-1}). \quad (1)$$

Autrement dit, la valeur de la séquence au temps i ne dépend que des K valeurs précédentes. Il s'agit d'une propriété de Markov d'ordre K. Elle est dite d'ordre variable car K n'est pas fixé. Une représentation usuelle d'un tel modèle est un *arbre de suffixes* où chaque nœud a pour parent son plus grand suffixe. Dans un tel arbre, chaque feuille représente une mémoire de la chaîne de Markov associée. Ce modèle est à l'origine de nombreuses applications, telles que la classification en familles des séquences de protéines (Bejerano et Yona (1999)). En effet, il permet d'estimer la probabilité de chaque sous-séquence de la base.

2.1 Arbre Probabiliste des Suffixes

Supposons que nous avons un ensemble S de séquences sur un alphabet fini Σ. Un arbre probabiliste des suffixes, ou *Probabilistic Suffix Tree* (PST), est un arbre des suffixes classique, muni de probabilités conditionelles associées à chaque nœud (Ron et al. (1996)). Plus précisément, chaque nœud est associé à une séquence s de la base. Il contient le nombre d'occurences de s dans la base et un vecteur de longueur $|\Sigma|$ des probabilités conditionnelles $P^T(\sigma|s)$ pour tout $\sigma \in \Sigma$. Comme la taille de l'arbre augmente de façon exponentielle avec la longueur des mémoires, il est élagué. A cet effet, une longueur maximale L est fixée pour l'arbre. Et les nœuds de fréquence faible sont également élagués, étant considérés comme négligeables.

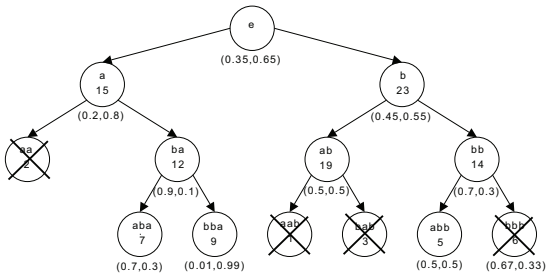

FIG. 1 – *Un exemple de PST sur l'alphabet binaire.*

Exemple 2.1 *La figure* 1 *montre un PST construit sur l'alphabet* $\{a, b\}$*. On pose une longueur maximale* $L = 3$*. Il s'agit de l'ordre maximum de la chaîne de Markov. On fixe également la fréquence minimale des séquences dans la base à* 4*. Les nœuds* aa*,* aab *et* bab *sont alors élagués car ils sont rares.*

Une fois l'arbre construit, les probabilités conditionnelles associées à ses nœuds sont utilisés pour estimer la distribution de chaque séquence, puisque pour toute séquence $s = s_1...s_\ell$, $P^T(s_1...s_\ell) = P^T(s_\ell|s_1...s_{\ell-1}) \times ... \times P^T(s_2|s_1) \times P^T(s_1)$.

Exemple 2.2 *En utilisant le PST de l'exemple précédent,*

$$\begin{aligned} P^T(aabb) &= P^T(b|aab) \times P^T(b|aa) \times P^T(a|a) \times P^T(a) \\ &= P^T(b|ab) \times P^T(b|a) \times P^T(a|a) \times P^T(a) \\ &= 0.50 \times 0.80 \times 0.20 \times 0.35. \end{aligned} \tag{2}$$

Une fois la probabilité de chaque séquence calculée, une mesure de similarité est introduite afin de distinguer les observations atypiques.

2.2 Mesure de similarité et théorie de l'information

Dans (Sun et al. (2006)), pour chaque séquence $s = s_1...s_\ell$, la mesure de similarité SIM_N est définie par :

$$SIM_N(s) = \frac{1}{\ell} \log P^T(s_1...s_\ell). \tag{3}$$

Afin de pouvoir estimer la similarité de n'importe quelle nouvelle séquence, les probabilités du PST sont lissées. La mesure SIM_N est normalisée, et n'est donc pas biaisée par la longueur de la séquence. De plus, sous certaines hypothèses, SIM_N possède une intéressante propriété asymptotique : supposons que les séquences sont générées par une source d'information. Cela signifie qu'elles sont à valeurs dans un alphabet fini, et qu'elles ont une distribution stationnaire (qui ne varie pas au cours du temps).

