Comparison of Jaspar, Transfac and Genomatix motif libraries on chip-seq data for 44 transcription factors

Michal Dabrowski[1]*, Norbert Dojer[2], Izabella Krystkowiak[1], Bartek Wilczynski[2], Bozena Kaminska[1]

[1] Nencki Institute of Experimental Biology, Warsaw, Poland
[2] Faculty of Mathematics Informatics and Mechanics, University of Warsaw, Poland

*presenting author, e-mail: m.dabrowski@nencki.gov.pl

For vertebrates, there are three major collections of TFBS motifs: public Jaspar and commercial Transfac and Genomatix. We compared performance of the three libraries, in terms of coverage, specificity, and sensitivity, using the chip-seq data for 44 transcription factors (TFs) as the positive sets, and third exons as the negative set. Each commercial library was used with its supplied scanner (Match and MatInspector), and all the three libraries were used with the same two open source scanners (Bio.Motif and matrix-scan).

The coverage (number of represented TFs) was highest for Genomatix (37), followed by Transfac (33), and by Jaspar (21). The average specificity and sensitivity was practically identical for all three libraries, when used with the same scanner. The two open-source scanners outperformed the two commercial scanners, by resulting in higher average sensitivity for the same average specificity. The use of Genomatix matrix families busted sensitivity, at the cost of a drop in specificity.

With Bio.Motif scanner, we analyzed the full ROC curves for all the motifs from the three libraries. We investigatied utility of different ways of parametrization of the ROC curves to automatically set the thresholds in a way that maximizes the balanced accuracy (average of specificity and the sensitivity). Our results demonstrate that the optimal value of the threshold is dependent on the information content of the motif.