R'MES: Finding exceptional motifs in sequences

Sophie S. Schbath, Mark M. Hoebeke

R'MES: finding exceptional motifs in sequences

Mark Hoebeke and Sophie Schbath

INRA, Unité Mathématique, Informatique et Génome, 78352 Jouy-en-Josas, France.
Mark.Hoebeke@jouy.inra.fr, Sophie.Schbath@jouy.inra.fr.
R'MES home page: http://migale.jouy.inra.fr/outils/mig/rmes
URL for accessing the code: https://mulcyber.toulouse.inra.fr/projects/rmes/

The R'MES project started in 1995. This is now the 3rd version. The main question R'MES addresses is "does this motif occur in that biological sequence with an expected frequency?" In other words, can we observe it so many times, or so few times, just by chance? Usually, when the answer is no, such a motif is a candidate to have a particular biological meaning. To do so, we calculate an exceptionality score for each word of a given length (or for each given set of words); this score is a one-to-one transformation of the corresponding p-value. The p-value is the probability that a random sequence having the same 1- up to $(m+1)$-letter word composition as the biological sequence contains as many occurrences of the given word. This probability is approximated thanks to rigorous statistical approximations of the word count distribution, namely either a Gaussian distribution (for frequent words) or a compound Poisson distribution (for rare words). Details about the statistical results on word counts in random sequences can be found in [1]. R'MES is getting enriched thanks to novel questions from the biologists. R'MES can now for instance compute an exceptionality score related to the skew of an oligonucleotide; the typical question is indeed "does this motif occur significantly more often on the leading strand than on the lagging strand?" At the moment, we are implementing the statistical tests proposed by [2] to compare motif exceptionalities between two different sequences. In the talk, we will illustrate how we have identified the Chi site of *Staphylococcus aureus* [3] and the matS site of *Escherichia coli* [4] thanks to R'MES.

[1] Robin, S., Rodolphe, F. and Schbath, S. (2005). *DNA, Words and Models.* Cambridge University Press.

[2] Robin, S., Schbath, S. and Vandewalle, V. (2007). Statistical tests to compare motif count exceptionalities. *BMC Bioinformatics.* 8:84 1–20.

[3] Halpern, D., Chiapello, H., Schbath, S., Robin, S., Hennequet-Antier, C., Gruss, A. and El Karoui, M. (2007). Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modelling. *PLoS Genetics.* 3(9) e153.

[4] Mercier, R., Petit, M.-A., Schbath, S., Robin, S., El Karoui, M., Boccard, F. and Espeli, O. (2008). The MatP/matS site specific system organizes the Terminus region of the E. coli chromosome into a Macrodomain. *Cell*, volume 135, Issue 3, 475–485.