Rappelons aussi que l'entropie d'une variable aléatoire est une mesure de régularité (Ash (1990)). Cette notion s'étend facilement au cas de deux variables ou plus, les concepts de distribution jointe et conditionnelle menant à ceux d'entropie jointe et conditionnelle. L'incertitude d'une source d'information est alors définie comme la limite de l'entropie conditionelle.

Supposons enfin que les séquences de la base S ont été générées par une unique source d'information, alors, d'après le théorème de Shannon-McMillan (Shannon (1948)), sous la condition d'ergodicité de la source, $-SIM_N$ converge vers l'incertitude de la source. Une preuve de ce résultat se trouven dans (Ash (1990)). Ainsi, quand ℓ est grand, $SIM_N(s)$ devrait être près de l'incertitude de la source, si s a réellement été générée par elle. Sinon, $SIM_N(s)$ sera loin des similarités des autres séquences. Par conséquent, l'inégalité de concentration de Bienaymé-Tchebycheff est utilisée pour déterminer des bornes au-delà desquelles les séquences sont susceptibles d'être des anomalies. Mais cette inégalité est moins performante pour les points éloignés de la moyenne qui sont précisément les outliers potentiels.

Dans (Sun et al. (2006)), des expérimentations sur des bases de protéines sont menées avec succès. Mais elles reposent sur une connaissance préliminaire de séquences typiques, puisque seules celles-ci sont utilisées pour construire le PST. En effet, tout d'abord, un modèle est construit sur une base de séquences, puis les auteurs déterminent si de nouvelles séquences sont des outliers par rapport à ce modèle. Cependant, dans notre approche, nous souhaitons extraire directement les anomalies de l'ensemble des séquences, puisque nous ne connaissons pas les séquences typiques. Donc, bien que la méthode présentée dans cette section soit très efficace pour mettre en évidence les différences de structures entre les familles de protéines, elle échoue lorsque l'on souhaite identifier les outliers parmi une base de séquences. De plus, bien qu'en partie réduite par un élagage selon la fréquence, la taille de l'arbre demeure problématique.

Par conséquent, notre proposition a pour but l'amélioration de l'approche de (Sun et al. (2006)). Et plus particulièrement, dans cet article, nous déterminons si une séquence s est un outlier étant donnés une base S et un seuil t.

3 Vers une approche plus générique

Dans cette section, nous détaillons notre approche. Plus précisément, nous introduisons :

- Un élagage supplémentaire de l'arbre avec un critère d'information, conduisant à une découverte systématique des anomalies, grâce à un modèle adéquat et réduit.
- L'utilisation d'une inégalité de concentration exponentielle pour la mesure de similarité, résultant en des seuils plus précis et ainsi en une meilleure séparation entre outliers et séquences typiques.

Nous adoptons les mêmes hypothèses de mémoire courte et de stationnarité, et nous considérons également le cas d'une alphabet Σ fini.

3.1 Elagage du PST avec le critère d'information d'Akaike

Nous avons vu dans la section 2 qu'un PST peut être élagué en deux étapes :

- Une longueur de branche maximale L est fixée. Tout nœud situé au-delà de L dans l'arbre est élagué.
- Tout nœud pour lequel la fréquence de la séquence associée est inférieure à un seuil donné est élagué.

En plus de ces deux procédures, dans (Ron et al. (1996)) un PST est construit de la façon suivante : un nœud est ajouté à l'arbre s'il diffère statistiquement de son père. Un critère basé sur l'information de Kullback-Leibler est utilisé à cet effet. L'information ou distance de Kullback-Leibler est parfois appelée entropie relative, et représente l'information perdue

quand une distribution est utilisée pour approximer une autre (Burnham et Anderson (1998)). La statistique d'erreur pour une lettre σ et une séquence s est définie dans (Ron et al. (1996)) par : $Err(\sigma s, s) = P^T(\sigma s) \times \sum_{\sigma' \in \Sigma} P^T(\sigma' | \sigma s) \log \frac{P^T(\sigma' | \sigma s)}{P^T(\sigma' | s)}$. Si $Err(\sigma s, s)$ est supérieur à un seuil donné, le nœud correspondant à σs est ajouté à l'arbre

