FINDING TRANSCRIPTION FACTOR BINDING SITES IN COREGULATED GENES BY EXHAUSTIVE SEQUENCE SEARCH

Kielbasa Sz.M., Korbel J.O., Beule D., Schuchhardt J., and *Herzel H.

Institute for Theoretical Biology, Humboldt University, Berlin, Germany

E-mail: h.herzel@itb.biologie.hu-berlin.de

*Corresponding author

Keywords: regulatory sequences, upstream regions, exhaustive search, Ras pathway

Resume

Motivation:

Growing amounts of gene expression data provide the possibility of finding coregulated genes by clustering methods. By analysis of the promoter regions of these genes, rather weak signals of transcription factor binding sites may be detected [Zhang, 1999]. We compare existing programs and own software on yeast clusters. Therefore, we introduce the new algorithm ITB, an integrated tool for box finding, which exhaustively analyses regular expression-like patterns in promoter sequences, allowing gaps and the matching of more than one base at any position within the candidates. The applicability of ITB to predict transcription factor binding sites in human promoter sequences is evaluated.

Results:

Three publicly available algorithms were compared to our program, particularly on yeast clusters. Moreover, ITB was tested on promoter sequences of coregulated human genes. ITB is capable of predicting several verified transcription factor binding sites in yeast.

Availability:

The program ITB is available upon request.

Introduction

Basic molecular biological processes are regulated by the specific interaction of proteins and short DNA sequences. Two different approaches are used to predict transcription factor binding sites: Exhaustive analyses of oligonucleotide frequencies, for instance, as described by van Helden et al. [van Helden, J. et al., 1998] and non-exhaustive optimization approaches using weight matrices, like Gibbs sampling, which was first described by Lawrence et al. [Lawrence, C.E. et al., 1993]. The program RSA-tools-oligo-analysis [van Helden, J. et al., 1998] compares the frequency of conserved words in a given set of promoter sequences to the frequency of those words in a training set. This method is sensitive in detecting conserved words, which are slightly over-represented in the given coregulated sequence set. Unfortunately, regulatory elements not having a conserved core sequence cannot be detected by this method. Weight matrix based methods like Gibbs sampling [Lawrence, C.E., et al., 1993] or MEME [Bailey, T.L. & Elkan, C., 1994] can predict elements without a conserved core. However, for a small number of sequences in the coregulated set, weight matrices are of limited use. Moreover, signals provided by DNA regulatory elements involved in transcription are weak and rather poly-(A), (T), or GC-rich regions of the promoter might be aligned by weight matrix-based prediction methods.

Methods and algorithms

Three publicly available tools were compared to our program ITB: The Gibbs Motif Sampler (developed by E.C. Rouchka & B. Thomson based on the work of Lawrence [Lawrence, C.E., et al., 1993], RSA-tools-oligo-analysis [van Helden, J. et al., 1998], and MEME version 2.2 [Bailey, T.L. & Elkan, C. 1994]. RSA-tools-oligo-analysis cannot be run with human DNA, since the scoring method of that program is dependent on the training set and human training sets cannot be chosen in the web interface of the program. Instead, we apply ITB, a program developed by our group, which is similar to RSA-tools-oligo-analysis when run in the "ACGT"-mode. For the analysis of Saccharomyces cerevisiae, ITB is trained with a collection of all 5'-UTRs of yeast of length 800, while all 271 human promoter sequences obtained from EPD, the eukaryotic promoter database [Perier, R.C. et al., 1998], are used as a training set for the analysis of human DNA.

Our program written in C++ compares frequencies of conserved elements in the given promoter set to the expected frequencies of these elements, which are estimated based on the training set by using Markov chain models of varying orders. The program performs an exhaustive search – scores are calculated for all possible
6-mers built from the alphabets 'ACGT' or 'ACGTWRKSYMN' (The meaning of these symbols used in the "extended mode" of the program are: W=A or T, R=A or G, K=G or T, S=C or G, Y=C or T, M=A or C, and N=any of ACGT.), depending on the mode requested. The score expresses, in the logarithmic scale, how improbable it is for the background model (with "all motifs equally distributed") to generate the observed number of occurrences of a motif in the coregulated set (formulae from [van Helden, J. et al., 1998]), with modifications for the extended alphabet. Finally, a list of the motifs with the best scores is created, containing the most over-represented patterns in comparison to the training set. The last step, which we currently achieve "by eye", is the removal of self-overlapping patterns (like AAARAA or ATAWAT) from the list.

We compare existing methods and own software on yeast clusters and evaluate the applicability of our algorithm for human promoter sequences. Zuber J. reported the coregulation of several genes of the H-Ras signal transduction pathway [Zuber, J., et al., 2000]. They kindly provided five promoter regions of those genes. Moreover, four promoter sequences of genes upregulated by Myc [Coller, H.A. et al., 2000] were extracted from EPD [Perier R.C. et al., 1998].

