Statistiques de motifs

Sophie S. Schbath

[21] G. Reinert, S. Schbath, and M. S. Waterman. Probabilistic and statistical properties of words : an overview. *J. Comput. Biol.*, 7 :1–46, 2000.
[22] F. Sanger. The Croonian Lecture, 1975. Nucleotide sequences in DNA. *Proc. R. Soc. Lond., B, Biol. Sci.*, 191 :317–333, Dec 1975.
[23] C. L. Sawyers. The cancer biomarker problem. *Nature*, 452 :548–552, Apr 2008.
[24] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270 :467–470, Oct 1995.
[25] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.*, 31 :64–68, May 2002.
[26] B. J. Strasser. Genetics. GenBank–Natural history in the 21st Century ? *Science*, 322 :537–538, Oct 2008.
[27] B. J. Strasser, L. Pauling, and F. Crick. A world in one dimension : Linus Pauling, Francis Crick and the central dogma of molecular biology. *Hist Philos Life Sci*, 28 :491–512, 2006.
[28] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 85(1) :267–288, 1996.
[29] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.
[30] N. Verzelen. Minimax risks for sparse regressions : Ultra-high-dimensional phenomenons. Technical report, Arxiv, 2010.
[31] M. Via, C. Gignoux, and E. G. Burchard. The 1000 Genomes Project : new opportunities for research and social challenges. *Genome Med*, 2 :3, 2010.
[32] M.S. Waterman, T.F. Smith, and W.A. Beyer. Some biological sequence metrics. *Adv. Math.*, 20 :367–387, 1976.
[33] J.D. Watson and F.H.C. Crick. A structure of deoxyribonucleic acid. *Nature*, 171 :964–967, 1953.

Statistiques de motifs

Sophie Schbath [1]

Cet article traite de l'identification de motifs d'ADN fonctionnels le long des génomes par des approches statistiques. Par motif d'ADN, on entend une courte suite de lettres (généralement pas plus d'une quinzaine) dans l'alphabet des nucléotides $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, et par motif fonctionnel, on entend un motif dont les occurrences sur le génome seront reconnues par une protéine qui se fixera alors sur l'ADN pour entrer en action.

L'identification de motifs d'ADN fonctionnels reste un problème biologique encore loin d'être résolu pour plusieurs raisons : (i) leur longueur est variable selon la nature de la protéine, (ii) la reconnaissance de chacune des lettres par la protéine peut être imprécise, autrement dit il n'y a pas nécessairement unicité d'une lettre à chaque position, (iii) l'activité protéique peut dépendre de la présence à proximité d'autres protéines elles-mêmes reconnaissant d'autres motifs ADN, (iv) les motifs fonctionnels ne sont généralement pas conservés d'un organisme à l'autre.

Bien souvent, ces motifs fonctionnels (appelés simplement motifs par la suite) se caractérisent par des répartitions bien particulières le long du génome, d'où la construction de critères ou scores dont on cherchera à mesurer la significativité : on cherche en effet à repérer des événements (ici des occurrences de motifs) qui

[1] INRA, Unité Mathématique, Informatique et Génome, Jouy-en-Josas, France.

auraient très peu de chance de se produire au hasard. Le « hasard » sera déterminé par des séquences aléatoires $X_1X_2\cdots X_\ell$ dont les lettres X_i seront tirées dans l'alphabet $\mathcal{A}$ selon un modèle probabiliste plus ou moins sophistiqué.
Typiquement, on utilise des modèles de chaîne de Markov d'ordre m : la probabilité de générer la lettre $b \in \mathcal{A}$ à la position i dépend des m lettres[2] $a_1a_2\cdots a_m$ qui précèdent, c'est-à-dire aux positions $i-m,\ldots,i-1$. Ces probabilités, dites *de transition*, des lettres $a_1a_2\cdots a_m$ vers la lettre b sont estimées à partir de la composition en mots de taille $m+1$ de la séquence d'ADN analysée. Ce modèle a l'énorme avantage de comparer ce que l'on observe dans la séquence d'ADN étudiée avec ce que l'on pourrait attendre dans des séquences aléatoires ayant en moyenne la même composition en lettres mais aussi en mots de tailles 2, 3, ..., $(m+1)$.

