 ISBN: 978-974-8285-62-7

The *Beacon Tools* for the Analysis of Gene Expression and Its Regulation

Federico Zambelli
Department of Biomolecular Science and Biotechnology
University of Milan, Italy
Federico.zambelli@unimi.it

Matteo Ré
Department of Computer Science
University of Milan, Italy.
matteo.re@unimi.it

Giulio Pavesi
Department of Biomolecular Science and Biotechnology
University of Milan, Italy
giulio.pavesi@unimi.it

ABSTRACT

Bioinformatic tools have become essential for researchers in modern molecular biology and medicine, allowing them to take advantage of the wealth of data currently available. In this article we present the tools developed by the Bioinformatics, Evolution, and Comparative Genomics (Beacon) group at the University of Milano, that provide a computational pipeline for the analysis of gene expression and its regulation. Through dedicated web interfaces users can retrieve expression data for their gene(s) of interest, as well as study their regulation, i.e. identify factors and sequence elements playing a key role for the modulation of their transcription. The tools are available through a dedicated web interface at either http://www.beacon.unimi.it/webtools http://159.149.109.11/webtools

Keywords

Bioinformatics, gene expression, microarray data, transcription factor, transcription factor binding sites.

1. INTRODUCTION

As more and more data become available for the scientific community, bioinformatics tools and methods are bound to play a major role in modern research. In particular, two sources of information are now commonly used in molecular biology: sequence data, and expression data. Genomic sequences can tell researchers where genes are located, while typical microarray based experiments give a picture on when genes are expressed, that is, transcribed into a RNA molecule (which in turn will be translated into a protein), and at which level, that is, measure the quantity of mRNA produced by a gene in any given condition.

Clearly, not all genes are simultaneously expressed at any time in any cell. The complete understanding of the complex mechanisms regulating gene expression is however far from having being reached. In particular transcription is finely modulated by living cells according for example to developmental stage, tissue type, external stimuli, and so on, and malfunctioning in this mechanism often leads to the onset of diseases like cancer. Key players in the regulation of gene transcription are *transcription factors* (TFs), which bind DNA in a sequence-specific fashion, usually in the region immediately preceding the transcription start site (called *promoter*), but also further away from the gene, in distal regulatory regions called *enhancers* or *silencers* for the effect they have on the transcription of the gene. Sites along the DNA bound by TFs are called *transcription factor binding sites* (TFBSs). The combined action of several different TFs allows the transcriptional machinery of the cell to recognize the starting points of genes along the genome and initiate transcription [1].

Since now several full genomic sequences are available, an ideal goal would be being able to annotate not only the gene repertoire on each one, but also all the TFBSs as well as the respective TFs, so to produce a complete picture of where genes are located, when and how much they are expressed, and which are the factors regulating them as well as their binding sites. For this task, bioinformatics tools can provide substantial help. Unfortunately, all sites recognized by a given TF have not to be perfectly identical to be bound. In other words, each TF recognizes a set of short (6-20 nucleotides) and similar short sequence fragments (*oligonucleotides*, or *oligos*). Thus, the reliable computational identification of TFBSs along a genomic sequence is far from being an easy task, and can be considered one of the most challenging and open problems in current computational biology.

2. RELATED WORK

Several databases devoted to gene expression data already exist, from species or condition specific ones to more general repositories like the Gene Expression Omnibus (GEO) maintained by the NCBI. In many cases, however, retrieving and analyzing the data is far from being an immediate task for a researcher not skilled in bioinformatics and statistics. Likewise, regulation of transcription is nowadays studied extensively, but while new approaches and algorithms are introduced almost on a daily basis, it is often very hard to understand the potential and the limitations of the different tools available, and to mine the wealth of existing literature to find out which tool is most suitable for a given analysis.

In our project we had a two-fold goal: first, make the retrieval of expression data and information therein as much immediate as possible, bypassing the need of expert knowledge of databases or involved statistical analyses. On the other hand, we wanted to build a set of different bioinformatic tools allowing researchers to investigate the regulation of their gene(s) of interest simultaneously from different viewpoints, merging and integrating different approaches. All in all, we provide a unique resource that can be used to perform a complete study on the expression of one or more genes and its regulation.

