INCLUSive: INtegrated Clustering, Upstream sequence retrieval and motif Sampling

Gert Thijs[1], Yves Moreau[1] , Frank De Smet[1], Janick Mathys[1], Magali Lescot[1], Stephane Rombauts[2], Pierre Rouze[3], Bart De Moor[1], Kathleen Marchal*[1]

[1] ESAT-SISTA/COSIC, KULeuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium

Email: {gert.thijs,kathleen.marchal}@esat.kuleuven.ac.be

URL: http://www.esat.kuleuven.ac.be/sista

Tel: +32/16321884

Fax: +32/16321970

* corresponding author

[2] Department of Plant Genetics, VIB, UGent, Ledeganckstraat 35, 9000 Gent, Belgium

[3] INRA associated laboratory, VIB, UGent, Ledeganckstraat 35, 9000 Gent, Belgium

keywords: clustering, motif finding, upstream sequence retrieval

Abstract

Summary:

INCLUSive allows automatic multistep analysis of microarray data (clustering and motif finding). The input consists of a data matrix containing the identification tags and the expression levels of the genes in the different profiling experiments. The clustering algorithm (adaptive quality-based clustering) groups together genes with highly similar expression profiles. The upstream sequences of the genes belonging to a cluster are automatically retrieved from GenBank and can be fed directly into Motif Sampler, a Gibbs sampling algorithm that retrieves statistically over-represented motifs in sets of sequences, in this case upstream regions of co-expressed genes.

Availability: Our tool is freely accessible for academic purposes on the Web at http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html

Contact: Gert.Thijs@esat.kuleuven.ac.be

Microarray experiments measure the global transcriptional behavior of the organism elicited by some signal. In such experimental setup, coregulated genes are likely to exhibit a similar expression profile. Clustering identifies such groups of co-expressed genes. Subsequent detection of a regulatory motif in the upstream regions of cluster members reflects a common mechanism of transcriptional regulation and therefore helps in discerning noisy and biologically relevant clusters.

However, motif finding algorithms are sensitive to noise (presence of sequences not containing the motif). Therefore multistep analysis requires the creation of clusters containing tightly co-expressed genes and a robust motif detection algorithm able to cope with noise.

Although many clustering and motif detection algorithms exist, the application described here offers several advantages. Firstly, it combines two algorithms that were specially designed for this multistep analysis. Secondly, it includes automatic retrieval of intergenic sequences. This automation relieves the user from the time-consuming task of manually selecting intergenic

regions. However, maximal interaction with the user is guaranteed. A schematic representation and detailed information on the interface is available at the web site..

Clustering

The high-throughput data can be passed to the algorithm via a web form (adequate data format is described on the web site). Adaptive quality-based clustering is used to generate groups of tightly co-expressed genes (De Smet et al., 2001).

The clustering algorithm necessitates two user-specified parameters: (1) quality criterium *s* (default value is 0.95) guarantees that the probability of a gene belonging to a cluster is at least *s*; (2) minimal number of genes in a cluster (default value is 3) merely serves presentational purposes and has no large effect on the core of the algorithm. By visually inspecting the cluster results, the user can decide whether other parameter settings are necessary. .

The URL of the cluster results page is reported by email. This page shows for each of the clusters retrieved by the algorithm: (1) a plot of the expression profiles of the cluster members, (2) the average expression profile of that cluster and (3) a list of the tags identifying the cluster members (accession number and gene name). Based on these tags the upstream regions of the corresponding genes can be retrieved automatically.

Selection of the upstream region

For each gene in a cluster the GenBank entry corresponding to the accession number is retrieved,parsed and indexed. The corresponding query gene name is matched to the indexed genes. Detailed warning messages inform the user if processing errors occur. If detected within the entry, the query gene is listed as successfully identified. If not detected, the user should manually identify an indexed gene that matches the query..

