A Novel Bioinformatics Technique For Predicting Condition-Specific Transcription Factor Binding Sites

Valmik Desai[1], Purvesh Khatri[1], Arina Done[1], Aviva Fridman[2], Michael Tainsky[2], Sorin Draghici[1,†]
[1]Department of Computer Science, Wayne State University, 431 State Hall, Detroit, MI 48202, USA
[2]Karmanos Cancer Institute, Detroit, MI 48202, USA
[†]To whom the correspondence should be addressed. Phone: 313-577-5484, Fax: 313-577-6868, E-mail: sorin@wayne.edu.

***Abstract*— The advent of high throughput sequencing and DNA microarray technologies along with the advances in bioinformatics have revolutionized biological research in the recent years. However, the precise mechanisms that control gene expression are largely unknown despite the numerous efforts to understand them. We describe a bioinformatics technique that can potentially identify condition-specific transcription factor binding sites. We applied our technique to cellular immortalization data set. Our analysis revealed similarities in upstream regions of CXCL gene family that explain condition-specific differential expression of genes CXCL1 and CXCL2, versus gene CXCL3.**

I. Introduction

High-throughput genome sequencing and mapping techniques continue to generate enormous amount of data describing gene structures (i.e., nucleotide sequences of the genes, their exons and introns) along with their genomic locations. Concurrently, high-throughput DNA microarrays have revolutionized the biotechnology by allowing the researchers to screen tens of thousands of genes simultaneously. The interpretation of the microarray results is further facilitated by Gene Ontology (GO)-based functional profiling approaches [1], [2] as well as biological pathway databases [3], [4] and pathway analysis tools [5], [6]. However, despite these advances in biotechnology and bioinformatics, the exact mechanisms responsible for gene regulation are largely unknown.

Gene regulation is accomplished by complex and tightly coordinated protein-DNA interactions between a number of transcription factors and the promoter region(s) of a gene. There have been large number of efforts since the early 1980s to predict transcription factor binding sites (TFBSs) *in vivo*, *in vitro*, and *in silico*. Some of the popular *in vivo* approaches include Chromatin Immunoprecipitation (ChIP) and DamID [7], [8], whereas *in vitro* approach include systematic evolution of ligands by exponential enrichment (SELEX) [7]. There are number of promoter sites databases developed [9], [10], [11] that contain promoter sites characterized through experimental techniques such as DNase I footprinting and gel retardation. However, these procedures are limited since they are driven by known or putative binding sites or assumptions about where a transcription factor may bind. In order to guide these wet bench techniques in providing valid results, a number of *in silico* approaches have been proposed including neural networks, evolutionary computation, fuzzy systems, etc. These *in silico* approaches can be broadly categorized in to 3 categories: (i) weight matrix-based approaches that use a weight matrix built from a set of known TFBSs to predict novel TFBSs [12], [13], [14], [15]; (ii) expectation maximization-based approaches that also use weight matrix. However, they differ from the weight matrix-based approaches in that the weight matrix is constantly updated using the novel sequences until it converges [16], [17], [18]; (iii) Gibbs sampling-based approaches, which are stochastic equivalents of expectation maximization-based approaches [19], and use maximum *a posteriori* log-likelihood ratio [20], [21], [22]. As disadvantages, weight-matrix based approaches require *a priori* knowledge of known TFBSs and Gibbs sampling-based approaches provide different results each time they are applied to the same data set due to their stochastic nature.

