Fouille de données du génome à l'aide de modèles de Markov cachés

Sébastien Hergalant * **, Bertrand Aigle *
Pierre Leblond*, Jean-François Mari**

*Laboratoire de Génétique et Microbiologie, UMR-UHP-INRA, IFR 110,
54506 Vandœuvre-lès-Nancy, France
{bertrand.aigle,pierre.leblond}@nancy.inra.fr,
**LORIA UMR-CNRS 7503, 54506 Vandœuvre-lès-Nancy, France
{hergalan,jfmari}@loria.fr
http ://www.loria.fr/~jfmari/ACI/

Résumé. Nous décrivons un processus de fouille de données en bioinformatique. Il se traduit par la spécification de modèles de Markov cachés du second-ordre, leur apprentissage et leur utilisation pour permettre une segmentation de grandes séquences d'ADN en différentes classes qui traduisent chacune un état organisationnel et structural des motifs d'ADN locaux sous-jacents. Nous ne supposons aucune connaissance *a priori* sur les séquences que nous étudions. Dans le domaine informatique, ce travail est dédié à la définition d'observations structurées (les k-d-k-mers) permettant la localisation en contexte d'irrégularités, ainsi qu'à la description d'une méthode de classification utilisant plusieurs classifieurs. Dans le domaine biologique, cet article décrit une méthode pour prédire des ensembles de gènes co-régulés, donc susceptibles d'avoir des fonctions liées en réponse à des conditions environnementales spécifiques.

1 Introduction

L'accumulation des séquences issues des projets de séquençage oblige la mise en œuvre de méthodes de fouille de données efficaces pour comprendre les mécanismes impliqués dans l'expression, la transmission et l'évolution des gènes. Nous nous intéressons aux modèles stochastiques et méthodes classificatoires permettant de prédire les séquences promotrices et autres petites séquences régulatrices chez les bactéries. Une manière de cerner notre ignorance vis à vis des motifs et segments d'ADN impliqués dans les mécanismes décrits plus haut est de modéliser l'évolution et la structuration du génome par des processus stochastiques capables d'apprentissage statistique nécessitant un minimum de connaissances *a priori*. Ces modèles stochastiques sont utilisés comme révélateurs d'organisations locales remarquables qu'un expert doit interpréter.

Nous nous intéressons à la localisation de sites de fixation de protéines. Ces sites de fixation – appelés TFBS (*Transcription Factor Binding Sites*) ou encore promoteurs transcriptionnels – sont constitués de trois séquences adjacentes de nucléotides :

$$N_x\text{—}N_y\text{—}N_z \quad \text{avec} \quad \begin{array}{c} N \in \{A, C, G, T\} \\ 3 \leq x, z \leq 9 \\ 0 \leq y \leq 25 \end{array}$$

N_x et N_z sont susceptibles d'être altérées par quelques substitutions tandis que N_y – appelé espaceur – est une chaîne de composition et de taille variable. L'ensemble N_x—N_y—N_z se situe en amont d'un gène et sert de site de fixation pour une protéine qui vient réguler son expression. Un régulon est un ensemble de gènes possédant nécessairement le même TFBS en amont.

2 Matériel et méthodes

2.1 Définition des HMM2

La modélisation stochastique est une approche mathématique pour prendre en compte la variabilité inhérente aux processus issus du vivant comme le sont la reconnaissance de la parole, ou la segmentation du génome. Un modèle stochastique particulier – le modèle de Markov caché (HMM pour *Hidden Markov Model*) – représente la suite des nucléotides par deux processus stochastiques : l'un caché, prenant ses valeurs sur un ensemble d'états et qui est une chaîne de Markov, l'autre visible prenant ses valeurs parmi les observations physiques : la séquence de nucléotides constituant l'ADN. La variabilité est capturée par la supposition que les observations ne sont pas uniquement associées aux états mais dépendent d'une densité définie sur chaque état. En reprenant les notations introduites par Churchill et Mury (Mury, 1997), nous définissons plus formellement un *HMM2* comme un HMM du type M2-M0 de la façon suivante :

- $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$, un ensemble fini comprenant N états ;
- $\mathcal{A}$ la matrice donnant les probabilités de transitions entre états : $\mathcal{A} = (a_{ijk})$ pour un *HMM2*, avec la contrainte :

$$\sum_k a_{ijk} = 1 \quad \forall\, i, j;$$

- $b_i()$ les lois des densités associées aux états s_i.

