Abstract

Unraveling the mechanisms that regulate gene expression is a major challenge in biology. An important task in this challenge is to identify regulatory motifs or short sequences in the DNA that serve as binding sites for transcription factors (TFs). The first computational methods developed for the discovery of regulatory motifs searched for an overrepresented motif in a set of genes that were believed to contain several binding sites for the same TF (e.g. a set of coregulated genes from a single genome). But with the growing number of sequenced genomes, detecting motifs through 'phylogenetic footprinting' became feasible and the next generation of motif discovery algorithms has therefore integrated the use of orthology evidence in addition to coregulation information. Moreover, the more advanced motif discovery algorithms explicitly model the phylogenetic relatedness between the orthologous input sequences and thus should be well adapted towards using orthologous information. In a first part of the study we evaluated the conditions under which complementing coregulation with orthologous information improves motif discovery for the class of probabilistic motif discovery algorithms with an explicit evolutionary model. We designed specific datasets, both synthetic and real, essential for the benchmarking of motif discovery algorithms that integrate orthologous information. Our results show that the nature of the used algorithm is crucial in determining how to exploit multiple species data in the best way to improve motif discovery performance. The use of an integrated evolutionary model that depends on reliable alignments of hard to align intergenic sequences seems to be the major bottleneck.In a second part of the study we developed a complete workflow for motif discovery in eukaryotes: PHYLO-MOTIF-WEB. This workflow is unique as it allows for integrating epigenetic information (e.g. nucleosome occupancy and histone modifications) to guide the motif search to putative regulatory regions in the DNA, a necessary step considering the long non-coding sequences in eukaryotes. An asymmetric clustering algorithm, FuzzyClustering, was developed to summarize the results of multiple advanced motif discovery algorithms into an ensemble solution. PHYLO-MOTIF-WEB is easy accessible for non-expert users through a web server. Finally, we applied PHYLO-MOTIF-WEB on a biological case to investigate the molecular mechanisms underlying the antiproliferative effects of vitamin D3 on both human and mouse cell lines. We predicted de novo the regulatory motifs of some known TFs that possibly can be involved in the vitamin D3 induced pathway. Further research is necessary to validate those predictions. Our results also show the potential of combining the results of multiple motif discovery algorithms, as a consequence of the diversity in their predictions.