In spite of these efforts, there are cases of transcription factors binding to sites not characterized by the calculated or published sequence [23]. One of the reasons for our lack of understanding of gene regulation mechanisms is the fact that transcription factors can bind to different sites for different actions. For instance, Motohashi et al. observed that MafK binds to a different promoter in neural cells than the mesodermal-specific promoter [24]. In this paper, we describe a bioinformatics technique to identify TFBSs for co-regulated genes. Our fundamental hypothesis is that co-regulated genes involved in related biological phenomena are regulated by similar transcription mechanism [13], [25], [26], [27], [28]. Our proposed technique uses GO-based functional profiling approach to group co-regulated genes, obtained from microarray data that are involved in same biological process or molecular function in the condition studied. Consequently, our technique can potentially identify condition-specific TFBSs. Identification of such condition-specific TFBSs can be immensely useful to drug developers in pharmacogenomics. Our proposed technique is implemented as a software tool, Promoter-Express (PE), and is a part of Onto-Tools.

II. Analysis methods

GO provides an organism-independent, controlled vocabulary organized in three principal categories: molecular functions, biological processes, and cellular components [29]. As defined by GO, a molecular function is an activity at the

molecular level; a biological process is a series of events accomplished by one or more ordered assemblies of molecular functions; and a cellular component is a component of a cell where a particular molecular function or a biological process takes place. Currently, an automatic functional profiling approach using GO, first proposed by Onto-Express [2], has become a *de facto* standard in microarray data analysis [1]. In the paper, we describe a technique that uses this functional profiling approach to classify genes involved in similar biological processes. However, note that the technique is equally applicable if genes are grouped by molecular functions.

The functional profiling identifies the biological processes that appear in the set of differentially regulated genes in a manner that is significantly different than expected by chance. Such processes are expected to be more relevant to the condition under study. For each of the genes involved in the selected biological process, we analyze the nucleotide sequence in its upstream region. By default, PE analyzes a nucleotide sequence 1500 base pairs upstream and 500 base pairs downstream from the start of an mRNA since most of the transcription factor binding sites are more likely to be in this region. However, PE allows the user to specify the region of the sequence to be analyzed. Note that in most cases, the start of an mRNA corresponds to its transcription start site (TSS). However, in case of some mRNAs, the TSS may not be annotated precisely, or not annotated at all. Therefore, in order to avoid any confusion, we use the term "start of the mRNA" instead of "transcription start site" PE retrieves the nucleotide sequence downstream of the start of an mRNA because some regulatory elements (e.g., downstream promoter elements [30]) are known to occur downstream of the start of the mRNA. Then, PE performs a pairwise sequence comparison using a sliding window approach. By default, PE uses a window size of 9 since most of the known transcription factor binding sites are 6 to 20 base pairs long [31]. However, PE also allows the user to specify the a different window size as an input parameter.

The result of the pairwise sequence comparisons is a number of exact matching subsequences found on both the input sequences. We define each of these exact matching subsequences as an *element*. It has been shown previously that two genes involved in a similar biological process and regulated by the same transcription regulation mechanism may require that these elements appear in the same order and approximately at the same distance on both genes [32]. Hence, after finding the exact matching elements in both genes, PE searches for the combinations of those elements (i.e., the exact matching subsequences) that appear in the same order with approximately the same distance among the elements on both genes (Figure 1). A set of elements that satisfies these criteria is defined as a *module*. Such a module represents a "footprint" of the transcriptional regulatory mechanisms at work in a specific biological context. However, we note that there are examples of TFBSs that have been swapped in order, and can be located at different positions. Consequently, our proposed technique may miss such TFBS (i.e., identify them as false negatives). However, it is difficult to estimate the number of false negatives at this moment due to the limited amount of information about the known TFBS. For instance, TRANSFAC database only contains 1,388 TFBSs for human.