Of the correctly identified genes the upstream DNA sequence is selected. The desired minimal length of the upstream region (default 300bp) is user defined. The search system preferentially selects intergenic regions (defined as the non-coding region between two

adjacent genes) based on one of the following annotations in the specified order of preference (1) primary transcript, (2) mRNA, (3) CDS, (4) gene. For a more detailed description of the definition of intergenic regions we refer to the web site. Depending on the availability of an intergenic region in the GenBank entry one of the following situations can be encountered:

1. When the GenBank entry is accurately annotated and the gene upstream of the gene of interest is present in the same entry, the intergenic region is added to the list of selected intergenic regions.
2. If the gene of interest is either the first (w-strand) or the last gene (c-strand) in the entry, the entry only contains an upstream region and no well-delineated intergenic region. However, if this upstream region is larger than the minimal user-defined length this upstream region is added to "the long upstream list". However, such gene will also be added to the *BLAST list* to search for a genomic entry containing the delineated intergenic region. For some organisms (e.g. prokaryotes) the intergenic region might be so small that an upstream region already comprises part of a yet unannotated coding region. A more clearcut delineation of the intergenic might therefore be useful.
3. If the retrieved upstream region is smaller than the desired minimal length, the gene is added to the *BLAST list*.

All query genes from the *BLAST list* are blasted at NCBI to retrieve more completely annotated GenBank entries containing an intergenic region.

Significant hits are extracted from the *BLAST* reports and the GenBank entries corresponding to these hits are parsed to find the gene overlapping with the query gene. If detected, the selected intergenic or upstream region is added to the corresponding list. Note that because of the procedure followed, the same gene can appear more than once in the final report. A well-documented intermediate report allows the user to define sequences desired for further analysis.

For the implementation of the automatic retrieval system, cgi-scripting and BioPerl modules (http://bio.perl.org/) are used. Note that this intergenic retrieval system can also be started from an independent form where the accession numbers and gene names can be entered.

The Motif Sampler

The Motif Sampler used in our integrated tool is an adapted version of the original Gibbs sampling algorithm (Lawrence et al., 1993) and has been described more extensively elsewhere (Thijs et al., 2001a,b). Summarizing, the Motif Sampler is a user-friendly implementation that allows detecting statistically over-represented motifs in a set of unaligned sequences. The algorithm determines in which sequences and at what positions a statistically over-represented motif is present in a given data set. The following parameters are user-specified: the motif length (default value = 8), the maximal expected number of occurrences of a given motif (default value = 1), the number of different motifs (default value = 6) and the allowed motif overlap (default value = 2).

As previously described (Thijs et al., 2001b), an appropriate organism dependent background model can considerably enhance the outcome of the Motif Sampler. If no higher-order model is available, the background model is derived from the input data.

The output of the Motif Sampler consisting, for each motif, of a position probability matrix, a motif logo, scores, and visualisation of the motif occurrences along the sequence is reported by email.

Conclusions

This web interface aims at integrating tools for the analysis of microarray data. The output of the clustering analysis will automatically be transformed into an input compatible with the Motif Sampler. Note that each of the described applications can also be used independently from each other.

Of course the current implementations are only the seed for a more complex integrated system. In this implementation the definition of the intergenic region relies on the annotation of the downloaded sequences. Mostly only translation start and stop are annotated such that the defined intergenic region encompasses the region between translation start and stop. This might pose severe drawbacks for non-compact genomes. Indeed, due to the sensitivity of the motif finding algorithms it is of outermost importance to delineate the region of interest as accurately as possible.

Acknowledgments

Gert Thijs is research assistant of the IWT; Kathleen Marchal and Yves moreau are post-doctoral researchers of the FWO; Prof. Bart De Moor is professor at the KULeuven. Pierre Rouze is Research Director of INRA, France. This work is partially supported by: 1. IWT project: STWW-980396; 2. Research Council KULeuven: GOA Mefisto-666; 3. FWO projects: G.0115.01; 4. IUAP P4-02. We like to thank Sigrid De Keersmaecker for helpful feedback on the prototypical system.

References

Altschul,S.F., Thomas,L., Madden,A., Schaffer,A., Zhang,J., Zhang,Z., Miller,W., and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25,** 3389-3402.

De Smet,F., Marchal,K., Mathys,J., Thijs,G., De Moor,B. and Moreau,Y. (2001) Adaptive quality-based clustering of gene expression profiles. submitted.

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

Thijs,G., Marchal,K., Lescot,M., Rombauts,S., De Moor,B., Rouzé,P. and Moreau,Y. (2001a) A Gibbs Sampling method to detect overrepresented motifs in upstream regions of co-expressed genes. In Proc. *Recomb'2001*, **5**, 305-312.

Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouzé,P. and Moreau,Y. (2001b) A higher-order background model improves the detection of promoter regulatory elements by Gibbs Sampling. *Bioinformatics*. In press.

Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L., and Rapp,B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **29**, 11-16.