A phylogenetic footprinting ensemble tool and a new scoring metric for evaluating the prediction of transcription factor binding sites in flies.

Karsten Hokamp, Yuliana Pozdeeva, Juan-Pablo Labrador
Smurfit Institute of Genetics, Department of Genetics, School of Microbiology and Genetics, Trinity College Dublin, Dublin, Ireland

Keywords
Ensemble, phylogenetic footprinting, transcription factor binding sites, scoring metric

Abstract
Motivation: Knowledge about the location of transcription factor binding sites (TFBSs) forms an important step to understand the regulatory machinery of a gene. Many programs exist that predict TFBSs, but their performance is widely variable, and the prediction metrics used do not accurately reflect real performance. Frequent recommendations include the use of phylogenetic footprinting and the combination of results from multiple programs to improve predictions and reduce the number of false positives.
Results: We introduce PhyloFootPrEns, the first ensemble method of phylogenetic footprinting tools for the prediction of transcription factor binding sites. Our ensemble tool is accessible through a user-friendly web-interface, which automatically extracts upstream sequences from any *D. melanogaster* gene and its orthologues and generates intuitive visualisations of the predicted sites. We also present a new scoring metric, the xF-score, which better reflects the wet-lab users' interests compared to traditional metrics. It is based on the F-score, but incorporates a mix of site-level and nucleotide-level measurements. We analysed the 2000 base pairs upstream regions of 18 fly genes. As a benchmark we used the experimentally validated TFBSs listed in the RedFly database combined with additional manually curated sites. Individual programs showed highly varying and often poor predictions. Additionally, individual programs show large performance variations for different genes, which highlights the need for test sets based on real biological rather than synthetic data. Our method achieves, on average, the best and most robust results, compared to those generated by individual tools and is also less variable across genes.