III. Results and Discussion

We applied our technique to 39 interferon-regulated genes after cellular immortalization [33]. The senescence was produced by the treatment of immortal Li-Fraumeni fibroblasts with 5'-aza-2'-deoxycitidine (5AZA-dC). We used the tree view in Onto-Express [2], [34], [35], [36] to categorize the 39 genes into groups such that each group consists of the genes involved in the same or closely-related biological processes or molecular functions as defined by GO [37]. One such group consisted of 3 genes, chemokine (C-X-C motif) ligand 1 (CXCL1), chemokine (C-X-C motif) ligand 2 (CXCL2), and interleukin 8 (IL8), involved in *cytokine activity*. Although the microarray experiment showed that these genes were co-regulated in the condition under study, they were not known to be involved in the same molecular function. Also note that the groups are not completely disjunct. In other words, the genes are not involved in one and only one molecular function (or a biological process). A gene can be involved in more than one molecular function (or a biological process). For instance, gene IL8 is also involved in *protein binding*. These genes were submitted to PE with the window size of 8 instead of the default window size in order to increase the sensitivity. PE detected a module consisting of 25 elements for CXCL1 and CXCL2 genes. The module extended from -209 to 498 from the start of the mRNA for CXCL1, and from -191 to 498 from the start of the mRNA for CXCL2. The part of the module that was downstream of the start of the mRNA consisted of the first and the second exons of the genes CXCL1 and CXCL2. This part of the module was excluded from the further analysis since the exons are not thought to play a regulatory role in transcription. The 5 elements considered in the further analysis are shown in Table I.

Next, we searched each of these 5 elements in TRANSFAC database [11] in order to see whether any element or part of an element that we identified contained a known TFBS. Table I shows the 15 genes that are known to have an element or a part of an element sequence as a TFBS, and the transcription factors that are known to bind to those sequences in TRANSFAC. Additionally, when we found an element sequence in TRANSFAC, we searched for transcription factors that are known to bind to the sequence. These are also shown in the same table.

As shown in Table I, each of the 5 elements contained at least one known TFBS. This confirmed the ability of our approach to detect true TFBS. Interestingly, part of the third element is a known TFBS for NF-kappaB in all 3 input genes (GGAATTTCCCT in IL8, and TCCGGGAATTTCCCTG in CXCL1 and CXCL2). It has been shown in the literature that these sites play crucial role in regulation of IL8 [38], [39] as well as CXCL1 and CXCL2 [40]. Further, we created functional profiles of the 15 genes using Onto-Express in order to verify whether these genes regulated by common mechanism, are involved in related biological processes. Figure 2 shows the results of functional profiling. *Cytokine activity* is

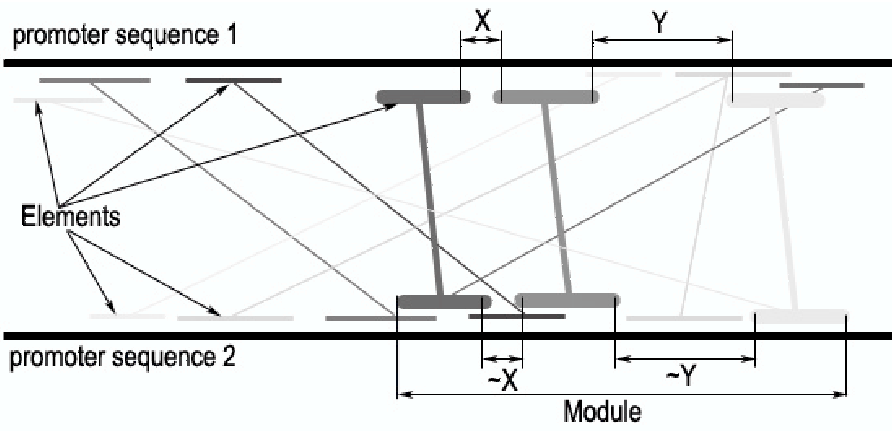

Fig. 1. Elements and module. The elements are represented by the small thin and thick lines. All the thick elements together represent a module. X, Y, ∼X, ∼Y represents the gap between adjacent elements of the module on corresponding promoter sequences.