Ainsi, l'information supplémentaire apportée par un fils à son père peut être mesurée. Comme l'information de Kullback-Leibler originelle est pondérée par la probabilité de σs, les nœuds correpondant aux séquences dont les probabilités d'obervation sont faibles sont élagués, qu'ils différent ou non de leur père. Etant rares, ils sont considérés comme négligeables. Dans (Sun et al. (2006)), seul ce dernier critère est appliqué. Mais ce qui est vrai à un niveau de l'arbre ne l'est pas nécessairement aux niveaux suivants, les distributions pouvant différer à une profondeur supérieure. Par conséquent, tous les descendants potentiels de chaque nœud élagué sont aussi testés.

Exemple 3.1 *Considérons à nouveau l'arbre binaire de la figure* 1. *On fixe le seuil à* 0.1. $Err(abb, bb) = 0.001$*, et* $Err(abb, bb) = 0$*. Les nœuds* abb *et* bbb *sont alors élagués, puisque leurs vecteurs de probabiliés conditionnelles sont similaires à celui de leur père, et qu'ils n'apportent donc aucune connaissance supplémentaire.*

La méthode d'élagage ci-dessus n'est pas utilisée dans (Sun et al. (2006)). Par conséquent, afin de réduire davantage la taille de l'arbre, nous utilisons un critère appelé *An Information Criterion* (AIC) et introduit dans (Akaike (1973)). Ce critère permet de trouver un compromis entre l'ajustement d'un modèle aux données et sa complexité. Soit $\mathcal{L}$ la fonction de vraisemblance d'un modèle, et k le nombre de ses paramètres. Alors ce critère est défini comme suit : $AIC = 2k - 2 \log \mathcal{L}$. L'AIC est relié à l'information de Kullback-Leibler. Il se base également sur la vraisemblance, et contient un terme supplémentaire afin de corriger le biais d'estimation asymptotiquement. Ce critère permet de comparer la distance de deux modèles potentiels emboîtés du "vrai" modèle inconnu, puis de choisir le plus proche (voir (Burnham et Anderson (1998)) pour plus de détails). Ainsi, dans un ensemble de modèles candidats, nous choisissons celui pour lequel l'AIC est le plus faible. En pratique, l'AIC peut être peu performant si le nombre de paramètres est élevé par rapport à la taille de la base de données. Ce problème est notamment soulevé dans (Sugiura (1978)). Par conséquent le critère d'information du second ordre est défini dans Hurvich et Tsai (1989) par :

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}, \tag{4}$$

où n est la longueur totale des données. L'AIC_c est adapté quel que soit le nombre de paramètres du modèle. Pour cela, nous utilisons toujours cette version corrigée dans nos applications.

Nous appliquons ce critère en deux étapes. Tout d'abord, notons M_L le modèle de Markov d'ordre fixe L. Nous choisissons le meilleur modèle au sens du critère parmi l'ensemble $\{M_L, L \geq 0\}$. Ainsi, une longueur maximale pour l'arbre est fixée. Ensuite, nous appliquons le même critère à chaque niveau père-fils : soit M_p le modèle basé sur le père et M_s celui basé sur ses fils, alors, $\Delta AIC_c = AIC_c(M_p) - AIC_c(M_s)$ exprime la différence entre ces deux modèles. Nous ajoutons tous ses fils à un père si cette différence est significative (Burnham et Anderson (1998)). Sinon, nous n'ajoutons aucun des fils à leur père.

Ainsi, nous obtenons un modèle de Markov d'ordre variable. Nous introduisons le critère d'Akaike corrigé dans l'algorithme proposé dans (Sun et al. (2006)). Nos expérimentations

montrent que la taille de l'arbre diminue considérablement avec ce nouveau critère, et que la qualité de discrimination est améliorée. En résumé, non seulement notre modèle possède un fondement statistique, mais il nous permet aussi de réduire le nombre de nœuds dans l'arbre. Dans la littérature, le critère d'Akaike est souvent associé à un autre critère connu, appelé critère d'information de Bayes (BIC). Il a été introduit dans (Schwarz (1978)). La différence entre les deux critères concerne le terme de correction, puisque $BIC = \log(n)k - 2\log\mathcal{L}$, où n est la longueur totale des données. Comme le sujet de notre travail n'est pas la comparaison de ces deux critères, nous nous bornons à remarquer qu'ils sont employés à des fins différentes : le but de l'AIC est de parvenir au meilleur compromis entre biais et variance, alors que le modèle sélectionné par le BIC converge vers le "quasi-vrai" modèle, le modèle parmi l'ensemble des candidats qui est le plus proche du vrai. Nous utilisons les deux critères, et obtenons des résultats comparables.