La matrice $\mathcal{A}$ est initialisée avec un ensemble de valeurs qui permettent de définir la topologie du graphe des transitions entre états : quelles sont les transitions autorisées ? aller simple ($a_{ijk} > 0$, $a_{kji} = 0$), aller-retour ($a_{ijk} > 0,\ a_{kji} > 0$), bouclage ($a_{ijj} \neq 0$), etc.

La modélisation du génome par des HMM est ainsi fondée sur deux principes : (i) le génome peut être découpé en segments par une chaîne de Markov et (ii) chaque segment est la réalisation d'un processus stationnaire représenté par une densité de probabilité sur l'espace des observations.

En modélisant de la sorte les segments composant le chromosome, on ignore la réalité de la constitution génétique comme résultat d'un processus organisé mais on peut utiliser des algorithmes (Boyer et al., 1990) d'apprentissage et de reconnaissance rapides. Cette façon de procéder est complémentaire d'une approche analytique et explicative fondée sur un mécanisme de raisonnement. En mesurant précisément par une probabilité ce que l'on qualifie au premier abord de hasardeux, on diminue l'indéterminisme de notre perception du processus et on peut faire apparaître des comportements explicables, donc prévisibles, qui pourront être réutilisés dans un mécanisme de raisonne-

ment ; ce mécanisme d'extraction et de réutilisation est un des principes de la fouille de données.

2.2 Les k-mer

Les observations les plus simples modélisées par les densités $b_i()$ sont les 4 nucléotides : $A_1 = \{A, C, G, T\}$. Si on considère les 16 paires possibles de nucléotides, on définit alors A_2 comme l'ensemble des *2-mer*, et ainsi de suite pour l'ensemble de *k-mer* A_k constitué des 4^k séquences de k nucléotides. La suite d'ADN peut être vue comme une chaîne de *k-mer* qui se recouvrent – on appelle l le décalage compté en nombre de bases entre deux *k-mer* successifs dans le chromosome – et qui constitue la suite des observations du HMM.

Par exemple si on traite la suite : TAGGCTAGGTG, avec k=4 et l=1, la chaîne observée est : TAGG-AGGC-GGCT-GCTA-CTAG-TAGG-AGGT-GGTG (1). Avec k=4 et l=4, elle devient TAGG-CTAG (2).

Nous définissons aussi les $k_1 - d - k_2 - mers$ $(1 < k_1 + k_2 < 7)$ constitués de deux *k-mer* espacés par d nucléotides. Il y a $4^{k_1+k_2} k_1 - d - k_2 - mers$ différents ; les d bases constituant l'espaceur n'intervenant pas dans la définition. Par exemple, la séquence TAGGCTAGGTG peut être vue comme une séquence de 2-3-2-mers. Dans ce cas, avec un décalage de 2, nous observons la chaîne TATA-GGGG-CTTG, qui est différente des deux autres chaînes (1) et (2). Cet article montre que les *k-d-k-mers* sont adaptés à la recherche de sites de fixation constitués par deux segments non adjacents de nucléotides situés en amont de gènes.

2.3 Estimation du *HMM2*

Dans cette section, nous décrivons la méthode qui permet de spécifier le *HMM2* qui rende le mieux compte des données au sens du maximum de vraisemblance. Nous ne contrôlons que la topologie du *HMM2*. Les matrices $\mathcal{A}$ et $\mathcal{B}$ (respectivement, les probabilités de la chaîne de Markov et les paramètres des densités) sont estimées par l'algorithme EM (Dempster et al., 1977).

Le but de la modélisation stochastique est d'élaborer un indice capable d'aider l'expert – le bioinformaticien – dans son travail de fouille. On désire calculer un indice concis et expurgé du bruit qui rende compte de l'organisation locale des nucléotides. Le *HMM2* est utilisé comme une machine à segmenter et à classer. Nous nous servons de la probabilité *a posteriori* de la transition $s_i \longrightarrow s_j \longrightarrow s_k$ entre $t-1$ et $t+1$ pendant l'observation de tout le chromosome. La figure 1 représente la probabilité de la transition "boucle" sur un état du *HMM2* dans Artemis (logiciel d'annotation de génome (Rutherford et al., 2000)) en même temps que l'annotation du génome reprenant les résultats des bases de données d'annotation issues du Web. Sur cet exemple, les pics de la probabilité sont des indices permettant à l'expert de retrouver une information connue : la présence de site de fixation d'une protéine régulatrice devant un gène.

