

Regulatory mechanisms in cell, Searching regulatory factors for genes

Presentation by Margus Lukk
Notes by Reetta Nylund

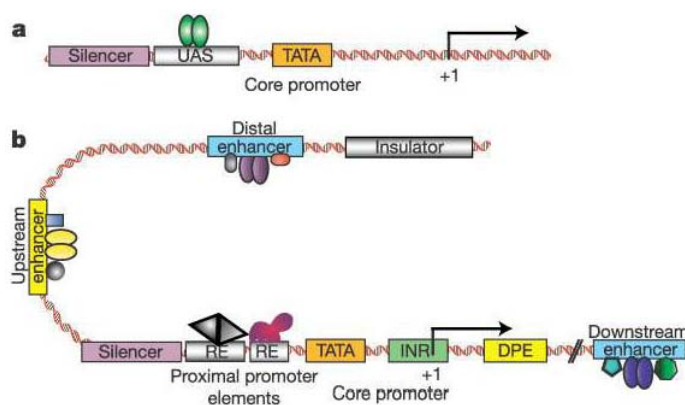
Lecture on Friday 22th April, 2005

Within past years the genetic sequence of several biological organisms has been revealed. The first eukaryotic cell, whose complete genome was sequenced, was yeast *S. cerevisiae* in the year 1996. The sequence consists of 16 chromosomes, which include more than 13 million bases of DNA and by far about 6300 open reading frames (ORF) possibly encoding for a gene have been identified. In 2000's also the genetic sequence of human has been identified and currently it is estimated to contain about 30000 genes. Thus, the number of genes among the organisms has been turned to be in the same level although the complexity of the organisms differs greatly. Consequently, the complexity of the systems can be explained by regulation.

Regulation of Transcription

Transcription is the process in which a one strand of a DNA molecule (i.e. typically a gene) is used as a template for a synthesis of complementary RNA. This RNA then goes through several processes (e.g. splicing) after which it is used as a template for translation process, which forms a polypeptide (i.e. typically a protein).

The transcription process is regulated on several ways. One part of the regulation is the *cis*-acting elements, which are located on the same strand of the gene (i.e. *cis*-strand). These elements are often binding substrates to other transcription regulating factors (e.g. several proteins) and, therefore, have an effect on gene expression on several ways. The typical activities of the modules are working as enhancers or silencers. Some of the typical activities of the elements



Bioinformatics

There are several databases which contain data for *in silico* studies. Genomic and mRNA sequences of various organisms can be found for instance from NCBI and EMBL. DNA binding matrices for various transcription factors can be found from Transfac and Jaspar. Several databases such as GEO and ArrayExpress also contain experimental data from microarray experiments.

A few studies have been published regarding to search of *cis*-acting elements. In these studies it has been searched for instance for over-represented sequences on genomic scale. Also comparative genomics has been used to search *cis*-acting elements from higher organisms which also might have a larger space between different elements. Also clustering has been used for searching of *cis*-acting elements. In clustering studies an assumption is made that similarly acting genes are regulated by same transcription factors through the same *cis*-acting elements. However, this might not be always true.

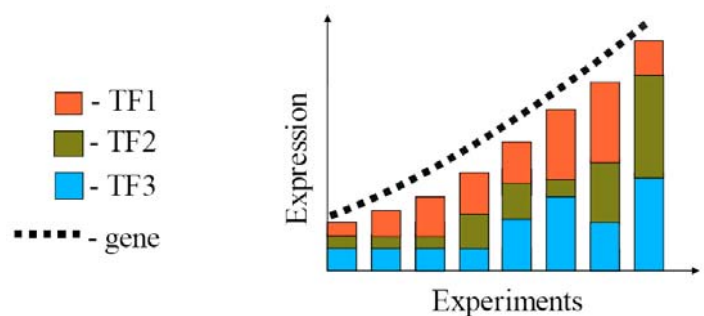
Also some *in silico* studies, in which transcription factors have been searched, have been done. However, the binding matrices are available only to a very limited number of transcription factors. In yeast data it has been searched for transcription factors that would best explain clusters of coexpressed genes, but it seems that no other methods has been used. Also it seems that there are not studies done in metazoans, just in simpler organisms.

In general, the research relating to transcription regulation as a form of a search of *cis*-acting elements and transcription factors is still very limited. Especially, not much has been done using higher eukaryotes, since by far they have turned to be too complicated and also an amount of available data remains to be very limited.

Proposal of new method

In the lecture also a new method for modeling gene expression was presented. The method assumes only main regulation mechanisms i.e. positive transcription regulation and expects that transcription factors act synergistically. One still needs to keep in mind that there are also other regulation mechanisms than transcription regulation and also there are several known negative transcription regulation cases (so called negative feedback loops). However, due to limited capacity it is not possible to take into account everything in a single model.

The method assumes that the higher level of transcription factors leads to higher expression of a gene (see attached figure). Also since the transcription factors are assumed to act synergistically their expression values are multiplied in the method. The computational capacity limits the number of multiplied transcription factors to three.

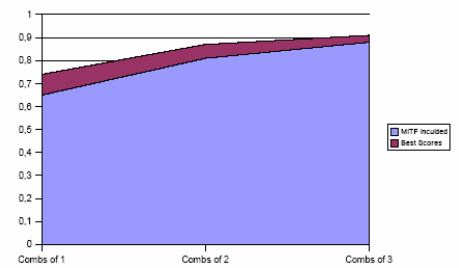


Spearman's Rank correlation coefficient is used to measure expression of genes explained by expression of transcription factors. The method is rather robust, but can be used to find correlation between gene expression and expression of transcription factors. However, it does not tell for instance how many of transcription factors really are regulating the gene expression. To obtain more combinations of the transcription factors some of the "invisible" regulators (i.e. regulators having high expression all the time) might need to be removed. Since a gene might have several enhances, which might act in different tissues differently, any data should not just be combined.

To test quality of the method a sample dataset was used. Dataset (47 arrays) for melanin producing cells was used to study a regulation of TYR gene. The results showed a good correlation between gene expression and expression of transcription factors such that found transcription factors are really known to be related to expression of TYR gene. The attached figure illustrates the results - MITF is known to regulate TYR gene and SOX10 is known to regulate MITF. The expression was not completely explained by the found transcription factors, but adding of higher number of transcription factors might fully satisfy the conditions.

The model was found to give good results, but there is still a need to improve it. There are several possibilities to do it, such as adding logical OR and NOT to it or adding sequence and protein-protein interaction data into model. However, several of these improvements might be difficult to do due to computational limitations.

In general, the new method explained the gene expression well in the used special case. However, in the higher eukaryotics the regulation is done in several levels (transcription, translation, post-translational) and plenty of work still needs to be done before the regulation in higher eukaryotics can be modeled computationally.



Observed:
 1. SOX10
 2. SOX10 SOX13
 3. SOX10 SOX13 JAZF1

MITF included:
 1. MITF
 2. MITF and several others
 3. MITF SOX10 SOX13