Par exemple, certains motifs se caractérisent par une fréquence anormalement élevée[3] sur le génome entier, ou seulement sur une partie du génome. Pour découvrir d'autres motifs potentiels ayant la même propriété statistique, on est amené à étudier la loi de probabilité du comptage d'un mot dans une chaîne de Markov (cf. section 1) et regarder ceux ayant un comptage significativement élevé (probabilité critique proche de zéro).
D'autres motifs se caractérisent par leur présence dans certaines régions caractéristiques du génome, par exemple en amont des gènes[4]. Si l'on considère alors toutes[5] les sous-séquences de quelques centaines de lettres situées en amont des gènes, cela se traduit par un nombre anormalement élevé de ces sous-séquences contenant au moins une occurrence du motif en question. Dans ce cas, on est amené à évaluer la probabilité qu'un motif donné soit présent (peu importe son nombre d'occurrences) dans une chaîne de Markov (cf. section 2).

D'autres types de questions statistiques relatives aux occurrences de motifs existent mais ne seront pas traitées dans ce dossier. On pourra cependant noter que la formalisation d'un certain nombre d'entre elles utilise non plus un modèle de séquences aléatoires, mais des processus ponctuels pour modéliser les occurrences elles-mêmes. On citera par exemple l'utilisation de processus de Poisson, pour détecter des régions anormalement riches ou pauvres en certains motifs ou pour tester si deux séquences sont aussi riches en un motif donné. Ou encore l'utilisation de processus de Hawkes pour détecter si les occurrences d'un ou plusieurs motifs présentent des distances favorisées ou évitées.

1. Comptage attendu ou anormal ?

Le problème est le suivant : on observe par exemple 762 occurrences du motif de longueur 8 gctggtgg dans une séquence d'ADN de longueur $\ell = 4\,638\,858$ (en fait le génome complet de *E. coli*) et on se demande si ce comptage ne serait

[2] La valeur de m est choisie par le modélisateur et permet de fixer en moyenne la composition des séquences aléatoires jusqu'aux mots de taille de $m+1$.

[3] C'est le cas du motif gctggtgg avec 762 occurrences le long du génome de la bactérie *Escherichia coli* long de $4.6\,10^6$ lettres, ou du motif aagtgcggt avec 740 occurrences le long du génome de la bactétrie *Haemophilus influenzae* de longueur $1.8\,10^6$.

[4] C'est le cas des sites de fixation des facteurs de transcription indispensables à la transcription des gènes en ARN.

[5] Plusieurs milliers, autant que le nombre de gènes par organisme.

pas significativement élevé. Intuitivement, en effet, si les 4^8 mots possibles de taille 8 avaient la même fréquence dans la séquence, on s'attendrait à les observer chacun 72 fois. Pour savoir si l'écart entre 762 (l'observé) et 72 (l'attendu) est significatif, il faut calculer la probabilité de l'événement $\{N \geqslant 762\}$ sous le modèle choisi, appelée *probabilité critique* (ou *p-value* en anglais). On se placera ici dans le modèle de chaîne de Markov d'ordre m ($0 \leqslant m \leqslant h-2$, où h est la longueur du mot étudié) qui permet de s'ajuster sur la composition de la séquence d'ADN en mots de taille 1 à $m+1$.

Le calcul de cette probabilité serait trivial si on connaissait la distribution du comptage N, mais cette dernière est complexe à obtenir du fait que l'on compte des occurrences qui potentiellement peuvent se chevaucher[6] dans la séquence. En effet, le comptage N est une somme de variables aléatoires de Bernoulli Y_i ($Y_i = 1$ si le motif est présent à la position i, et 0 sinon) non indépendantes. Sa loi n'est donc pas une loi binomiale comme l'on aurait pu être tenté de dire.

Plusieurs approches ont été proposées soit pour calculer exactement la probabilité critique soit pour l'approcher. L'une des approches exactes consiste à calculer la distribution du temps d'attente T_n de la n-ième occurrence du mot, pour tout $1 \leqslant n \leqslant 762$, puis d'utiliser le principe de dualité suivant : $\mathbb{P}(T_{762} \leqslant \ell) = \mathbb{P}(N \geqslant 762)$. La distribution exacte du temps d'attente s'obtient par récurrence [4] ou via sa fonction génératrice [10] qu'il convient ensuite de développer en série de Taylor. Néanmoins, la valeur exacte de la probabilité critique n'est réellement calculable numériquement que pour des séquences de quelques dizaines de milliers de lettres et pour des ordres de modèles très faibles, 0 ou 1, ce qui en limite grandement l'usage pour l'analyse de génomes entiers.