FIG. 1 – *Représentation d'une probabilité de segmentation donnée par le HMM2 dans Artemis*

2.4 Spécification de la topologie du *HMM2*

Nous nous limitons à l'étude de *HMM2* ergodiques, c'est à dire dans lesquels toutes les transitions entre états sont possibles. Il s'agit donc d'identifier le nombre d'états susceptibles de donner la classification la plus utile pour l'expert.

La distance entre deux états est représentée par la divergence entre les deux densités $b_i()$ et $b_j()$.

$$div(i,j) = \int_x b_i(x) \ln \frac{b_i(x)}{b_j(x)} dx + \int_x b_j(x) \ln \frac{b_j(x)}{b_i(x)} dx$$

Nous utilisons des *HMM2* de 3 à 6 états utilisant des observations du type *k-mer* (k = 2, 3, 4, 5, 6) et/ou 2-d-2, 3-d-3 avec 0 ¡ d ¡ 25. Le déplacement entre deux *k-mer* est toujours de $l = 1$. A partir d'un *HMM2* de trois états dans lequel toutes les transitions sont équiprobables et les densités des lois uniformes, nous utilisons l'algorithme Forward-backward pour obtenir une estimation selon le maximum de vraisemblance. La divergence entre toutes les paires d'états est calculée. Si la plus grande divergence est inférieure à un seuil donné, nous recommençons l'apprentissage avec un modèle possédant un état supplémentaire. Nous cherchons à faire apparaître ainsi des états captant des irrégularités locales dans la distribution des nucléotides donc bien différents d'un état rendant compte d'une organisation "moyenne". Une situation dans laquelle on n'arrive pas à faire apparaître un état différent des autres est une situation d'échec de classification par les *HMM2*. Nous tentons alors, un autre apprentissage en envisageant une autre définition des observations. Cette situation n'a pas été rencontrée sur les corpus étudiés.

Deux *HMM2*, appelés *HMM2+* et *HMM2-* sont construits : l'un pour le sens direct, l'autre pour le sens complémentaire inverse du chromosome. (les deux brins sont symétriques et s'enroulent pour former la double hélice d'ADN constituant le génome).

L'ensemble des programmes est écrit en C++ et dérive du logiciel[1] CarrotAge (Le Ber et al., 2004) utilisé initialement en agronomie pour la recherche de successions de cultures.

2.5 Recherche automatique des pics

La recherche des pics de probabilité *a posteriori* pour un état donné se fait en utilisant une fenêtre glissante de 200 pb (paires de bases) se recouvrant de 100 pb avec la fenêtre voisine. Dans chaque fenêtre sont calculées des statistiques (moyenne, variance, min et max). La construction du modèle a été faite de telle sorte qu'un état au moins possède des fluctuations de cette probabilité. Un pic est défini par sa valeur maximum qui doit être significativement supérieure à la moyenne calculée dans la fenêtre et par sa longueur p qui doit vérifier l'inégalité : $2k - 1 < p < 12$, k étant la taille des *k-mer* observés. La valeur 12 correspond à la largeur maximum d'un pic recherché. En effet, les *HMM2*– bien que supérieur aux *HMM1* (Mari et al., 1997) dans ce domaine – ont une piètre capacité à modéliser les durées de longs segments. Sous chaque pic, nous extrayons un motif qui doit être ensuite classé.

2.6 Clustering des motifs

La classification des motifs se fait à l'aide de deux programmes effectuant des alignements entre chaînes. Le premier (Multalin (Corpet, 1988)) effectue une classification hiérarchique de motifs sous-jacents aux pics fondée sur la distance calculée entre deux motifs. Ce programme autorise l'alignement de séquences courtes sans pénalités aux extrémités. Le nombre maximum de substitutions de nucléotides peut être spécifié et les trous interdits. Le résultat de Multalin est un ensemble de classes représentées par leur consensus. Le deuxième, Fasta effectue une classification incrémentale sur les paires de motifs espacés comme le montre l'algorithme 1.

