Identification of Regulatory Elements from Gene Expression Data without Clustering

Mathieu Lajoie, Olivier Gascuel, Laurent Brehelin

Identification of Regulatory Elements from Gene Expression Data without Clustering

Mathieu LAJOIE[1], Olivier GASCUEL[1] and Laurent BRÉHÉLIN[1]

Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS, Université Montpellier 2, France
{lajoie,gascuel,brehelin}@lirmm.fr

Keywords Gene regulation, motif discovery, transcription factor, k-nearest neighbors.

1 Background

In silico discovery of regulatory motifs (RM) can be seen as a feature extraction problem. Given a set of nucleic acid sequences that are mapped to some expression data, the goal is to find a concise set of motifs that is the most informative with regards to that mapping. Apart from few exceptions (e.g. [1]), one of the most common approaches is to use a clustering algorithm to partition the expression dataset, and to apply on each cluster one of the numerous algorithms that have been designed to find over-represented motifs in a predefined set of sequences, such as MEME [2] or AlignACE [3]. However, the partition induced by this clustering rarely corresponds to a biological reality. Firstly, expression data is inherently noisy, and determining the "real" number of clusters is considered to be one of the most difficult problems in classification. Secondly, different RM usually have overlapping gene sets, which cannot be appropriately modeled by a single partition. In addition, these algorithms rely on statistical models of sequence background, which have been reported to produce many false positives [4,5], especially with repeat-rich and atypical genomes. This is the case, for example, with *Plasmodium falciparum*, whose A+T content almost reaches 90% in the intergenic regions.

The FIRE [4] and GEMS [5] algorithms have been designed for finding RM from whole genomes and high dimensional datasets without models of sequence background. However, they both rely on a clustering of the expression data and are subject to the aforementioned criticisms. The two methods differ in the way the dependency between the presence of a motif and the expression profile of the corresponding gene is measured. GEMS uses the hypergeometric distribution to measure motif enrichment in each co-expression cluster, while FIRE computes the mutual information between the presence/absence of a motif and the cluster membership of the corresponding genes. These two approaches can be seen as two extremes of a simple model, which only assumes that RM must show some kind of statistical dependency with the expression data. The hypergeometric approach is a local criterion, as it considers motif enrichment in a single cluster at a time, while the mutual information approach is a global one including the contributions of all the clusters.

2 Method and Results

In this work, we show that the hypergeometric distribution and the mutual information criteria can be used without requiring any clustering, using the notion of *motif density* in expression space. Namely, rather than considering the number of genes that contain a motif in each cluster, we compute motif densities locally around each gene with a k-nearest neighbors approach. For the hypergeometric criterion, the score of a motif is then defined as the negative logarithm of the lowest p-value observed among all these neigborhoods. For the mutual information, the score is obtained by summing over the density estimate of each gene, instead of each cluster.

We compared the original and new version of both criteria on two *S. cerevisiae* and three *P. falciparum* gene expression datasets. All possible 8-mers were enumerated and scored with the four objective functions, and a false discovery rate (FDR) was estimated for different score thresholds using a shuffling procedure. For the original criteria, we used the k-means algorithm with different number of clusters ($3-10, 20, 30, 40$), and kept the clustering that yields the best sensitivity at 0.1 FDR. Fig. 1 shows the number of 8-mers identified by the new

and original criteria under various FDR for two datasets. We see that avoiding the clustering step results in a significant increase of the sensitivity over the original methods. Overall, we observed a significant improvement on the five datasets for the hypergeometric criteria, and on four datasets for the mutual information.