Des approximations de la probabilité critique sont donc plutôt utilisées en pratique. Deux types d'approches ont été poursuivies : d'une part celles visant à approcher la distribution du comptage par des lois paramétriques explicites ; d'autre part celles qui approchent directement la queue de la distribution par des approches de grandes déviations [1]. Ces dernières sont en effet pertinentes lorsque les motifs sont très exceptionnels. Parmi les distributions approchées, on peut noter la loi gaussienne pour les motifs plutôt fréquents [2], des lois de Poisson composées pour les motifs plutôt rares [8], voire la loi binomiale pour des motifs non périodiques (dont les occurrences ne peuvent jamais se chevaucher).

Si l'on reprend le problème mentionné au départ de ce paragraphe, il s'avère que la probabilité critique d'observer au moins 762 occurrences du motif `gctggtgg` dans une chaîne de Markov de longueur $\ell = 4\,638\,858$ et ayant en moyenne la même composition en mots de taille 1 à 7 que le génome de *E. coli*, vaut environ $8.7\,10^{-27}$ (en utilisant l'approximation gaussienne). Autrement dit, ce motif est significativement fréquent le long de ce génome (seul deux autres motifs de taille 8 ont une probabilité critique encore plus petite). D'un point de vue biologique, ce motif est vital à la bactérie puisqu'il intervient dans la réparation du génome en cas de cassure des brins d'ADN.

[6] Par exemple, le motif périodique `atgatga` a deux occurrences chevauchantes dans la séquence `gtatgatgatgattc` : l'une à la position 3 (soulignée), l'autre à la position 6 (en italique).

2. Présence attendue ou anormale ?

Ici ce qui nous intéresse est de savoir si le fait qu'un motif soit présent dans une séquence (plutôt courte) soit exceptionnel, c'est-à-dire non dû au hasard, ou pas. Pour cela, on peut bien sûr calculer ou approcher la probabilité $\mathbb{P}(N \geqslant 1) = \mathbb{P}(T_1 \leqslant \ell)$ en utilisant les approches décrites dans le paragraphe précédent.

Dans le cadre de la recherche de sites de fixation de facteurs de transcription, on est très vite amené à considérer l'occurrence de 2 motifs $\mathbf{m}_1$ et $\mathbf{m}_2$ (voire plus) à une certaine distance d plus ou moins variable dans une séquence ($d_1 \leqslant d \leqslant d_2$). Par exemple, `ttgactt` suivi de `ataataa` 16 à 18 lettres après. On s'intéresse donc à l'occurrence du motif composite résultant, noté $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$. On pourrait considérer ce motif composite comme un ensemble de longs motifs « simples » (de taille 30 à 32 dans l'exemple) et leur appliquer les résultats précédents, mais la distance est généralement trop grande, produisant un ensemble de motifs « simples » beaucoup trop grand (de l'ordre de 10^{18} dans l'exemple). Il faut donc recourir à des méthodes ad-hoc pour traiter ces motifs composites.

Voici trois approches qui ont été proposées pour évaluer la probabilité qu'un motif composite donné $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$ soit présent dans une chaîne de Markov.

– On peut approcher la probabilité $\mathbb{P}(N = 0) = 1 - \mathbb{P}(N \geqslant 1)$ par le produit $(\mathbb{P}(Y_i = 0))^{\ell-h+1}$ où Y_i vaut 1 si le motif composite (i.e. la première lettre du motif) apparaît à la position i dans la séquence, ou 0 sinon, et h est la taille du motif composite [13] ce qui revient à faire comme si les variables Y_i étaient indépendantes. On peut sensiblement améliorer l'approximation en considérant une dépendance d'ordre 1 entre les variables Y_i ce qui ramène à calculer le produit $\mathbb{P}(Y_1 = 0)(\mathbb{P}(Y_i = 0 \mid Y_{i-1} = 0))^{\ell-h}$ [6][7]. Le calcul de $\mathbb{P}(Y_i = 0)$ pour un motif composite est moins trivial que pour un motif « simple » mais s'obtient à partir de la distribution du temps d'attente entre deux motifs « simples » (problème lié au paragraphe précédent). En effet, le motif composite $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$ apparaît en position i si et seulement si le premier motif $\mathbf{m}_1$ apparaît en position i et le 2$^{\text{e}}$ motif $\mathbf{m}_2$ apparaît à une distance comprise entre d_1 et d_2 après le 1$^{\text{er}}$ motif. Le calcul de $\mathbb{P}(Y_i = 0 \mid Y_{i-1} = 0)$ est plus technique mais faisable pour un motif composite composé de 2 motifs simples (le travail reste à faire pour 3 motifs ou plus).