Une fois ce modèle sélectionné, les probabilités conditionnelles du PST sont utilisées afin de calculer $SIM_N(s)$ pour chaque séquence s de S.

3.2 Des bornes plus précises pour la concentration de SIM_N

Dans la section précédente, nous avons vu comment les mesures de similarités sont calculées à partir du PST. Puis, pour détecter les outliers dans la base, l'inégalité de Bienaymé-Tchebycheff est utilisée dans (Sun et al. (2006)) : soient $\mathbb{E}(SIM_N)$ et $\mathbb{V}(SIM_N)$ l'espérance et la variance de la variable aléatoire SIM_N. Alors, $\mathbb{P}\{|SIM_N - \mathbb{E}(SIM_N)| \geq t\} \leq \frac{\mathbb{V}(SIM_N)}{t^2}$ pour tout $t > 0$.

Les outliers sont les séquences dont la similarité se trouve loin des autres, en dehors des bornes définies ci-dessus. Mais, bien que satisfaisante pour les points proches de la moyenne, cette inégalité est connue pour être peu adéquate pour les observations loin de la moyenne, *i.e.* les outliers potentiels. Pour ces derniers, les inégalités de concentration de type exponentiel sont particulièrement adaptées. Nous nous intéressons en particulier à l'inégalité de Bennett (1962) :

Théorème 3.1 *Soient $X_1, ..., X_\ell$ des variables aléatoires réelles indépendantes et centrées, et telles que $|X_i| \leq c$ avec une probabilité de un. Soit $S_\ell = \sum_{i=1}^{\ell} X_i$ et $\sigma^2 = \frac{1}{\ell}\sum_{i=1}^{\ell}\mathbb{V}(X_i)$. Alors pour tout $t > 0$,*

$$\mathbb{P}\{S_\ell > t\} \leq \exp\left(-\frac{\ell\sigma^2}{c^2}\, h\left(\frac{ct}{\ell\sigma^2}\right)\right), \tag{5}$$

où la fonction h est définie pour $u \geq 0$ par $h(u) = (1+u)\log(1+u) - u$.

Nous considérons les variables $X_i = \log P^T(s_i|s_1...s_{i-1})$, $1 \leq i \leq \ell$. Elles sont bornées puisque les probabilités conditionnelles de l'arbre sont lissées. Comme l'inégalité de Bennett est de type exponentiel, nous obtenons des résultats de concentration plus précis pour la mesure de similarité. En effet, nos expérimentations montrent qu'en ce qui concerne la détection d'outliers, les bornes obtenues avec l'inégalité de Bennett sont meilleures que celles issues de l'inégalité de Bienaymé-Tchebycheff.

4 Expérimentations et analyse

Afin de valider notre approche expérimentalement, nous avons considéré la base de données (Bateman et al. (2000)), qui contient environ 9300 familles de protéines, sur l'alphabet des acides aminés de taille 20. Pfam est connue pour couvrir de nombreuses familles de protéines (Ferreira et Azevedo (2007)). Nous utilisons le logiciel R (Team (2006)), muni du package Bio3D (Grant et al. (2006)) pour lire les données dans le format FASTA. Dans (Sun et al. (2006)), il a été judicieusement observé qu'une bonne mesure de similarité devrait détecter la différence de structure entre deux familles. Pour cela, un PST est construit sur une famille, et les similarités de chaque séquence sont calculées afin d'obtenir des bornes. Ensuite, le même arbre est utilisé pour le calcul des similarités des membres des autres familles, afin de savoir combien d'entre eux se trouvent à l'extérieur des bornes. Nous avons mené des expérimentations similaires, en comparant les résultats obtenus avec ou sans élagage selon l'AIC, et en utilisant les inégalités de Bienaymé-Tchebycheff et de Bennett. Toutes ces méthodes nous donnent des résultats satisfaisants, ce qui suggère que tous les modèles de Markov d'ordre raisonnable fonctionnent bien à cet effet. Cependant, notre but ici est de détecter quels sont les outliers parmi un ensemble de séquences, sans savoir quels membres sont typiques et devraient donc être utilisés pour construire l'arbre.