Implementation and results

Coregulated gene sets of yeast used in our analysis are identical with sets used by van van Helden, J. [van Helden, J. et al., 1998]. Table 1 lists human gene sets used in the analysis. Table 2 shows the results of a preliminary analysis (without adjusting/optimizing parameters) of the performance of MEME, Gibbs sampler, RSA-tools-oligo-analysis, and ITB in upstream regions of coregulated yeast genes ("gene families"). ITB predicted most previously characterized elements correctly (top scoring element). For the GCN family, the previously characterized element was ranked at position two. MEME and the Gibbs sampler failed to predict some of the previously characterized elements. The programs were compared in a single run. Table 3 shows the type of motifs predicted by ITB run in the "extended-mode" and their rank based on the score.

Table 1. The genes used in analyses of human promoter regions.

| Human protein | Genes regulated via the human protein |
|---------------|--------------------------------------|
| Myc           | EP11114, EP15041, EP36018, EP37001    |
| H-Ras         | LOX, LOXL1, LOXL2, TSP1               |

Table 2. Exhaustive methods and non-exhaustive, optimizing approaches are used to predict hexanucleotides, as parts of previously characterized regulatory elements of yeast. An ‘x’ indicates that the previously characterized element is correctly predicted (ranked on the top position) by the program.

| Gene family | MEME | Gibbs sampler | RSA-tools-oligo-analysis | ITB ("ACGT-mode") | ITB ("extended-mode") |
|-------------|------|---------------|--------------------------|--------------------|------------------------|
| NIT         | –    | –             | –                        | x                  | x                      |
| MET         | –    | –             | –                        | x                  | x                      |
| PHO         | x    | –             | x                        | x                  | x                      |
| PDR         | x    | –             | –                        | x                  | x                      |
| GCN         | x    | –             | –                        | x                  | –                      |

Table 3. Highly ranked hexanucleotides as computed by the program ITB are similar to the consensus sequences of previously characterized sites taken from TRANSFAC [Wingender et al., 1996] or from publications by Katzmann D.J. [Katzmann, D.J. et al., 1996], and Kuras L. [Kuras, L. et al. 1996].

| Family | Previously characterized element | Element predicted by ITB | Inf. cont. | Score  | Rank |
|--------|----------------------------------|--------------------------|------------|--------|------|
| NIT    | GATAAG¹                         | KATMRS                   | 8.1        | 25.6   | 1    |
| MET    | TCACGTG¹                        | CMCRYR                   | 9.0        | 29.2   | 1    |
| PHO    | CACGTKNG¹                       | AYKWGS                   | 8.4        | 25.0   | 1    |
| PDR    | TTCGCCGGAA²                      | CCRYGG                   | 12.3       | 33.7   | 1    |
| GCN    | RTGACCTCATNS²                    | AYGACK                   | 11.6       | 15.1   | 2    |

Coregulated via

| Myc         | CACGTG¹ (binding core of Myc) | MMCKKG                  | 25          |
| Ras         | not characterized yet         | SYSTST                  | 1           |

Discussion

Although the first results in yeast are promising, in order to predict motifs in human gene sets we are extending the program mentioned above. Currently we develop an algorithm, which is aware of self-overlapping words and allows an exhaustive search for patterns containing gaps. Moreover, we need more human promoters of coregulated genes as well as better training sets. Since only a minority of human promoters have been experimentally verified, promoter prediction algorithms could be used (reviewed by Fickett & Hatzigeorgiou, 1997) to obtain more sequences.
References

1. Bailey, T.L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California.

2. Coller, H.A., Grandori, C., Tamayo, P., Colbert, T., Lander, E.S., Eisenman, R.N., and Golub, T.R. (2000). Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. Proc. Natl. Acad. Sci. USA, 97, 3260-3265.

3. Fickett, J.W., and Hatzigeorgiou, A.G. (1997). Eukaryotic promoter recognition. Genome Res., 9, 861-878.

4. Katzmann, D.J., Hallstrom, T.C., Mahe, Y., and Moye-Rowley, W.S. (1996). Multiple Pdr1p/Pdr3p Binding Sites Are Essential for Normal Expression of the ATP Binding Cassette Transporter Protein-encoding Gene. J. Biol. Chem., 271, 23049-23054.

5. Kuras, L., Cherest, H., Surdin-Kerjan, Y., and Thomas, D. (1996). A heteromeric complex containing the centromere binding factor 1 and two basic leucine zipper factors, Met4 and Met28, mediates the transcription activation of yeast sulfur metabolism. EMBO J., 15, 2519-2529.

6. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wooton, J.C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science, 262, 208-214.

7. Perier, R.C., Junier, T., and Bucher, P. (1998). The eukaryotic promoter database EPD. Nucleic Acids Res., 26, 353-357.

8. van Helden, J., Andre, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J. Mol. Biol., 281, 827-42.

9. Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res., 24, 238-241.

10. Zhang, M.Q. (1999). Large-scale gene expression data analysis: a new challenge to computational biologists. Genome Res., 9, 681-688.

11. Zuber, J., Tchernitsa, O.I., Hinzmann, B., Schmitz, A.C., Grips, M., Hellriegel, M., Sers, C., Rosenthal, A., and Schafer, R. (2000). A genome-wide survey of RAS transformation targets. Nat. Genet., 24, 144-52.