3. *HUNTED*: HUMAN TISSUE EXPRESSION DATABASE

The database interface allows users with little or no knowledge of database query languages or systems, nor any experience in statistical analysis of expression data, to retrieve data for their genes of interest or sets of genes showing similar or specific expression patterns. At the moment, the database includes two major datasets, for the tissue-specific expression of human genes in normal and tumoral cells. The "normal tissue database" contains the data presented in [2], while tumor data was collected by the Expression Project for Oncology (expO) project (https://expo.intgen.org/geo/). These two sources, which are part of an ongoing research project in co-operation with a biomedical company, have been chosen for their reliability and for the fact that data were generated on the same platform, thus allowing for the direct comparison of the expression of genes in normal and in cancer cells.

3.1.1 Retrieving a single gene

Users can query the database for the expression profile of a given gene in either dataset or in both, obtaining a graphical representation of its variation. To assess the tissue specificity of the gene investigated, absolute expression values have been transformed into relative values. Let e_i be the expression level of gene in tissue i, and μ, σ mean and standard deviation of the expression level of the gene across all tissues available. We transformed the expression level e_i into its z-score:

$$z_i = \frac{e_i - \mu}{\sigma}$$

In this way, the specificity of the expression of the gene can be directly assessed: a gene mainly expressed in a given tissue will have a significantly high z-score, and vice versa for a gene not expressed in a tissue. An example is presented in section "Experimental Results".

3.1.2 Retrieving co-expressed genes

A typical analysis performed on expression data is to single out sets of genes which share a similar expression profile, by grouping them into clusters. Several clustering algorithms have been proposed, each one with its pros and cons. In our implementation, we chose instead two simpler approaches.

In the first, coupled with the single gene query option, the interface outputs a list of genes ranked according to the correlation of their expression with the expression of the query gene, that is, by measuring how close their expression profile is. As distance we chose the Pearson correlation coefficient, which is independent of absolute expression values but rather considers the similarity in the variation of expression of two genes in two or more experiments. In this case, thus, users starting with a gene of interest can easily retrieve its "expression neighbors".

The second option we included is to query the database for genes whose (absolute or relative) expression is above or below a given threshold in a given condition. Thus, for example, users can retrieve all genes over- or under-expressed in a given tissue, or for which the expression have the greatest change between normal and tumor tissues. The difference from the first strategy is that in this case the comparison is performed within a single tissue (experiment), rather than on the variation of the expression across all tissues. In either case, users can retrieve a set of genes whose expression is modulated in a similar fashion. Other than providing valuable information by itself, this result leads in turn to the natural question of which are the TFs that are responsible for the common pattern of expression detected.

4. THE TOOLS

As previously described, transcription is modulated by TF binding the DNA in a sequence-specific manner. From a computational viewpoint, the problem is usually tackled in three different ways [3]:

- If available, a descriptor of the binding specificity of a given TF can be used, in order to scan genomic sequences and to predict its binding sites in the proximity (promoter) of one or more genes.
- Promoter sequences from a set of co-regulated genes can be analyzed looking for recurrent similar sequence elements, or motifs: since these genes are co-regulated, the recurrent motifs could represent TFBSs for their common regulators.
- Genome comparisons have shown how, in closely related species like mammals, not only genes tend to be conserved in sequence, but also the sequence elements responsible for their regulation. Thus, for example, a promoter region fragment conserved between human and mouse is likely to be or contain binding sites for TFs regulating the corresponding gene.

All these approaches are usually based on the same underlying principle: redundancy yields information. The sequences investigated should contain similar oligos (likely to be bound by the same TF), and a similar set should be not found if the sequences investigated were

unrelated or picked at random, that is, came from genes with different expression patterns. Our pipeline includes three different tools for this task, each one covering one of the three points outlined above.

4.1 Pscan

The binding specificity to the DNA of TFs can be studied with different lab approaches, from mutagenesis to footprinting to more modern techniques like SELEX and chromatin immunoprecipitation. Several studies of this kind have been published, and the result is that for many TFs of interest we have at our disposal a collection of sites experimentally known to be bound by the TF in vivo or in vitro.

The next logical step is to employ the TFBS collection available for a given TF for building a *descriptor* of its binding specificity on the DNA, that in turn can be used to scan DNA sequences looking for potential novel binding sites. Since TFBSs for the same factor have usually the same size, the most immediate way is to align the sites, in order to determine which nucleotides are favored at each position of the sites. In this way, a *frequency matrix* (or *profile*) can be built, describing the frequency with which each nucleotide appears at each position of the sites available [4]. Figure 1 shows an example, together with the "logo" representation of the matrix that gives an immediate idea on which nucleotides are most conserved at each position of the sites [5]. In turn, the frequency matrix can be used to assess the likelihood, for a given oligo of the same size of the matrix, to be a binding site for the same TF for which the matrix was built. Analogously, a DNA sequence (e.g. a promoter) can be scanned with the matrix looking for oligos that fit it well, likely to be instances of binding sites for the TF.

