# Analyse de la distribution de motifs dans le génome et le protéome.

**Calcul scientifique, calcul exact, algorithmique symbolique-numérique.**


| | |
|---|---|
| **Contact :** | Jean-Guillaume.Dumas@imag.fr, tél : 04 76 51 48 66. |
| **Laboratoire :** | Jean Kuntzmann, tour IRMA, BP 53 X |
| | 51, av. des Mathématiques, 38041 Grenoble. http://ljk.imag.fr |
| **Rémunération :** | Standard par le laboratoire. |
| **Poursuite en thèse :** | Thèse en lien avec le projet européen OpenDreamKit.org. |


Afin de déterminer si des occurrences de motifs dans de longues chaînes ont une pertinence statistique, il est nécessaire d'analyser les probabilités d'apparition de ces motifs rapportées à une distribution uniforme. Une approche classique consiste à utiliser un automate et une chaîne de Markov pour reconnaître les motifs puis d'étudier la matrice de transition pour en extraire les informations sur la distribution [5, 6].
Le stage consiste à étudier cette matrice de transition par des méthodes hybrides formelles-numériques. En effet, ces matrices de transition sont structurées (matrices stochastiques, forte proportion de zéros) et mal conditionnées. La difficulté est de déterminer des itérés de ces matrices à des ordres de grandeur très différents pour de très grands itérés (de l'ordre de dizaines de millions pour le protéome, de centaines de millions pour les chromosomes).
L'approche considérée pourra combiner itérations modulaires et polynomiales bivariées [2] avec du calcul approché de pgcd, ou de reconstructions rationnelles [4, 3, 1]. Cette méthode sera ensuite utilisée pour étudier les facteurs de transcription du chromosome humain 5* et des signatures PROSITE† du protéome.

## References


[1] Annie Cuyt and Wen-shin Lee. Sparse interpolation of multivariate rational functions. *Theoretical Computer Science*, 412(16):1445–1456, April 2011, special issue on Symbolic Numeric Algorithms. http://www.researchgate.net/publication/220154902_Sparse_interpolation_of_multivariate_rational_functions/file/79e4151152b8f4c411.pdf

[2] Jean-Guillaume Dumas and Grégory Nuel. Sparse approaches for the exact distribution of patterns in long multi-states sequences generated by a markov source. *Theoretical Computer Science*, special issue on Symbolic Numeric Algorithms, vol. 497, p. 22-42, (2013). http://arxiv.org/abs/1006.3246.

[3] Erich Kaltofen and Zhengfeng Yang. On exact and approximate interpolation of sparse rational functions. In Christopher W. Brown, editor, *ISSAC'2007, Proceedings of the 2007 ACM International Symposium on Symbolic and Algebraic Computation, Waterloo, Canada*, pages 203–210. ACM Press, New York, July 29 – August 1 2007. http://www.math.ncsu.edu/~kaltofen/bibliography/07/KaYa07.pdf

[4] Erich Kaltofen, Zhengfeng Yang, and Lihong Zhi. Approximate greatest common divisors of several polynomials with linearly constrained coefficients and singular polynomials. In Jean-Guillaume Dumas, editor, *ISSAC'2006, Proceedings of the 2006 ACM International Symposium on Symbolic and Algebraic Computation, Genova, Italy*, pages 169–176. ACM Press, New York, July 2006. http://www.math.ncsu.edu/~kaltofen/bibliography/06/KYZ06.pdf

[5] Pierre Nicodème, Bruno Salvy, and Philippe Flajolet. Motif statistics. *Theoretical Computer Science*, 287(2):593–617, September 2002. http://hal.inria.fr/docs/00/07/30/74/PDF/RR-3606.pdf

[6] Grégory Nuel. On the first k moments of the random count of a pattern in a multi-states sequence generated by a Markov sourc. *Journal of Applied Probability*, 47(4):1105-1123, 2010. http://arxiv.org/abs/0909.4071


*http://www.pseudogenes.org/data/human/build36/genome/chr5.fa
†http://expasy.org/prosite