A TOOL FOR REVEALING COMPOSITE MODULES, REGULATING GENE EXPRESSION ACCORDING TO MICRO ARRAY DATA

[1]Konovalova T.G, [1]Ivanov V., [1]Valeev T., [1]Cheremushkin E., [1]Beschastnov E., [1]Komashko V., [1]Lobanova M., [1,2]Kel A.

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk

[2] BIOBAS, GmbHWolfenbuettel

A novel computational method based on artificial intelligence algorithms was developed to study composite structure of promoters of co-expressed genes. Our method enabled the identification of combinations of multiple transcription factor binding sites regulating the concerted expression of genes. Now we have console application implementing functionality described above. We started development of graphical users interface for wide biologist usage.

Introduction

Functionally related genes involved in the same molecular-genetic, biochemical, or physiological process are often regulated coordinately by specific combinations of transcription factors. Such regulation is provided by precisely organized binding of a multiplicity of transcription factors (TFs) to their target sites (cis-elements) in regulatory regions. Cis-element combinations provide a structural basis for the generation of unique patterns of gene expression. Specific combinations of cis-elements (composite modules) for the vast variety of gene functional classes have to be determined to achieve goals of functional genomics and gene expression analysis[1]. Mass data on gene expression coming from the microarray experiments provide valuable information to deduce gene regulatory mechanisms. Groups of co-expressed genes can be revealed from these data using various clustering techniques. We can search for potential TF binding sites using a collection of known weight matrices. The most promising are techniques searching for specific combinations of TF binding sites that correlate with gene expression patterns. We have developed a new method for revealing class-specific composite modules in promoters of functionally related or coexpressed genes.

Composite Modules model

We define a composite module CM as a set of TF weight matrices with given matrix cut-offs and other parameters which is associated with a specific functional type of gene regulatory regions. CMs are characterized by the following parameters: K, the number of PWMs in the module (typically 3 to 10), cut-off values $q_{cut-off}^{(k)}$, relative importance values $\phi^{(k)}$ and maximal number of best matches $\kappa^{(k)}$ that are assigned to every weight matrix k (k=1,K) in the CM. These K matrices are selected by the program from a library of all considered matrices. We use different profiles including the profile vertebrate_minFN62.prf, which includes 410 matrices for different transcription factors of vertebrate organisms (TRANSFAC® rel. 6.4)[2].

When all these parameter settings can be defined, a “composite module score” (CM score) is given for any sequence X using the following equation:

$$F_{CM}(X) = \sum_{k=1,K} \phi^{(k)} \times \sum_{i=1}^{\kappa^{(k)}} q_i^{(k)}(X)$$

where $q_i^{(k)}(X)$ are the $\kappa^{(k)}$ best scoring sites found in the sequence X by the matrix (k).

The algorithm takes as an input two sets of sequences (the set which is analyzed and a background set) and a set of weight matrices for transcription factors. In all sequences from these sets potential sites a searched, using weights matrixes. For each sequence X function F (X)

(reflecting frequency of occurrence of the certain combinations of sites in sequences) is calculated and distributions of F(X) is built for each sets

Classification of any new regulatory sequences to one of two types (+) or (-) is carried out by calculation constructed functional for this sequence.

The further the received distributions are one from another, the further identification will be more effective.

Another approach use the whole set of promoters with different levels of expression. In this case distribution of F(x) must be as much correlated with expression pattern as possible.

Search of such optimum combination is conducted with the help of metropolis [3] or genetic algorithms.

Algorithms

At the first stage of the metropolis algorithm one complex is randomly created. It is set S from K matrices m1…mK with cut-offs c1…cK appropriate to them.

For each promoter X and matrix mi from our complex we calculate

$$F(x) = \sum_{m_i \in S} w_{ij}$$

where wij is weight of a j site for matrix mi. Further dividing function R is calculated: R=(F+ - F-)/(δ++δ-) where F+ and F- -average, and both δ+ and δ- - dispersions of distributions of F values on samples Q+ and Q-accordingly.

On each following step i complex S(i-1) undergoes random mutation by change of a site or increase / reduction of its cut-off value, so that we have new complex S(i), and value R(i) is calculated. For two values R(i-1) and R(i) the following condition is checked :

if R (i) > R (i-1) we accepts S (i) otherwise we accept S (i) with probability r (i)/r (i-1) or S (i-1) with probability 1-R (i)/R (i-1). If we decide to accept S(i-1) that means that S(i)= S(i-1). Information about S(i) is stored for statistics and we go to the next step.

Thus complexes that most frequently occur in our distribution will divide our sets in most accurate way.

For the analysis of distribution we used several types of statistics.

The first variant takes into account single matrixes with a cut-off got in the certain interval ri . Second type records whole complex matrixes with corresponding intervals off cut-offs.

We have also applied a genetic algorithm for revealing a best complex. In that case for given S and k we generate a random ‘pool’ of ‘genomes’ G. Each genome is a set of matrixes with all appropriate parameters. These genomes in the pool are subject to mutation and selection. Selected genomes (best according to the dividing function R) are duplicated several times until the initial number of genomes in the pool is reached. After numerous rounds of such selection we retrieve the best genome Sbest

Method to determine CMs correlated with gene expression.

The method is based on a modification of the algorithms described in the previous sections. A program searches for a combination of TF matrices that minimizes a goal function G that is a weighted sum of the least square deviation between logarithms of gene expression values and the score function FCM described in the previous section and a rank correlation coefficient

$$\alpha \times \sum_{genes} (LR - F_{CM})^2 + \beta \times \rho_{rank}(LR, F_{CM}) \rightarrow \min$$

here $LR = \log_2(ex1/ex2)$ which is the log-ratio of the expression values of a particular gene in the experiment versus control, e.g. a signal log-ratio.

Testing.

In the beginning, using artificial sets of data, we have estimated quantity of iterations needed for successful division of sets. We used three sets of functionally related genes - brain-, muscle-and liver-specific. On a first step after 120000 iterations 7 best matrixes were selected. Then we select complex of 3 factors with their cut-offs got in the certain intervals, appeared the most probable after 6.000.000 iterations. Results are shown in the following table.

Liver-specific	Muscle-specific	Brain-specific
V$HNF1_01 [0.8693 , 0.9347]	V$HEB_Q6 [0.9897 , 1.0000] ,	V$MAZ_Q6 [0.9420 , 0.9710]
V$AP3_Q6 [0.9500 , 1.0000]	V$CDX2_Q5 [0.8900 , 0.9450],	V$TEL2_Q6 [0.9687 , 0.9843]
V$YY1_01 [0.9697 , 1.0000]	V$SRF_C [0.8740 , 0.9160]	V$CDX2_Q5 [0.9450 , 1.0000]

Here we can see that algorithm have selected matrixes for factors that are known to play important role in corresponding genes group regulation, such as HNF1 for liver or SRF for muscle differentiation.

Further work

Now we have console application implementing functionality described above. We started development of graphical users interface for wide biologist usage.

Literature

1. Kel-Margoulis OV, Ivanova TG, Wingender E, Kel AE. Automatic annotation of genomic regulatory sequences by searching for composite clusters.Pac Symp Biocomput. 2002;:187-98. PMID: 11928475
2. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhäuser, R., Prüß, M., Schacherer, F., Thiele, S. and Urbach, S. «The TRANSFAC system on gene expression regulation. » *Nucleic Acids Res.* 29, 281-283
3. Aris-Brosou S. How Bayes tests of molecular phylogenies compare with frequentist approaches. Bioinformatics. 2003 Mar 22; 19(5):618-24.