Nous considérons alors la famille HCV_core de la base Pfam, qui compte plus de 3000 membres, auxquels nous ajoutons quelques séquences appartenant à la famille NADHdh. Dans cet article, nous présentons les résultats obtenus pour de tels jeux de données. Le premier, que l'on note D_1, contient 30 séquences issues de la famille NADHdh, soit environ 1% de la base. Le deuxième jeu, appelé D_2, contient 300 séquences de NADHdh, représentant 10% environ du total. Nous construisons un PST sur ces jeux de données, et vérifions si la mesure de similarité distingue bien les séquences atypiques (elles devraient être de la famille NADHdh).

Tout d'abord, nous sélectionnons le modèle global (l'ordre maximal de la chaîne de Markov) en utilisant l'AIC_c et le BIC. Nous considérons quatre modèles de Markov d'ordre croissant $\{M_L, 0 \leq L \leq 3\}$. Le tableau 1 montre les résultats obtenus pour D_1. D'après les deux

Modèle	AIC_c	BIC
M_0	10.2×10^5	10.2×10^5
M_1	6.4×10^5	6.4×10^5
M_2	1.6×10^5	2.9×10^5
M_3	3.2×10^5	5.5×10^5

TAB. 1 – *Critères d'information pour des modèles de Markov d'ordre* 0 *à* 3.

critères, nous choisissons le modèle d'une chaîne de Markov d'ordre 2 correspondant au score le plus faible. Intéressons-nous aux histogrammes des similarités obtenues avec M_0, M_1, M_2 et M_3. La figure 2 montre une estimation de la distribution des similarités des séquences typiques et atypiques, selon le modèle utilisé. M_2 sépare le mieux les deux groupes de mesures de similarité, et donc une inégalité de concentration adéquate devrait permettre de pouvoir identifier les outliers parmi les données, comme nous le verrons plus tard. Au contraire, les autres modèles laissent les deux groupes "déborder" l'un sur l'autre, ce qui rend la distinction difficile. Nous voyons ainsi que le modèle le plus complexe de la liste n'est pas adéquat, pas

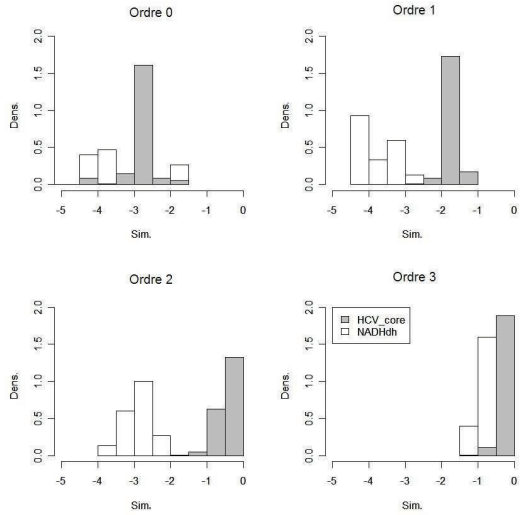

FIG. 2 – *Comparaison des modèles de Markov d'ordre* 0 *à* 3.

plus qu'un modèle trop simple tel que M_0. Le choix du modèle doit se faire sur la base d'un critère approprié.