Algorithm 1 *Algorithme de recherche de TFBS.*

for all classe trouvée par Multalin **do**
  construit le consensus de cette classe
  recherche dans le génome entier les occurrences de cette séquence
end for
Sélectionne les paires d'occurrences (s_1, s_2) espacées de d bp. Appelle ce triplet $s_1 + s_d + s_2$
Recherche dans tout le génome les occurrences de $s_1+s_d+s_2$. Ceci élimine les (s_1, s_2) non espacés convenablement
Effectue un alignement par Fasta entre les paires de séquences $s_1 + s_d + s_2$ pour permettre un regroupement de séquences légèrement dégénérées.
Sélectionne celles qui apparaissent plus de 3 fois
Sélectionne $s_1 + s_d + s_2$ trouvée dans les régions intergéniques et/ou à moins de 600 bp en amont d'une ORF
Regroupe dans une classe tous les gènes trouvés en aval d'une séquence $s_1 + s_d + s_2$

[1] Licence publique Gnu

3 Expérimentations

3.1 Le génome d'étude

Les *Streptomyces* sont des bactéries filamenteuses du sol qui revêtent un intérêt économique fort compte tenu de l'importance des produits de leur métabolisme secondaire. Elles sont en effet la source principale d'antibiotiques parmi les micro-organismes. L'espèce *Streptomyces coelicolor* A3(2) présente un chromosome linéaire de 8,7 méga paires de bases (8,7 Mb), qui se caractérise par un taux global en bases G + C de 72%. Chez *S. coelicolor*, environ 12% des 7825 gènes prédits (ou ORF pour *Open Reading Frame*) se répartissent dans différentes familles de régulateurs transcriptionnels. Les facteurs sigma y représentent une classe particulièrement importante, avec 65 gènes prédits chez *S. coelicolor* et 60 chez *S. avermitilis*. Moins de dix facteurs sigma ont été étudiés jusqu'à présent chez les *Streptomyces*, et seulement quelques régulons (ensembles de gènes co-régulés) ont été définis par des approches expérimentales en biologie. De même, un nombre très faible de régulateurs transcriptionnels (activateurs ou répresseurs) a été étudié jusqu'alors. La très grande majorité des motifs nucléotidiques sur lesquels agissent ces facteurs de transcription reste donc à définir.

Notre approche bio-informatique vise à extraire et classifier les motifs nucléotidiques présents dans les régions intergéniques des génomes des bactéries. Cette approche exhaustive permettra de définir des motifs régulateurs en faisant abstraction du bruit notamment généré par la superposition de motifs dans les gènes à régulation complexe. La validation biologique expérimentale permettra ensuite de définir les gènes co-régulés.

L'algorithme d'extraction des irrégularités locales fait apparaître 4 fois plus de pics dans les régions intergéniques que dans les gènes. Bon nombre de pics se situe au voisinage des sites de fixation des protéines intervenant dans la régulation de l'expression des gènes. Ce sont des séquences d'ADN situées en amont des gènes, organisées en paires espacées, et qui sont spécifiquement reconnues par les protéines régulatrices (les facteurs de transcription comme les facteurs sigma) en question. Chaque régulateur transcriptionnel possède sa séquence d'ADN cible qui lui est propre. Celle-ci est définie sous la forme d'un consensus. Les gènes co-régulés par un facteur de transcription possèdent donc tous la même séquence en amont. Après l'extraction de ces pics, notre travail a consisté à classer les segments nucléotidiques sous-jacents pour voir dans quelles mesures ils pourraient être des TFBS connus ou à découvrir.

3.2 Validation de la méthode sur un régulon connu : SigR

Chez *S. coelicolor*, SigR est un des 65 facteurs sigma prédits ou connus. Il est impliqué à un niveau clé dans les mécanismes de réponse au stress oxydant, en co-régulant (positivement ou négativement) 30 gènes de cette bactérie. Ceci a été montré expérimentalement (Paget et al., 2001). La séquence consensus reconnue par SigR est $GGGAAT - N_{18} - GTTN$ (structure type $s_1 + s_d + s_2$). En nous basant sur ces résultats, nous avons testé nos algorithmes sur 3 génomes bactériens du groupe des actinomycètes : *S. coelicolor*, *S. avermitilis* et *Mycobacterium tuberculosis*, proche parente des deux premières et pathogène pour l'homme (vecteur de la tuberculose). Nous utilisons pour cela des *HMM2* à *3-mers* ou *3-d-3-mers* ($0 \leq d \leq 25$), qui représentent le

mieux la réalité du code biologique et fournissent les meilleurs résultats en termes quantitatifs (ni trop, ni trop peu de pics) et qualitatifs (localisation, spécificité et largeur des pics).

