Detecting motifs in biological networks by local concentration

Etienne Birmelé

Laboratoire Statistique et Génome, UMR CNRS 8071, INRA 1152
Tour Evry 2, 523 place des Terrasses de l'Agora 91000 EVRY France
etienne.birmele@genopole.cnrs.fr

Abstract: Studying the topology of biological networks by using statistical means has become a major field of interest in the last decade. One way to deal with that issue is to consider that networks are built from small functional units called *motifs*, which can be found by looking for small subgraphs whose numbers of occurrences in the whole network of interest are surprisingly high [4]. Some of the existing methods compute a huge number of random networks with the same degree distribution as the biological one [4], but seem not to be tractable for motifs on more than four vertices. Others [2,5] rely on a sampling algorithm and the calculation of a Z-score. Nevertheless, the distribution of the number of small subgraphs is more heavy-tailed than a gaussian and therefore those methods may lead to false positives.

Another issue is that a small graph can appear as over-represented in a network because it contains an over-represented subgraph, which is in fact the biological relevant structure. Moreover, Dobrin and al. [1] show that the motifs in the yeast transcriptional regulatory network aggregate. For both reasons, we will consider a new definition of a motif: given a small graph $\mathbf{m}$ and an occurence in the network of one of its subgraphs $\mathbf{m}'$, we will look for an over-representation of the number of occurences of $\mathbf{m}$ extending the given occurence of $\mathbf{m}'$. In other words, a motif will be defined by a local over-representation rather than by a global one.

In order to avoid time-consuming simulations, we also use a mixture model on graphs as the null model. The parameters of the mixture are estimated using new results on bayesian mixture models for graphs [3].

That approach allows us to developp a statistic which detects motifs avoiding false positives and with a significant improvement in terms of running time. We apply it to the Yeast gene interaction data and show that the known biologically relevant motifs are found again and that our method gives some more informations than the existing ones.

References

[1] R. Dobrin and al. Aggregation of topological motifs in *escherischia coli* transcriptional regulatory network. *BMC Bioinformatics*, 5:10, 2004.

[2] N. Kashtan and al. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20-11:1746, 2004.

[3] P. Latouche, E. Birmelé, and C. Ambroise. Bayesian methods for graph clustering. *preprint SSB*, 2008.

[4] R. Milo and al. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[5] S. Wernicke and F.Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.