TFBSs prediction by integration of genomic, evolutionary, and gene expression data

Ambesi-Impiombato A (1,2), Bansal M (1,3), Rispoli R (1), Liò P (4), di Bernardo D (1,3)

(1) Telethon Institute of Genetics and Medicine, Tigem, Napoli
(2) Department of Neuroscience, University of Naples "Federico II", Napoli
(3) SEMM, European School of Molecular Medicine, Naples, Italy
(4) Computer Laboratory, Cambridge University, Cambridge, UK

Motivation

Control of gene expression is essential to the establishment and maintenance of all cell types, and is involved in pathogenesis of several diseases. However, biological mechanisms underlying the regulation of gene expression are not completely understood, and predictions via bioinformatics tools are typically poorly specific. We have developed and tested a computatonal workflow to computationally predict Transcription Factor Binding Sites on proximal promoters of vertebrate genes. Finally we applied the workflow to a cluster of genes found to respond significantly to p63 overexpression.This dataset consists of microarray gene expression at 15 time-points in primary murine keratinocytes.

Methods

Our approach for the prediction of regulatory elements is based on a search for known regulatory motifs retrieved from TRANSFAC, on DNA sequences of genes' promoters. Genomic information is retrieved from ensembl database (www.ensembl.org) and compara for orthology information. Predictions are computed independently on different species and the final scores are integrated using a weighted sum calibrated on the phylogenetic distances between the species. These predictions are further refined using logistic regression to integrate data from co-regulated genes. For the purpose of this analysis each matrices were scored using a 3rd order Markov Model trained on a large number of intergenic regions upstream of randomly selected genes.

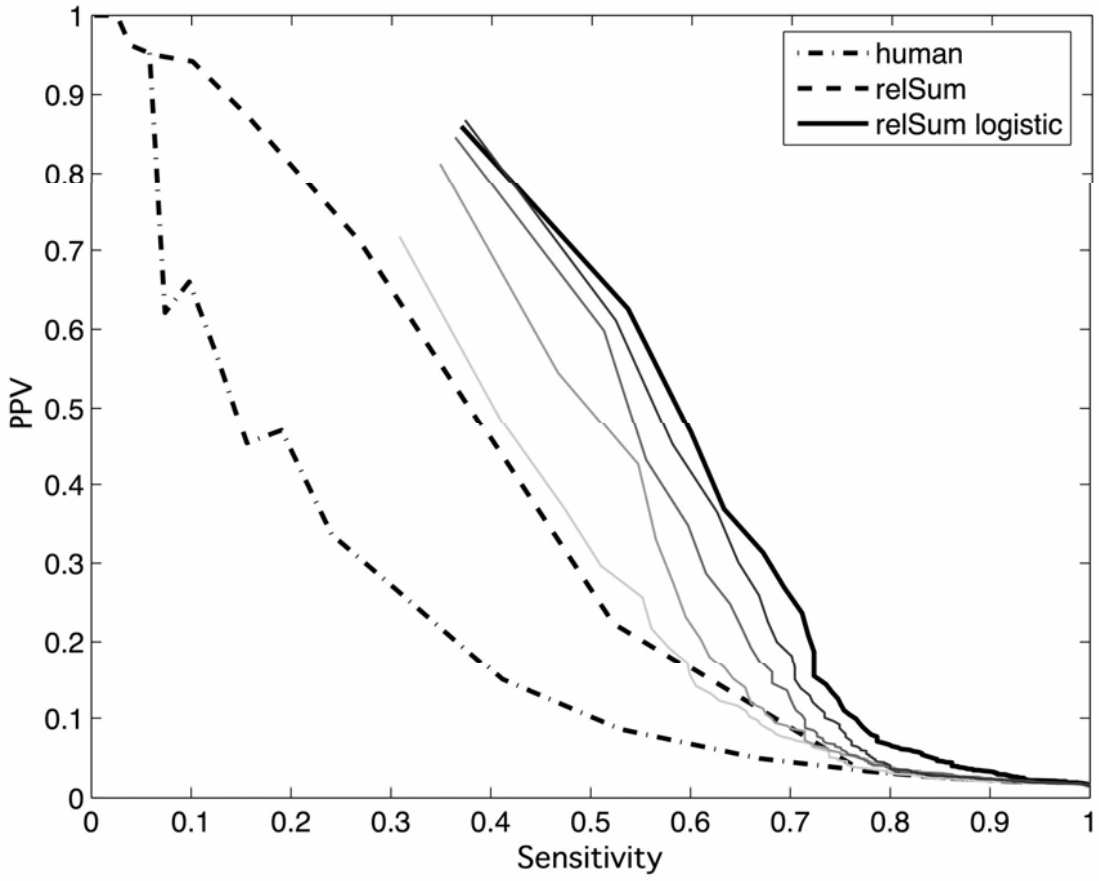
Results

We show the advantages of integrating genomic data with information based on evolutionary conservation, as well as gene expression data. Consistent results were obtained on a large simulated dataset consisting of 13050 simulated promoter sequences (performance shown in figure 1), on a set of 161 human gene promoters for which binding sites are known. Key factors of our approach include the integration of predictive scores obtained on promoters of ortholog genes from multiple species, and the possibility to include a priori information such as that available from quantitative or qualitative gene expression data, by fitting a logistic regression. A robustness of the logistic regression was evalutated by progressively misassigning genes to the co-regulated group. Our results on simulated datasets show that integrating information from multiple data sources, such as genomic sequence of genes' promoters, conservation over multiple species, and gene expression data, indeed improves the accuracy of computational predictions.

Contact email: ambesi@tigem.it

References

- Tadesse MG, Vannucci M, Lio P: Identification of DNA regulatory motifs using Bayesian variable selection. Bioinformatics 2004, 20:2553-2561.
- Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J: Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 2006, 124:47-59.
- Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: Reverse engineering of regulatory networks in human B cells. Nat Genet 2005, 37:382-390.


human
relSum
relSum logistic
PPV
Sensitivity
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1