Searching regulatory factors for genes in higher eukaryotes

Margus Lukk

April 15, 2005

Abstract

The ultimate goal in research of higher eukaryotes is to understand their gene regulation. This paper describes new method for finding a set of regulatory factors for a gene. The method is based on common belief that transcription factors act synergically. The reliability of the method is tested on expression data of melanin producing cells. Preliminary results as well as strengths and weaknesses of the method are discussed.

Introduction

The complexity of metazoans is mainly result of alternative splicing and complex gene regulation. Both mechanisms have been studied over long period of time and general framework for transcriptional and translational regulations have been established. Despite of more than twenty years of research on transcription factors (TFs) and cis-acting promoter elements [2], regulation mechanisms are fully understood only for a tiny fraction of genes.

A typical gene in multicellular organism may contain several cis-regulatory modules, located 5' and 3' or even in introns [1]. These modules can be classified to enhancers, silencers or insulators. The latter is defined as enhancer-blocking sequences and is typically located between the distal enhancer and the promoter [3]. Regulatory modules may also be either initiation or elongation specific, or both. Each cis-regulatory module contains multiple binding sites for at least three different sequence specific transcription factors that trigger or silence the gene expression within particular tissue or cell type [4]. Therefore different sets of TFs may result expression of the same gene under different conditions. Since early days of research in eukaryotic transcription machinery, TFs have been divided to two classes. While general TFs join with RNA polymerase II and constitute together basal transcription apparatus, sequence specific TFs recognize short consensus elements in only these promoters and cis-regulatory modules which contain their appropriate binding sites [5]. In addition to TFs, there exist also transcription co-activators and co-repressors which act as mediators between sequence specific TFs and general/basal transcriptional machinery. The whole initiation complex is therefore huge containing nearly 100 polypeptides in its core plus additional sequence specific TFs[6].

Until recent, most of the regulatory discoveries have come from wet lab. As wet lab work is extremely time and resource consuming, the amount of new knowledge has been slow to come. With rapidly growing public databases of expression and sequence data, the hope for speeding up the research is now on computational methods. Traditional approach for combining microarray and sequence data has been clustering the genes based on expression similarities and then searching enriched sequence patterns in upstream regulatory regions from cluster members. Such studies are driven by the assumption, that similarly acting genes are likely to be regulated by common TF. Regardless from its relative success, the strategy is far from perfection as genes sharing the same regulatory motif may not cluster together and vice versa [7]. Several other algorithms have been developed to detect cis-acting elements without need to cluster microarray data[8][9].

Cis-acting elements have been successfully mined also from genomic sequences only. Methods applied on unicellular organisms are based on a property that binding site of a TF appears in many places in the genome and is therefore over-represented compared to random motifs [10]. For metazoans, cis-elements have been searched with the help of comparative genomics [11]. However in *Drosophila*, success has been reached also by looking for cis-regulatory modules instead of separate elements. A regulatory module has over-representation of cis-acting elements and allows thus to distinguish between background and module [15].

All described strategies can point regulatory regions of a gene, but they still do not say which transcription factors are actually using these. Binding sequences are known only for a fraction of TFs. Even worse, as individual eukaryotic sequence specific factors bind DNA with relatively low specificity, the regulatory power is often result of several factors acting together[12][6].

Several approaches, applied to yeast microarray data, have been more precise by actually trying to answer the question "which TFs regulate which genes". Seagal et al partitioned the expression data based on candidate set of regulator genes. As a result, transcriptional modules and so called regulation program, which contains TFs that best explain the expression of the module, was obtained [13]. Also Wang et al. used data partitioning in the first step, and then continued with combining expression modules with regulatory motifs for which sequence specific TFs are known[14].

For metazoans, similar studies are still missing. Here, a new method for finding TFs that regulate a gene is presented and applied on microarray data from human.

Method

Similarly to other earlier works, expression values of genes are treated as concentrations of active gene products. The logic of finding TF gene relations is opposite with all previous studies in which TF that could best explain small tight cluster of supposedly co-regulated genes is searched. Here the goal is to find group of TFs which best explains expression of one gene in particular cell type and conditions.

Let matrix M be gene expression data matrix, where each row contains expression values of some gene i and each column represents hybridization experiment j. Subsequently, let m_{ij} be expression value of gene i in experiment j in matrix M. We also denote the set of all genes as G and set of all TFs as $F, \subset G$. For each $g_i \in G$ there exists a set of transcription factors F_i that regulates the particular gene. Although the gene product can be sometimes involved in controlling its own expression, we consider such cases as exceptions and define $g_i \notin F_i$. Finally, the goal is to find set F_i for any g_i given M.

Sequence specific TFs are known to act synergically [6] and the therefore contribution of all TFs for g_i is defined as

$$\hat{F}_i = \prod_{k=1}^{F_i} m_{kj}.$$

Let it be reminded that $\hat{F}_i$ shows the principle how single expression value from a set of TFs is computed in one array system. In case of M with j experiments, each value in vector which represents TFs contributions, is computed separately for position j like in one expression array system. Higher number of TFs in F_i and higher concentration of each TF means increased expression of g_i. Although expression profile of $\hat{F}_i$ has due to the model far higher range of "expression values", in ideal it should still have perfect correlation with g_i. The correlation measure of choice is the highly robust Spearman's rank correlation. Advantage of Spearman's rank correlation coefficient comes from its usage of ranks instead of original values, a property which allows perfect correlations even in cases when two expression vectors have different scales.