Although quite naïve (for example, dependencies among the nucleotides of the sites are not modeled), this idea can be very useful in many circumstances, for example when one has put together a set of co-regulated or co-expressed genes. TFs for which "good sites" (according to the corresponding profile) can be found in a statistically significant way in the promoter of the genes are likely to be responsible for their co-expression. The tool we included in our web server, named Pscan, is based exactly on this principle. It takes as input gene identifiers, retrieves automatically the corresponding promoter regions (whose size can be in turn defined by the user), and scans them with the TF profile collections available in specialized databases like JASPAR [6] and TRANSFAC [7].

Sequence analysis is performed by computing for each matrix the best possible score on each promoter sequence (that is, the score of the highest scoring oligo, according to the matrix). Then, for each matrix, mean and standard deviation on the input gene set (the sample) are compared to the mean and standard deviation of the same matrix on the whole promoter set of the organism from

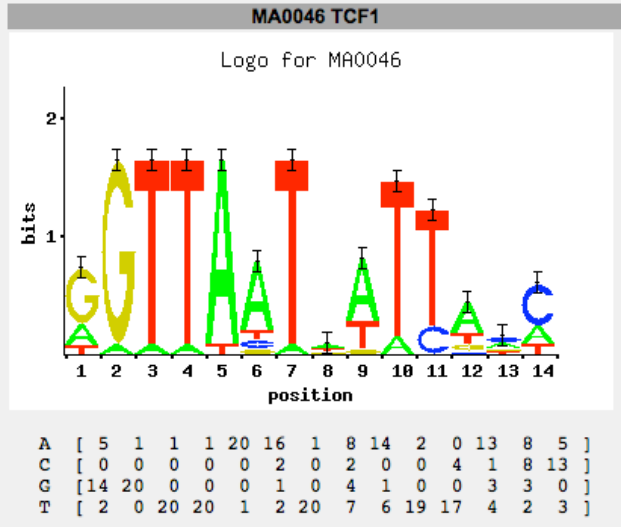

Figure 1. Frequency matrix (profile, bottom) for TF TCF1, retrieved from the JASPAR database [6], and its "Sequence Logo" representation [5].

which the sequences were taken (the universe). The comparison is performed with a z-test, and yields for each matrix available a p-value, representing the probability of obtaining the same result on a set of random genes of the same size of the input set [8]. Standard p-value thresholds for statistical testing (i.e. 0.05 or 0.01) can be employed to single out significant results, that is, TFs for which the matrix has good enough sites in the promoter sequences investigated, candidate to be responsible for the co-expression of the genes.

4.2 Weeder

The number of TFs in the human genome has been estimated to be in the thousands. However, frequency matrices describing TFBSs are available for a limited subset of TFs, since for example the JASPAR database contains less than 100 matrices for mammals and TRANSFAC a few hundreds (many of which redundant, that is, different matrices are available for the same TF, and many not reliable enough, since built with small collections of sites). Discovery of putative regulatory elements can be anyway performed *ab initio*, that is, without resorting to descriptors of the binding specificity of TFs. Given a set of promoters from co-regulated genes, the general idea, as stated before, is to single out one or more sets of oligos similar enough to one another to be considered binding sites for the same TF, and to evaluate whether the same set would appear with the same degree of conservation in a set of sequences picked at random.

Since TFBSs are short and often quite degenerate, the problem formalized in this way is very challenging, independently from the algorithm used for its solution. The tool we include in out pipeline, called Weeder [9], can be however considered as the state of the art in this field [10], and has been used several times many research groups worldwide with good results. The idea is to enumerate all possible oligos of feasible length (6-12 nucleo-

tides) and to collect their approximate occurrences in the sequences (allowing a feasible number of substitutions). Then, the algorithm evaluates how likely is each oligo to appear with the same number of occurrences and degree of conservation in a set of random sequences from the same organism [9]. Also in this case, users have only to input the promoter sequences of their genes. Computation takes longer in this case, and the results can be received by e-mail.