Procédons maintenant à la deuxième étape de notre stratégie d'élagage. Nous avons vu dans la section 3 qu'une fois la profondeur maximale pour l'arbre trouvée, nous pouvons également utiliser le critère localement, pour chaque nœud. Nous élaguons le PST selon l'AIC_c au niveau local, et obtenons un histogramme similaire à celui du modèle M_2(figure 3). L'arbre possède désormais 312 nœuds au lieu de 368. Comme les anomalies sont mises en évidence de manière comparable, et que le coût de calcul du critère local est important (il s'agit d'une somme sur tout l'alphabet), la question de l'utilité de cette seconde étape se pose. Cependant, lorsque l'on a affaire à de très grands alphabets, il peut être souhaitable de sélectionner un modèle de Markov niveau par niveau, sans avoir d'abord à fixer une profondeur maximale, puis construire tout l'arbre, et enfin élaguer les nœuds le cas échéant. Nous construisons donc le PST sur la même base mais en utilisant le seul critère ΔAIC_c, ce qui signifie que nous ne fixons pas de longueur maximale pour les mémoires. Le PST ainsi obtenu possède 515 nœuds seulement pour une profondeur maximale de 3, et montre la même efficacité pour détecter les outliers, comme le montre l'histogramme de la figure 4. En résumé, lorsque nous élaguons l'arbre des suffixes avec ce seul critère local, nous obtenons un modèle parcimonieux et qui détecte bien les anomalies. Cependant, il est habituellement recommandé de d'abord sélectionner un modèle global (Burnham et Anderson (1998)). Cette approche doit donc être utilisée avec précaution.

Une fois que le modèle a été choisi, nous cherchons à obtenir des bornes pour déterminer si une observation est un outlier par rapport à un seuil donné. Pour cela, sous le modèle M_2, nous appliquons l'inégalité de Bennett avec les seuils correspondants à la réelle proportion

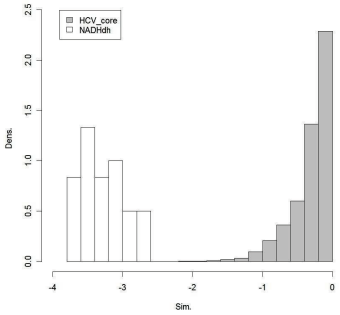

FIG. 3 – *Histogramme obtenu en utilisant le critère local pour $L = 2$.*

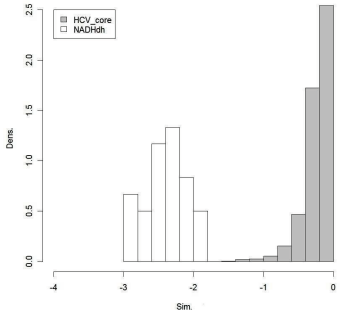

FIG. 4 – *Histogramme obtenu en utilisant le critère local sans profondeur maximale.*

d'anomalies dans D_1 et D_2. Puis nous comparons ces bornes à celles obtenues avec l'inégalité de Bienaymé-Tchebycheff avec un seuil de 11% comme il est recommandé dans (Sun et al. (2006)),ce qui revient à fixer la borne à 3 écarts-type de la moyenne. Nous obtenons les résultats des tableaux 2 et 3 qui contiennent les pourcentages d'anomalies extraites pour chaque base selon chaque méthode. Pour le premier jeu de données, les deux inégalités mènent à des résul-

Inégalité	Seuil	Vrais	Faux
Bienaymé-Tcheb.	0.11	100.0	0.3
Bennett	0.01	100.0	0.5

TAB. 2 – *Pourcentages de vrais et faux outliers hors des bornes pour D_1.*

Inégalité	Seuil	Vrais	Faux
Bienaymé-Tcheb.	0.11	5.0	0.7
Bennett	0.10	100.0	4.0

TAB. 3 – *Pourcentages de vrais et faux outliers hors des bornes pour D_2.*

tats similaires. Cependant, le seuil utilisé pour l'inégalité de Bennett a plus de sens puisqu'il correspond à la proportion d'outliers de D_1. Mais, en général, il est impossible de savoir à l'avance combien il y a de séquences atypiques dans la base. Toutefois, un histogramme tel que ceux de la figure 2 donne une indication. Pour le deuxième jeu de données, l'inégalité de Bennett est clairement plus performante que celle de Bienaymé-Tchebycheff, pourtant à un seuil très proche. Pour D_2, de meilleurs résultats pourraient être obtenus en changeant le seuil de l'inégalité de Bienaymé-Tchebycheff, d'autant plus que les 3 écarts-type n'ont pas de fondement théorique. En effet, la figure 5 montre les bornes obtenues par les inégalités de Bennett et de Bienaymé-Tchebycheff pour les mêmes seuils. La ligne en pointillés représente la similarité la plus grande des outliers. Nous voyons qu'elle correspond à un seuil de 0.06 pour l'inégalité de Bennett, et de 0.29 pour celle de Bienaymé-Tchebycheff. Par conséquent, nous pourrions utiliser cette dernière avec ce seuil pour détecter les anomalies. Cependant le seuil de l'inégalité de Bennett correspond à l'intuition que l'on peut avoir à propos de ces outliers, puisqu'il nous informe que leurs similarités auraient eu moins de 6% de chances de se trouver en dehors des bornes si elles avaient été typiques.