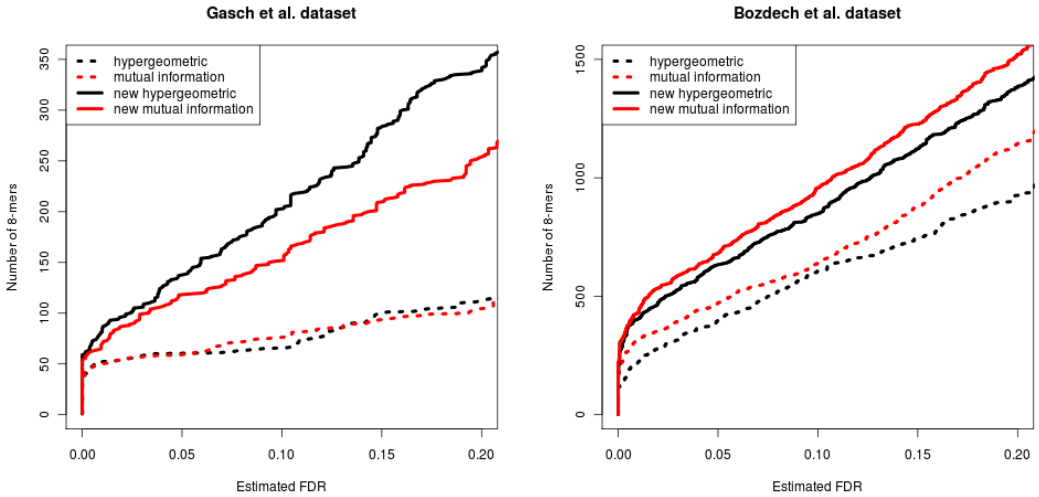

Figure 1. Number of predicted 8-mers, according to different estimated FDR thresholds, for (left) the yeast dataset [6] and (right) the *P. falciparum* dataset [7]. The best results of the original scoring functions (hypergeometric and mutual information) are achieved with 9 and 7 clusters (respectively) in yeast, and 3 and 7 clusters (respectively) in *P. falciparum*.

Using yeast Protein Binding Microarray (PBM) datasets [8], we further show that our continuous approaches also improve prediction. A True Positive Rate (TPR) that measures the fraction of predicted 8-mers bound by a transcription factor in the PBM experiments was computed for the four approaches. As we observed with the estimated FDRs, the new criteria outperform the original ones in this experiment. For the Gasch dataset presented in Fig. 1, the TPR is 55% for the 200 highest scoring 8-mers returned by the original criteria, whereas it reaches 65% for the new versions. Finally, we showed that using motif densities presents several advantages compared to the clustering approach. In addition to the increased sensitivity, it provides a simple way of comparing different motifs and predicting the functionality of individual motifs occurrences. All these methods have been implemented in a software called RED^2, for *Regulatory Elements Discovery from Raw Expression Data*. Motifs are represented as IUPAC strings of arbitrary lenght, allowing an easy and comprehensive analysis of a wide range of expression data.

References

[1] H.J. Bussemaker, H. Li and E.D. Siggia, Regulatory element detection using correlation with expression. *Nature genetics*, 27(2):167–174, 2001.

[2] T.L. Bailey and C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *ISMB. International Conference on Intelligent Systems for Molecular Biology*, volume 2, page 28, 1994.

[3] J.D. Hughes, P.W. Estep, S. Tavazoie, and G.M. Church, Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of molecular biology*, 296(5):1205–1214, 2000.

[4] O. Elemento, N. Slonim, and S. Tavazoie, A universal framework for regulatory element discovery across all genomes and data types. *Molecular cell*, 28(2):337–350, 2007.

[5] J.A. Young, J.R. Johnson, C. Benner, S.F. Yan, K. Chen, K.G. Le Roch, Y. Zhou, and E.A. Winzeler, In silico discovery of transcription regulatory elements in Plasmodium falciparum. *BMC Genomics*, 9(1):70, 2008.

[6] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein, and P.O. Brown, Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11(12):4241, 2000.

[7] Z. Bozdech, M. Llinás, B.L. Pulliam, E.D. Wong, J. Zhu, and J.L. DeRisi, The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol*, 1(1):E5, 2003.

[8] G. Badis, E.T. Chan, H. van Bakel, L. Pena-Castillo, D. Tillo, K. Tsui, C.D. Carlson, A.J. Gossett, M.J. Hasinoff, C.L. Warren, et al, A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Molecular cell*, 32(6):878–887, 2008.