4.3 WeederH

WeederH is an algorithm that compares genomic sequences from different species for the discovery of regions conserved by evolution that are likely to play a key functional role. While the latter fact is clearly known for protein-coding regions, it has been discovered how conserved regions can be easily found in non-coding DNA as well. For example, a typical human-rodents comparison of promoter regions can single out conserved TFBSs, or a comparison of whole inter-genic regions can identify conserved elements that can function as distal regulatory elements like transcription enhancers or silencers.

Phylogenetic footprinting, that is, the comparison of non coding DNA regions from different species aimed at the detection of conserved functional elements has thus become a routine technique in modern biology. Several issue are nevertheless still open in algorithm design for this task. In WeederH we introduced a measure of "relative conservation", that is, looks for sequence regions that are significantly more conserved than the surrounding ones. What the program requires as input are sequences of any length from orthologous genes, e.g. the promoters of the genes or the intergenic regions upstream or downstream of the genes [11].

To perform an analysis with WeederH users just have to input two or more orthologous sequences from different species. The program detects putative conserved regulatory elements in a bottom up fashion, that is, first looks for conserved 8-12 base pair sequences, and then tries to combine them into longer elements. "Relative conservation" is measured according to both sequence similarity and position with respect to the gene: the idea in fact is that a functional element should have been preserved by evolution both in sequence (to be bound by the corresponding TFs) and in position, so to arrange spatially the machinery for transcription initiation. The output is a ranked list of the most conserved sequence fragments, as well as whole significantly conserved regions likely to represent silencers or enhancers.

The analysis performed with WeederH can provide useful indications when researchers want to study the regulation of a single gene, and in turn the most conserved regions can be examined for good matches with the profiles available with Pscan. Alternatively, a whole set of genes can be analyzed one by one, and an analysis with Weeder can be performed on "masked" sequences, that is, sequences in which only the most conserved regions are kept (masked sequences are automatically output by WeederH as well). In this way, false positives can be significantly reduced, by limiting the analysis only to those portions which are conserved throughout different species, and thus more likely to play a functional role.

5. EXPERIMENTAL RESULTS

The performance and scope of applicability of the various tools have been discussed in previous articles [8-11]. Here, we illustrate with a working example how their results can be integrated to perform a comprehensive analysis of the expression profile of a given gene as well as its regulation.

The albumin (ALB) human gene encodes a soluble, monomeric protein which comprises about one-half of the blood serum protein. Albumin functions primarily as a carrier protein for steroids, fatty acids, and thyroid hormones and plays a role in stabilizing extracellular fluid volume. Mutations in the ALB gene on chromosome 4 result in various anomalous proteins. Albumin is synthesized in the liver as preproalbumin, which in turn implies that the gene encoding for this protein has a strong liver-specific expression, and in turn has to be regulated by liver-specific transcription factors like hepatocyte nuclear factors. In the following we show how all these pieces of information can be gathered by using our database as well as our tools.

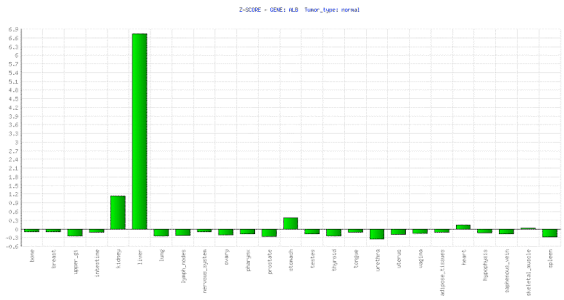

Figure 2 Expression of the albumin gene (z-scores) in human tissues, showing a marked peak in liver cells.

5.1.1 Retrieving Albumin Expression Data

As described before, expression data can be retrieved in two ways. The most straightforward one is to query the database for the ALB gene in normal tissues, which yields the plots shown in Figure 2. ALB shows a marked peak in liver-specific expression (z-score > 6.5). Furthermore, a set of genes with highly similar expression profile (Pearson correlation > .9) can be detected, ideal candidates for a study of factors governing their co-expression, in this case liver-specific expression.