– On peut décomposer le temps d'attente avant la première occurrence du motif composite comme une somme aléatoire de temps d'attente indépendants entre motifs simples [11]. Il faut néanmoins faire l'hypothèse que chacun des motifs simples ne peut apparaître plus d'une fois dans le motif composite. Dans le cas du motif composite $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$, on obtient la décomposition suivante (cf. figure 1) : il faut attendre le premier motif $\mathbf{m}_1$ puis (**) attendre le prochain motif parmi $(\mathbf{m}_1, \mathbf{m}_2)$; s'il s'agit de $\mathbf{m}_2$, on vérifie si la distance qui le sépare de $\mathbf{m}_1$ est bonne (entre d_1 et d_2) auquel cas on a trouvé le motif composite ; si la distance n'est pas bonne, il faut rechercher le prochain motif $\mathbf{m}_1$ et repartir de (**) ; s'il s'agissait du motif $\mathbf{m}_1$ alors on repart de (**). Connaissant les lois du temps d'attente avant la prochaine occurrence de $\mathbf{m}_1$ et du temps d'attente avant l'occurrence d'un des motifs $(\mathbf{m}_1, \mathbf{m}_2)$ [5], [10], on en déduit la loi du temps d'attente du motif composite. Cette approche n'est cependant valable que pour des motifs composites à 2 motifs.

[7] $P(A|B)$ désigne la probabilité de l'événement A sachant l'événement B.

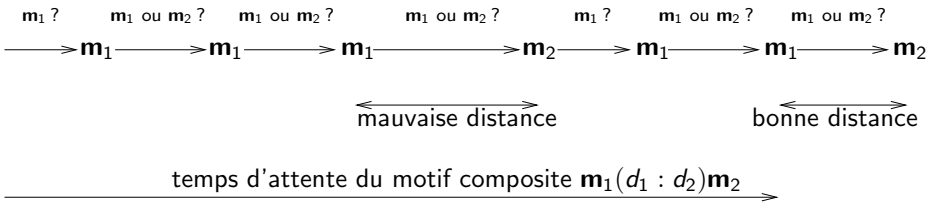

FIG. 1. Décomposition du temps d'attente avant l'occurrence du motif composite $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$ comme la somme de temps d'attente indépendants entre les occurrences de $\mathbf{m}_1$ et $\mathbf{m}_2$.

– L'approche précédente peut se reformuler en termes de construction d'un processus semi-markovien dans lequel les états du processus correspondent aux différentes étapes nécessaires pour obtenir une occurrence du motif composite [12]. Pour 2 motifs il n'y a que 3 états : (*état 1*) le motif $\mathbf{m}_1$ est atteint (l'état suivant peut être l'état 1 ou l'état 2), (*état 2*) le motif $\mathbf{m}_2$ est atteint mais à la mauvaise distance de $\mathbf{m}_1$ (dans ce cas, l'état suivant sera l'état 1) et enfin (*état 3*) le motif $\mathbf{m}_2$ est atteint à la bonne distance de $\mathbf{m}_1$ (cet état est un état absorbant du processus semi-markovien). Si on reprend l'exemple de la figure 1, la succession des états serait la suivante : $\rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 2 \rightarrow 1 \rightarrow 1 \rightarrow 3$. Le temps d'attente du motif composite est donc égal au temps d'absorbance du processus semi-markovien[8] dont on sait calculer les probabilités de transition et les lois des temps de séjour dans chaque état grâce aux lois des temps d'attente entre motifs simples. Cette approche est facilement généralisable à un nombre quelconque de motifs simples.

3. Conclusion

Le lecteur intéressé par plus de détails mathématiques pourra se reporter au livre [7] ou aux chapitres d'ouvrages [3] et [9].

4. Références

[1] NUEL, G. (2004). LD-SPatt : Large Deviations Statistics for Patterns on Markov Chains. *Journal of Computational Biology*, **11**, 1023–1033.