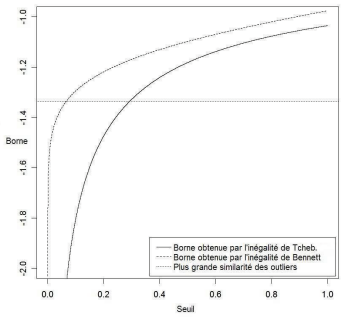

FIG. 5 – *Bornes pour les deux inégalités de concentration pour tous les seuils.*

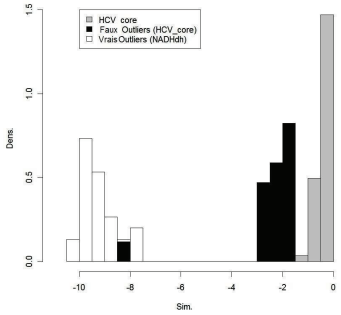

FIG. 6 – *Mesures de similarité pour les séquences typiques, les vrais et les faux outliers.*

Seuils	Nombre de nœuds	Outliers non détectés
5	1318	0.37
10	782	0.03
15	603	0.0

TAB. 4 – *Résultats avec un critère d'élagage basé sur la fréquence pour D_1.*

Le pourcentage de vrais outliers détectés est optimal. Les résultats pour les fausses anomalies est très correct (moins de 4%), cependant quelques séquences typiques demeurent en dehors des bornes, par exemple, 17 par l'inégalité de Bennett pour D_1. Cela représente le coût de la bonne performance de l'inégalité pour les vrais outliers. Pour remédier à ce problème, nous recommençons le même processus de contruction et d'élagage de l'arbre, mais sur une base réduite dont les vrais et faux outliers ont été enlevés. Ensuite, nous regardons à quel point les observations mises à l'écart sont mises en évidence par le nouveau modèle. Les similarités des faux outliers, excepté une seule séquence, sont maintenant beaucoup plus proche de ceux des séquences typiques, alors que les similarités des vrais outliers se détachent clairement. La figure 6 illustre ce résultat pour D_1.

Finalement, l'élagage du PST avec le critère AIC_c, couplé à une inégalité de concentration adaptée, donne des résultats satisfaisants. Mais dans (Sun et al. (2006)), un critère d'élagage basé sur la seule fréquence était utilisé. On peut donc se demander s'il ne pourrait pas mener à une détection comparable. Le tableau 4 montre les résulats d'un tel élagage pour D_1 et une profondeur d'arbre $L = 4$, en utilisant l'inégalité de Bennett au seuil 1%. Pour des seuils inférieurs à 15, tous les outliers ne sont pas détectés. Pour un seuil supérieur à 15, toutes les anomalies sont en dehors des bornes, mais l'arbre peut être plus grand que celui obtenu en élaguant avec le critère d'information. En résumé, cette méthode mène à des résultats comparables et parfois même meilleurs pour la taille de l'arbre. Cependant, aucune information sur le seuil n'est donnée, pourtant ce dernier dépend de la taille de notre jeu de données, du nombre

d'anomalies, et de la structure même des séquences. En effet, pour D_2, en fixant le seuil de fréquence minimale à 15 et en utilisant l'inégalité de Bennett au seuil 10%, 66% des outliers ne sont pas détectés. La qualité de la détection est donc très variable alors que l'élagage selon un critère d'information permet d'identifier systématiquement les outliers.

Dans cette section, nous avons présenté les résultats de notre approche sur des bases de protéines. Notons que nous avons également mené des expérimentations aussi efficaces sur d'autres familles de la base Pfam, et obtenu des résultats également pertinents.