A dual approach, instead, can be to query the database for genes showing liver-specific expression, by retrieving all genes whose expression (z-score) is greater

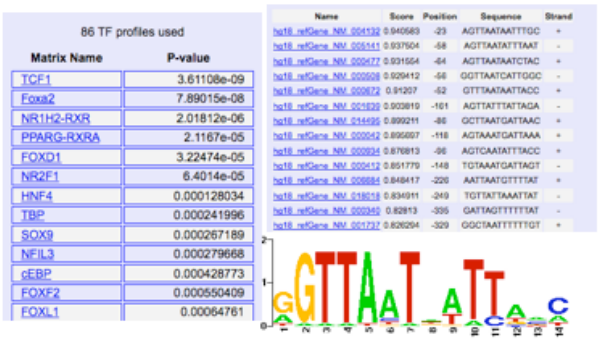

Figure 3. Ranking of the highest scoring matrices (left) in the liver-specific gene set, and predicted TFBSs in each promoter for factor TCF1 (HNF1A, top right) and logo representation of the TCF1 binding site (bottom right).

than a given threshold. In this case we would have as a results that albumin ranks among the genes preferentially expressed in liver, and in turn, this set of genes could be singled out for further analysis on their regulation.

5.1.2 Finding TFs and TFBSs regulating albumin

Once a set of genes showing a marked liver-specific expression has been selected, their promoters can be analyzed for binding sites of known TFs with Pscan, or for conserved sequence motifs likely to play a functional role with Weeder and WeederH.

First of all, we submitted to Pscan albumin as well as the set of genes with expression correlation higher than .9, choosing to analyze the region between -450 base pairs to +50 base pairs with respect to the transcription start sites of the genes. The result is shown in Figure 3, with the list of TFs included in the JASPAR database ranked by increasing p-value. At the top of the list we have matrices corresponding to Hepatocyte nuclear factors (TCF1, HNFs, FOX), which are a group of transcription factors expressed predominantly in liver that ensure liver specific expression of their target genes, as well as other factors known to be involved in the regulation of albumin and other liver-specific genes like proliferator activated receptors (PPARs) and retinoid x receptors (RXRs). The low p-value returned by the z-test on these matrices are clear indicators of the significance of the results obtained, as well as of the likelihood of the corresponding TFs to be involved in the regulation of the genes.

On the other hand, the ab initio analysis performed by Weeder on the same set of promoters discovered significantly conserved motifs that resemble closely the ones employed by Pscan, occurring at the same locations. An example is shown in Figure 4, in which it can be clearly seen how the highest scoring motif reported by Weeder (the alignment of the motif instances found, top) is virtually equal to the motif describing the binding specificity of TCF1, the highest scoring matrix reported by Pscan.

The regulation of the albumin gene can be further investigated by comparing it to its orthologues in other species, for example to mouse and rat albumin genes. As described before, WeederH does not limit the analysis to the promoter region, but can be applied to longer sequences spanning whole intergenic regions. Thus, in this case, we not only want to determine which are the most conserved (and most likely to be functional) sites in the promoter (for which Pscan and Weeder already gave good indications), but also to determine whether there exists a conserved region outside the promoter likely to regulate the transcription of the gene as well. As a matter of fact, transcription regulation in metazoa is not limited to promoters, as in yeast and plants, but it often involves sequence regions that can be located at several thousands of base pairs from the gene itself. In literature, a region of this kind, enhancing transcription, has been reported to be located approximately 10,000 base pairs upstream of the albumin transcription start site [12] .

We submitted to WeederH the 25,000 base pairs upstream of the albumin human gene, as well as a region of the same size taken upstream of the albumin orthologous genes of mouse and rat. The bottom up analysis performed by WeederH can be read at different levels. In the promoter region, we could validate the fact that most of the TFBSs predicted by Pscan (for the highest scoring matrices in the z-test) and Weeder in the albumin promoter are actually conserved in mouse and rat, further evidence to the functionality of the sites. At an higher level, the output of WeederH singled out that, apart from the promoter, there exist another evolutionarily conserved region located approximately at -10,000 base pairs from the transcription start site of the gene, in correspondence to the annotated enhancer. If this analysis were performed lacking any information on the existence of distal enhancers, the conserved region detected would be a good candidate for experimental in vivo or in vitro testing.

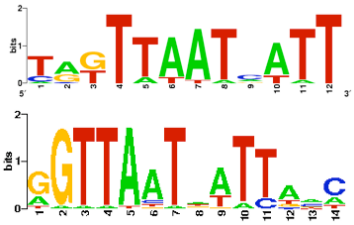

Figure 4. Highest scoring motif output by Weeder on the liver-specific gene promoter set (top) and representation of the binding specificity of TCF1 (bottom).