[2] PRUM, B., RODOLPHE, F. and TURCKHEIM, À. (1995). Finding words with unexpected frequencies in DNA sequences, *Journal of the Royal Statistical Society series B*, **57**, 205–220.

[3] REINERT, G., SCHBATH, S. and WATERMAN, M. (2005). *Applied Combinatorics on Words*. volume 105 of *Encyclopedia of Mathematics and its Applications*, chapter Statistics on Words with Applications to Biological Sequences. Cambridge University Press.

[4] ROBIN, S. and DAUDIN, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters, *J. Appl. Prob.* **36,** 179–193.

[5] ROBIN, S. and DAUDIN, J.-J. (2001). Exact distribution of the distances between any occurences of a set of words, *Ann. Inst. Statist. Math.* **36,** 895–905.

[8] Le processus est semi-markovien dans le sens où les lois des temps de séjour dans chaque état ne sont pas géométriques, contrairement au cas markovien.

[6] Robin, S., Daudin, J.-J., Richard, H., Sagot, M.-F. and Schbath, S. (2002). Occurrence probability of structured motifs in random sequences, *Journal of Computational Biology*, **9**, 761–773.
[7] Robin, S., Rodolphe, F. and Schbath, S. (2003). *ADN, mots et modèles*. BELIN.
[8] Schbath, S. (1995). Compound Poisson approximation of word counts in DNA sequences, *ESAIM : Probability and Statistics*, **1**, 1–16.
[9] Schbath, S. and Robin, R. (2009). *Scan Statistics – Methods and Applications*. (J. Glaz, I. Pozdnyakov, and S. Wallenstein, ed.), chapter How can pattern statistics be useful for DNA motif discovery ? Statistics for Industry and Technology. Birkhauser.
[10] Stefanov, V.T. (2003). The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models : an algorithmic approach, *J. Appl. Prob.* **40,** 881–892.
[11] Stefanov, V.T., Robin, S., and Schbath, S. (2007). Waiting times for clumps of patterns and for structured motifs in random sequences. *Discrete Appl. Math.* **155**, 868–880.
[12] Stefanov, V., Robin, S. and Schbath, S. (2011). Occurrence of structured motifs in random sequences : Arbitrary number of boxes. *Discrete Appl. Math.* **159**, 826–831.
[13] Sandve, G.K. and Abul, O. and Drablos, F. (2008). Compo : composite motif discovery using discrete models. *BMC Bioinformatics*, **9** :527.

Segmentation pour l'analyse de puces CGH

Emilie Lebarbier [1] et Franck Picard [2]

Chaque espèce possède un nombre caractéristique de copies des chromosomes : chez la grande majorité des êtres vivants, comme chez l'homme, chaque chromosome est présent en deux copies (organisme diploïde) mais d'autres organismes peuvent être polyploïdes (ayant plus de 2 copies de chaque chromosome), comme par exemple la pomme de terre ou l'huître avec 4 copies. Une déviation de ce nombre de copies par rapport au nombre normal pour l'espèce (surnuméraire ou manquante) entraîne de ce fait un déséquilibre du nombre de copies des gènes, pouvant être à l'origine de maladies majeures. Un exemple classique chez l'humain est la trisomie 21, qui se caractérise, comme son nom l'indique, par la présence de 3 copies du chromosome 21. La détection de ces défauts chromosomiques est donc un élément majeur dans l'établissement du diagnostic et/ou du traitement de la pathologie. Si la détection se place à l'échelle du chromosome, elle est rendue possible grâce au traditionnel caryotype. Cependant, la perte ou le gain peut ne toucher qu'une portion du chromosome. Et lorsque cette anomalie chromosomique est de très petite taille, elle peut passer inaperçue. C'est en 1992 que l'étude de ces « petites » anomalies connaît un essor considérable grâce à la mise au point d'une nouvelle technique, l'hybridation génomique comparative ou CGH. La résolution a été nettement améliorée par l'utilisation de la technologie des microarrays (microarrays CGH ou puces CGH) permettant la détection d'aberrations d'environ 50kb. Après avoir été exclusivement appliquées à l'étude de la génomique du cancer, les microarrays CGH sont aujourd'hui utilisées dans d'autres études de génétique humaine.

[1] UMR Agroparistech/INRA MIA 518.
[2] LBBE, UMR CNRS 5558 Université Lyon 1 Villeurbanne, France.