En amont des 30 gènes régulés par SigR chez *S. coelicolor*, 84 motifs sont extraits. Parmi ceux-ci, 47 décrivent les motifs types s_1 et s_2. Les 30 motifs s_1 trouvés permettent d'inférer la position des 30 promoteurs reconnus par SigR (table 1). Le motif s_2 n'est pas toujours extrait. Ceci est dû au fait que la séquence est trop petite pour pouvoir être modélisée de façon convenable par un *HMM2* utilisant des *3-mers* dans leur définition.

Sur la figure 1, nous observons une segmentation globalement homogène le long de la séquence, interrompue par la présence d'hétérogénéités locales dans les régions intergéniques. Cadrons notre attention sur la région intergénique en amont du gène *folE*. Celui-ci est un des 30 gènes contrôlés par SigR. Deux des trois pics décelables dans cette région se trouvent aux positions des motifs s_1 et s_2 (appelés respectivement boîtes -35 et -10). Le dernier pic a une signification inconnue.

Parmi les 37 pics ne décrivant pas les promoteurs reconnus par SigR, 22 décrivent d'autres promoteurs pour ces 30 gènes ou d'autres motifs de fixation dont le sens biologique est démontré.

Génome	*Nombre de gènes*	s_1	s_2	TFBS détectés
S. coelicolor	30 (SigR)	30	17	30
S. avermitilis	23 (SigR)	23	15	23
M. tuberculosis	13 (SigH)	13	10	13

TAB. 1 – *Détection par HMM2 de promoteurs connus chez 3 bactéries actinomycètes.*

Chez *S. avermitilis*, les promoteurs SigR n'ont pas été définis mais un test de similarité de séquences permet de proposer un jeu de 23 gènes homologues à ceux régulés par SigR chez *S. coelicolor*. L'hypothèse de départ est que les mécanismes de régulation sont vraisemblablement comparables entre génomes voisins. Cette modélisation permet de détecter les 23 promoteurs de type SigR correspondant (cf. table 1).

Chez *M. tuberculosis*, SigH est le régulon homologue à SigR (Manganelli et al., 2002). Il comporte au moins 13 gènes possédant en amont une séquence reconnue par SigH (très similaire de la séquence de type SigR). Ces 13 promoteurs sont également extraits avec cette méthode (table 1).

Enfin, les motifs s_1 et s_2, pris isolément, sont largement distribués sur les génomes étudiés. Cependant, ces *HMM2* ne réagissent qu'en présence de la paire de motifs convenablement espacés, dans un environnement nucléotidique intergénique spécifique que l'on pourrait qualifier de "région promotrice". C'est cette propriété remarquable qui fait tout l'attrait de cette méthode.

3.3 Classification des motifs extraits de *S. coelicolor*

Pour cette étude, les modèles *HMM2+* et *HMM2-* utilisent des *3-mers* avec $l = 1$. Nous considérons une portion de 1/8 du génome de *S. coelicolor* (1,15 Mb) pour laquelle 3000 motifs intergéniques ont été extraits à partir des 2 *HMM2*. Les classes trouvées

par la méthode de clustering décrite dans cet article représentent 229 pics/3000 et permettent de prédire :

- Des promoteurs reconnus par des facteurs sigma connus (SigB, WhiG, SigR).
- Des promoteurs reconnus par des régulateurs transcriptionnels qui font actuellement l'objet d'expérimentations biologiques encore non publiées (régulon PhoR / PhoP, définition d'une classe plus large pour le régulon SigR). Les résultats obtenus *in vitro* confirment la validité de cette méthode *in silico*.
- Des promoteurs potentiels reconnus par des régulateurs transcriptionnels hypothétiques et déjà prédits par d'autres méthodes de recherche de motifs promoteurs (Li et al., 2002; Studholme et al., 2004).
- Cinq régulons potentiels entièrement nouveaux. Deux d'entre eux sont potentiellement régulés par deux facteurs sigma hypothétiques pour lesquels il est possible de déduire une fonction biologique. Une des deux classes de gènes situés en aval des séquences extraites serait impliquée dans les processus de sporulation et de développement de *S. coelicolor*. L'autre contrôlerait plusieurs régulateurs transcriptionnels différents pour fournir une réponse plus généralisée.