5.1.3 Results summary

All in all, by applying all tools included in the pipeline we could derive most of the information currently available for the *albumin* gene. More in particular:

1) It has a marked liver-specific expression, and by retrieving a set of genes with similar expression profile we could determine a set of transcription factors potentially able to confer liver-specific expression to genes.

2) Sequence analysis of the promoters yielded sequence motifs that correspond to sites recognized by the TFs identified at the previous step. If this information were not available, we would have identified some of the sequence elements in the genes' promoters fundamental for the regulation of their expression.

3) Most of the TFBSs predicted by Pscan and Weeder in the albumin promoter are conserved in mouse and rat.

4) The albumin gene has a distal regulatory element, that corresponds to a evolutionarily conserved region located about 10,000 base pairs upstream of the transcription start site of the gene. With WeederH we identified a genomic region significantly conserved, likely to play a functional role in its regulation.

Clearly, albumin is a very well studied gene, with an ideal expression profile in a tissue (liver) for which most of the factors involved (as well as several of their binding sites along the genome) have already been identified and characterized. However, it serves as good example of the potential of the tools we present, that can significantly narrow down the scope of the necessary experimental validations that have to be performed to confirm the predictions.

6. CONCLUSIONS AND FUTURE WORK

Bioinformatic analysis is a first fundamental step in any modern study in molecular biology. Our gene expression database, as well as the tools for the identification of common regulators and regulatory sequence elements, provide a comprehensive set of instrument that allow researchers to investigate and characterize gene expression and its regulation. Clearly, each of the tools can be used separately: the Weeder algorithm, for example, has been introduced a few years ago, and now it can be considered one of the de facto standards for the identification of conserved sequence motifs. WeederH has been presented last year, and it is gaining more and more users. Pscan and the expression database are the latest additions, that complete the picture and we believe are going to be of great interest for researchers. All the tools are freely available (with downloadable standalone versions) for academic and non-profit users. We are currently considering the possibility of releasing commercial custom-tailored software packages.

The gene expression database is still at an early stage of development, and will be soon augmented with new datasets in human and other species. Furthermore, we are currently testing novel algorithms that bypass the two-step process of gene selection followed by sequence analysis, by integrating in a single approach expression and sequence data. Finally, once every tool has produced its output, the merging of the results is still mostly left to the user. We are currently working on having the web interfaces performing this step in an automated way, so to make the usage of the tools more user-friendly and less prone to human-biased evaluation of the results.

7. REFERENCES

[1] M. Levine and R. Tjian, "Transcription regulation and animal diversity," *Nature*, vol. 424, pp. 147-51, 2003.

[2] R. B. Roth, P. Hevezi, J. Lee, D. Willhite, et al., "Gene expression analyses reveal molecular relationships among 20 regions of the human CNS," *Neurogenetics*, vol. 7, pp. 67-80, 2006.

[3] K. D. MacIsaac and E. Fraenkel, "Practical strategies for discovering regulatory DNA sequence motifs," *PLoS Comput Biol*, vol. 2, pp. e36, 2006.

[4] G. D. Stormo, "DNA binding sites: representation and discovery," *Bioinformatics*, vol. 16, pp. 16-23, 2000.

[5] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Res*, vol. 14, pp. 1188-90, 2004.

[6] A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, et al., "JASPAR: an open-access database for eukaryotic transcription factor binding profiles," *Nucleic Acids Res*, vol. 32, pp. D91-4, 2004.

[7] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, et al., "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes," *Nucleic Acids Res*, vol. 34, pp. D108-10, 2006.

[8] G. Pavesi and F. Zambelli, "Prediction of over Represented Transcription Factor Binding Sites in Co-regulated Genes Using Whole Genome Matching Statistics," *Lecture Notes in Computer Science*, vol. 4578, pp. 651-658, 2007.

[9] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole, "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes," *Nucleic Acids Res*, vol. 32, pp. W199-203, 2004.

[10] M. Tompa, N. Li, T. L. Bailey, G. M. Church, et al., "Assessing computational tools for the discovery of transcription factor binding sites," *Nat Biotechnol*, vol. 23, pp. 137-44, 2005.

[11] G. Pavesi, F. Zambelli, and G. Pesole, "WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences," *BMC Bioinformatics*, vol. 8, pp. 46, 2007.

[12] Y. Kajiyama, J. Tian, and J. Locker, "Characterization of distant enhancers and promoters in the albumin-alpha-fetoprotein locus during active and silenced expression," *J Biol Chem*, vol. 281, pp. 30122-31, 2006.