F_i for g_i can be found as follows:

$$F_i = argmax_{F_i}(corr(g_i, \hat{F}_i))$$

Computation of similarly between g_i and all sets in powerset of F is not feasible as, for example, in humans $|F| \approx 4000$. The search space can still be reduced by leaving aside all regulators which expression does not vary much. These "invisible" regulators may in reality have important regulatory function, but as their regulatory activity is dependent on factors like phosphorylation, sumolation etc. they remain invisible in expression data. Nevertheless, cases of $|F_i| > 3$ are not computable on average desktop PC.

Results

Method described above was applied on melanin expression dataset (47 Affymetrix U133 arrays). The dataset contains expression measurements from various normal and cutaneous melanin producing cells (dataset not published). The goal was to find transcriptional regulators for tyrosinase (TYR) known as key enzyme of melanin production.

Set of regulators was extracted from Affymetrix website. 3882 probes that had GO annotation "regulators of transcription" were selected.

Before the method was applied on melanin expression data, distribution of correlation scores was first studied on randomized data. The expression data matrix was randomized by switching places of random pair of expression values such that in average every value in every position was moved 100 times. Resulting matrix and the set of regulators were applied on search of TFs for a randomly picked gene. Correlations of up to combinations of 3 TFs were measured. Observed correlation scores were normally distributed in range from -0.5 to 0.5 regardless from the size of F_i.

The expression of TYR was best correlated with transcription factor SOX10 (score 0.74). From pairs of TFs, the maximum score was 0.87 (SOX10 and SOX13). Finally, from triplets , SOX13, SOX10 and JAZF1 had score 0.91.

Also performance of MITF, the known TYR regulating TF was studied. Correlation between TYR and MITF expression profiles was 0.65. Several TFs combined with MITF gave correlation score 0.81. The best triplet including MITF had score 0.88 (MITF, SOX10, SOX13).

Discussion

The method did not recognize MITF as the regulator of TYR while it should have been. Instead it detected SOX10, the regulator MITF, and SOX13 which up to current knowledge does not have role in melanin production. It may be that size of $|F_i|$ that maximizes the score is greater than 3 and contains also MITF. However, given the present data and search space reduction method, it can not be computationally proved. Or another possibility, the model is oversimplification of real life and can never generate true results.

Observation of SOX10 instead of MITF shows stronger influence of distant regulatory relations than expected. The content of the set of TFs that maximizes the score may still not be completely wrong. Further study of the reliability of the method is necessary.

One of the weaknesses of the model is that it assumes expression of all TFs (logical "AND") at least in some level in all experiments. In case of one non expressed TF in F_i, the whole model collapses. Multiplication with zero is however avoided as microarray expression data shows some weak expression due to background noise even in cases when the expression actually is missing. As sometimes whole family of TFs may bind to certain binding motif, including logical OR to the model would be justified. Also, negative regulation, logical NOT respectively, is common in transcriptional regulation. However, it is questionable if such improved model would be computable in practice.

It is not clear if including new biological data from other domains than expression data would somehow help to reduce the search space. It is at least worth of considering. Surely, the attempts to mine regulatory relations between regulating TFs and their regulatory targets must continue!

References

[1] Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424:147-151.

[2] McKnight SL, Kingsbury R (1982) Transcriptional control signals of a eukaryotic protein-coding gene. *Science* 217:316-324.

[3] Conte C, Dastugue B, Vaury C (2002) Coupling of enhancer and insulator properties identified in two retrotransposons modulates their mutagenic impact nearby genes. *Molecular and Cellular Biology* 22:1767-1777.

[4] Davidson EH (2001) Genomic Regulatory Systems: Development and Evolution. *Academic Press.*

[5] Lewin B (1999) Genes VII. *Oxford University Press.*

[6] Kadonaga JT (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 116:247-257.

[7] Li H, Wang W (2003) Dissecting the transcription networks of a cell using computational genomics. *Current Opinion in Genetics & Development* 13:611-616.

[8] Bussemaker HJ, Li H, Siggia ED (2001) Regulatory element detection using correlation with expression. *Nature Genetics* 27:167-171.

[9] Liu XS, Brutlag DL, Liu JS (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology* 20:835-839.

[10] Brazma A, Johanssen I, Vilo J, Ukkonen E (1998) Predicting regulatory elements *in silico* on a genomic scale. *Genome Research* 8:1202-1215.

[11] Frazer KA et al (2003) Cross-species sequence comparisons: a review of methods and available resources. *Genome Research* 13:1-12.

[12] Walter J, Dever CA, Biggin MD (1994) Two homeo domain proteins bind with similar specificity to a wide range of DNA sites in Drosophila embryos. *Genes Development* 8:1678-1692.

[13] Segal E. et al (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* 34:166-176.

[14] Wang W. et al (2002) A systematic approach to reconstructing transcription networks in Saccharomyces cerevisiae. *PNAS* 99:16893-16898.

[15] Rajewsky N, Vergassola M, Gaul U, Siggia ED (2002) Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics* 3:30-43.