L'apport de ces connaissances est primordiale pour le biologiste, qui voit son panel de conditions expérimentales restreint, et peut ainsi entreprendre de tester ces résultats.

4 Conclusions et Perspectives

Nous proposons ici une nouvelle méthodologie sans *a priori* pour construire et utiliser des *HMM2* permettant de localiser des sites de fixation de protéines régulatrices dans les séquences d'ADN des procaryotes. Sur trois génomes de bactéries actinomycètes, l'étude de régulons connus a montré la capacité qu'ont ces *HMM2* à décrire de courts motifs d'ADN (5 à 12 pb) riches en sémantique biologique. La classification de ces motifs permet également de prédire des nouvelles classes de gènes co-régulés, potentiellement testables par le biologiste.

Cependant, cette méthode repose sur une classification hiérarchique elle-même basée sur l'alignement multiple de séquences courtes. Ceci est mal réalisé par les méthodes courantes d'alignement de séquences, dont le degré de similarité est calculé à partir de scores obtenus pour un nombre limité de symboles. La classification utilisée est séquentielle et chaque classifieur utilise des données propres. Dans cette optique, une meilleure définition des consensus de classe, ainsi qu'une fusion de résultats classificatoires sont envisagées. A plus long terme, l'adaptation de la méthode à des génomes de structuration différentes sera une démarche importante à mettre en œuvre.

Références

Boyer, A., Martino, J. D., Divoux, P., Haton, J.-P., Mari, J.-F., and Smaili, K. (1990). Statistical Methods in Multi-Speaker Automatic Speech Recognition. *Applied Stochastic Models and Data Analysis*, 6(3) :143–155.

Corpet, F. (1988). Multiple Sequence Alignment with Hierarchical Clustering. *Nucl. Acids Res*, 16(22) :10881–10890.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum-Likelihood From Incomplete Data Via The EM Algorithm. *Journal of Royal Statistic Society, B (methodological)*, 39 :1 – 38.

Le Ber, F., Mari, J.-F., Benoît, M., Mignolet, C., and Schott, C. (2004). Carrotage, a software for mining land-use data. In *Fourth International Workshop on Environmental Applications of Machine Learning - EAML'2004, Bled, Slovenia.*

Li, H., Rhodius, V., Gross, C., and Siggia, E. D. (2002). Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl. Acad. Sci. USA*, 99(18) :11772–11777.

Manganelli, R., Voskuil, M. I., Schoolnik, G. K., Dubnau, E., Gomez, M., and Smith, I. (2002). Role of the extracytoplasmic-function sigma factor sigma(H) in *Mycobacterium tuberculosis* global gene expression. *Molecular Microbiology*, 45(2) :365–374.

Mari, J.-F., Haton, J.-P., and Kriouile, A. (1997). Automatic Word Recognition Based on Second-Order Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, 5 :22 – 25.

Mury, F. (1997). *Comparaison d'algorithmes d'identification de chaînes de Markov cachées et application à la détection de régions homogènes dans les séquences d'ADN.* Thèse de doctorat, Université René Descartes, Paris V.

Paget, M. S. B., Molle, V., Cohen, G., Aharonowitz, Y., and Buttner, M. J. (2001). Defining the disulphide stress response in *Streptomyces coelicolor* A3(2) : identification of the σ^R regulon. *Molecular Microbiology*, 42(4) :1007–1020.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M., and Barrell, B. (2000). Artemis : sequence visualization and annotation. *Bioinformatics*, 16(10) :944–945.

Studholme, D. J., Bentley, S. D., and Kormanec, J. (2004). Bioinformatic identification of novel regulatory DNA sequence motifs in *Streptomyces coelicolor*. *BMC Microbiol.*, 4(1) :14.

Summary

We describe a data mining process in Bioinformatics based on second-order hidden Markov models. After an automatic, unsupervised training, these models perform a segmentation of a whole genome in several classes that contain specific genetic features. We do not make any *a priori* assumption on the genetic organisation. This work is devoted to the definition of structured motifs (the k-d-k mers) that allow the localization of irregularities. It also describes a method to predict co-regulated sets of genes.