5 Conclusion et perspectives

Dans cet article nous avons proposé une méthode pour détecter des anomalies dans les bases de séquences. Nous avons déterminé si une obervation est atypique selon une base S et un seuil t. Notre approche est une extension de celle de (Sun et al. (2006)) : elle consiste à construire un arbre de suffixes sur la base et à utiliser une mesure de similarité. En effet, lorsque les données sont générées par une unique source d'information, la convergence de cette mesure vers l'incertitude de la source nous assure qu'elle est appropriée. Cependant, à la fois la taille de l'arbre et l'extraction exacte des outliers demeuraient problématique. Par conséquent, nous avons étendu cette méthode à travers un élagage supplémentaire de l'arbre grâce au critère d'information d'Akaike, afin de réduire sa taille et d'améliorer le modèle, et à par l'utilisation de l'inégalité de concentration de Bennett afin de borner plus précisément la mesure de similarité. Ces ajouts ont permis une détection plus efficace car systématique des anomalies. En effet, la qualité de la discrimination a été améliorée alors que la taille de l'arbre est contrôlée. Pour nous en assurer, nous avons testé notre méthode sur des bases de séquences de protéines. Ainsi, nous avons posé des bases pour la détection d'anomalies dans un cadre plus général, afin de permettre l'extension de notre méthode à des structures de données plus complexes, telles que les motifs séquentiels dans les séquences de données (Agrawal et Srikant (1995)).

Remerciements : Nous remercions le Pr. André Mas pour ses conseils avisés. Merci également à Yoann Pitarch pour son aide pour les expérimentations.

Références

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In P. S. Yu et A. S. P. Chen (Eds.), *Eleventh International Conference on Data Engineering*, Taipei, Taiwan, pp. 3–14. IEEE Computer Society Press.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In . F. Petrox, B.N. & Caski (Ed.), *Second International Symposium on Information Theory*, pp. 267–281.

Ash, R. (1990). *Information Theory*. Dover Publications.

Barnett, V. et T. Lewis (1994). *Outliers in Statistical Data*. John Wiley.

Bateman, A., E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, et E. L. Sonnhammer (2000). The pfam protein families database. *Nucleic Acids Res. 28*, 263–266.

Bejerano, G. et G. Yona (1999). Modeling protein families using probabilistic suffix trees. In S. Istrail, P. Pevzner, et M. Waterman (Eds.), *Proc. 3rd Ann. Conf. Computational Molecular Biology (RECOMB)*, Lyon, France, pp. 15–24. ACM Press.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association 57*, 33–45.

Burnham, K. P. et D. R. Anderson (1998). *Model Selection and Inference : A Practical Information-Theoretic Approach*. Springer-Verlag Telos.

Ferreira, P. G. et P. J. Azevedo (2007). Chapter vi : Deterministic motif mining in protein databases. In F. Masseglia, P. Poncelet, et M. Teisseire (Eds.), *Successes and New Directions in Data Mining*.

Grant, B., A. Rodrigues, K. ElSawy, J. McCammon, et L. Caves (2006). Bio3d : An r package for the comparative analysis of protein structures. *Bioinformatics 22*, 2695–2696.

Hawkins, D. (1980). *Identification of Outliers*. Chapman and Hall.

Hurvich, C. M. et C. L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika 76*(2), 297–307.

Knorr, E. M. et R. T. Ng (1998). Algorithms for mining distance-based outliers in large datasets. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pp. 392–403.

Ron, D., Y. Singer, et N. Tishby (1996). The power of amnesia : Learning probabilistic automata with variable memory length. *Machine Learning 25*(2-3), 117–149.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal 27*, 379–423 and 623–656.

Sugiura, N. (1978). Further analysis of the data by akaike's information criterion and the finite corrections. *Communications in Statistics : Theory and Methods 7*, 13–26.

Sun, P., S. Chawla, et B. Arunasalam (2006). Mining for outliers in sequential databases. In *Proc. 6th SIAM Int. Conf. Data Mining*, pp. 94–105.

Team, R. D. C. (2006). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Summary

Recently, biological sequential databases have increased in both size and number. In this context, identifying the outliers is essential. To extract them, most of approaches use a sample of known typical sequences to build a model. However, such a database is not often at hand. Besides, the existing methods remain greedy in terms of memory usage. In this paper we propose a new approach, based on a variable order markov model and on a measure of similarity. We add to existing methods a pruning criterion to control the size of the search space and its quality, and a sharp inequality for the concentration of the measure of similarity, to better sort the outliers. We prove the feasability of our approach through a set of experiments.