cis-Regulatory Element Prediction in Mammalian Genomes

Asim Siddiqui, Gordon Robertson, Misha Bilenky, Tamara Astakhova, Obi L. Griffith, Maik Hassel, Keven Lin, Stephen Montgomery, Mehrdad Oveisi, Erin Pleasance, Neil Robertson, Monica C. Sleumer, Kevin Teague, Richard Varhol, Maggie Zhang and Steven Jones
Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Research Centre, British Columbia Cancer Agency, Vancouver, BC, Canada
asims@bcgsc.ca

Abstract

The identification of cis-regulatory elements and modules is an important step in understanding the regulation of genes. We have developed a pipeline capable of running multiple motif prediction methods on a whole genome scale.

Using gene expression datasets to identify co-expressed genes and the Ensembl Compara database orthologues, we assemble input sequence sets comprised of the upstream regions of a target gene, its orthologues and co-expressed genes on the premise that such genes will share promoters by evolution (orthologues) or share regulatory control mechanisms (co-expressed genes). Co-expressed genes are identified by an approach that combines Pearson distances from multiple gene expression datasets derived from multiple experimental approaches and calibrated against the GO database. Our pipeline runs a number of established motif detection algorithms with a range of parameter settings on the input dataset. We integrate the diverse result sets by scoring motifs with a method-independent function. For each target gene, we assign p-values to the motif score by running the discovery pipeline on multiple sets of input sequence containing the target gene, non-coexpressed genes and "fake" orthologues generated by neutral numerical evolution.

We have predicted 30,636 motif binding sites in human for 4,182 genes and an initial set of 472 motif binding sites in mouse for 92 genes with $p < 0.001$. The positive predictive value against a library of biologically confirmed regulatory sites approaches 0.4 at the highest p-value threshold.

Predicted regulatory elements and other resources from the project are available at www.cisred.org.

1. Introduction

The identification of cis-regulatory elements, the sites to which transcription factors bind and thereby control gene expression, remains a difficult task. Many methods have been created, each with its own advantages and disadvantages and, in general, these methods have a low positive predicative value (PPV) and low sensitivity (S_n) [1].

Rather than create a new method, we have chosen to focus our efforts on refining the input dataset and developing methods to assess the validity of the analysis. We describe a high-throughput discovery system that predicts regulatory elements for mammalian genomes using a suite of regulatory element prediction algorithms. We have developed a framework for identifying and incorporating co-expressed and orthologous input gene sets.

2. Materials and Methods

2.1 Construction of Input Gene Sets

For each target gene, a set of orthologous and co-expressed genes are identified. Orthologous genes are identified by combining data from a number of databases including ENSEMBL Compara[2] and KEGG[3]. Putative orthologues for unannotated genomes are also included.

We identify co-expressed genes using public gene expression data from many sources, utilizing the Pearson Correlation coefficient to detect genes with similar patterns of expression [4].

The input set comprises the upstream regions of the target gene, the orthologous genes and co-expressed genes are extracted (1500 base pairs after removal of repeat regions). Further work is underway to improve the identification of the transcriptional start site.

2.2 Generation of a Background Model

For each target sequence set, we generate a large set of 'random' sequences. Co-expressed genes are replaced genes with no co-expression relationships to the target gene and synthetic orthologous sequences are generated by DUNE, a neutral evolution simulator that transforms a target sequence without the influence of selective constraints.

The random gene sets are used to provide a measure of the false discovery rate of motifs.

2.3 Parallel discovery runs using multiple methods

We run a number of motif discovery methods (presently Meme[5], Consensus[6], MotifSampler[7]) on the input and random gene sets. Each method is run under a variety of parameter settings to improve sensitivity. This part of the pipeline requires the most CPU and is run over a ~400 CPU computing cluster. Despite these resources, the pipeline requires several weeks to generate genome wide results.

2.4 Scoring and assignment of confidence values to motifs

Discovered motifs from a target set and its random sequence sets are assigned method-independent (MI) scores. The MI-scoring function is optimized against a library of biologically known transcription factor binding sites (KSL). The current KSL is derived from TRANSFAC v9.1 and comprises 758 sites for 177 genes. For each target gene, we use the distribution of MI motif scores from the random set to transform MI scores for target set motifs into p-values.

3. Results and Discussion

Predictive runs have been completed on a set of 4,182 human genes yielding 30,636 binding sites with $p < 0.001$. For these sites, there are 19,387 unique consensus sequences. Further work is underway to improve the clustering of similar motifs.

Figure 1 plots the nucleotide-level specificity and sensitivity and site-level PPV for the motifs as a function of their p-value. We consider a prediction to covers a site if there is at least a three base pair overlap.

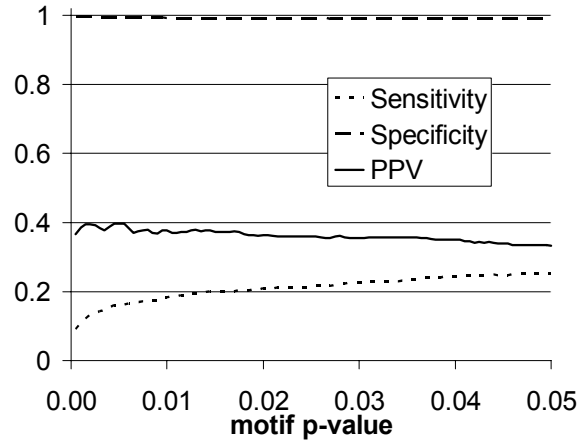

Figure 1. Method's performance against a set of known sites

Although, the PPV for our method is high, approaching 0.4 at a more stringent p-value cutoff, the sensitivity remain low. With the prediction pipeline in place, we plan to refine our input sets, background model, parameter setting and method used against the KSL to improve sensitivity and the PPV.

4. References

[1] M. Tompa, et al., "Assessing computational tools for the discovery of transcription factor binding sites," *Nat Biotechnol*, vol. 23, pp. 137-44, 2005.

[2] E. Birney, et al., "An overview of Ensembl," *Genome Res*, vol. 14, pp. 925-8, 2004.

[3] M. Kanehisa, et al., "The KEGG resource for deciphering the genome," *Nucleic Acids Res*, vol. 32, pp. D277-80, 2004.

[4] O. L. Griffith, et al., "Assessment and Integration of Publicly Available SAGE, cDNA Microarray, and Oligonucleotide Microarray Expression Data for Global Coexpression Analyses," *Genomics*, accepted.

[5] T. L. Bailey and C. Elkan, "The value of prior knowledge in discovering motifs with MEME," *Proc Int Conf Intell Syst Mol Biol*, vol. 3, pp. 21-9, 1995.

[6] G. Z. Hertz and G. D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences," *Bioinformatics*, vol. 15, pp. 563-77, 1999.

[7] G. Thijs, et al., "A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling," *Bioinformatics*, vol. 17, pp. 1113-22, 2001.