TABLE I

TRANSFAC SEARCH RESULTS FOR EACH ELEMENT IN THE TRANSCRIPTION MODULE FOR GENES CXCL1 AND CXCL2. FOR EACH ELEMENT, THE TABLE SHOWS THE GENE(S) IN WHICH IT IS KNOWN TO BE A TFBS, AND THE TRANSCRIPTION FACTOR KNOWN TO BIND TO THE ELEMENT. THE ACTUAL BINDING SITE OF A TRANSCRIPTION FACTOR IS CAPITALIZED. NOTE THAT MELANOMA GROWTH-STIMULATING ACTIVITY ALPHA, MELANOMA GROWTH-STIMULATING ACTIVITY BETA, AND MELANOMA GROWTH-STIMULATING ACTIVITY GAMMA ARE SYNONYMS OF GENES CXCL1, CXCL2, AND CXCL3, RESPECTIVELY.

Element sequence	**Genes found in TRANSFAC**	**Transcription factors**
tcgCCTTCCTTCCg	delta-globin (HBD)	
actcGGGATCGATCTGGA	melanoma growth-stimulating activity alpha (CXCL1)	HMG I, HMG Y
ctccgGGAATTTCCctggcccgg	glycoprotein hormone alpha-subunit (GCA)	glucorticoid receptor (GR), NR3C1
ctccgGGAATTTCCCTggcccgg	interleukin-8 (IL8)	NF-KappaB, NF-KappaB complex, p50/p65
cTCCGGGAATTTCCCTGgcccgg	melanoma growth-stimulating activity alpha (CXCL1), melanoma growth-stimulating activity beta (CXCL2), melanoma growth-stimulating activity gamma (CXCL3)	NF-KappaB, NF-KappaB complex, p50/p65
ctccgggctttccagccccaaccATGCATAAAaggggttcgc	chorionic somatomammotropin (CSH1)	GHF-2, LSF-1, Pit-1, PUF-1, PUF-I, Pit-1a
ctccgggctttccagccccaaccATGCATAAAaggggttcgc	apolipoprotein B (APOB)	
ctccgggctttccagccccaaccATGCATAAAaggggttcgc	growth hormone gene-1 (GH1)	Pit-1, GHF-1, Pit-1b, pituitary-specific positive transcription factor 1
ctccgggcTTTCCAGCCccaaccatgcataaaaggggttcgc	interleukin-3 (IL3)	(Jun)2, AP1, Fos/Jun, PEA1 (mouse)
ctccgggcTTTCCAGCCccaaccatgcataaaaggggttcgc	transferrin (TF)	
CTCCGGGCTTtccagccccaaccatgcataaaaggggttcgc	lyn (LYN)	(Ig)NF-A, alpha-H1, NF-A1, NF-III, OBP100, Oct-1, oct-B1A, oct-B1B, OTF-1, TRF, Oct-1B
ctccgggctttccagccCCAACCATGcataaaaggggttcgc	alpha1-globin (HBA1)	(Jun)2, AP1, Fos/Jun, PEA1 (mouse)
tctcggagagccacaGAGCCCGGGCC	fibronectin (FN1)	CTF, NF-1, NF-I, TGGCA-binding protein

a crucial molecular function in *immune response*. As shown in Figure 2, the top 7 biological processes are also very important biological processes involved in immune response, and thus validated our hypothesis that co-regulated genes involved in similar biological processes are regulated by a common mechanism.

We used RepeatMasker [41] and CENSOR [42] to detect any repetitive sequences in the module. Both tools align the input sequence with a database of already known repetitive DNA sequences. Both tools were used with their default settings. Both RepeatMasker and CENSOR confirmed that there were no repetitive sequences in or around the module region of both genes.

However, our putative TFBS could be a repetitive sequence that is just not detected by RepeatMasker and CENSOR. Therefore, we performed a genome-wide search of the module sequence in order to observe whether the module is specific to CXCL1 and CXCL2 genes only, and if not, at what other locations it is found in the genome. We were also interested in finding out why CXCL3 was not reported as differentially regulated in the microarray experiment, although it contains the same known NF-kappaB binding site as CXCL1 and CXCL2.

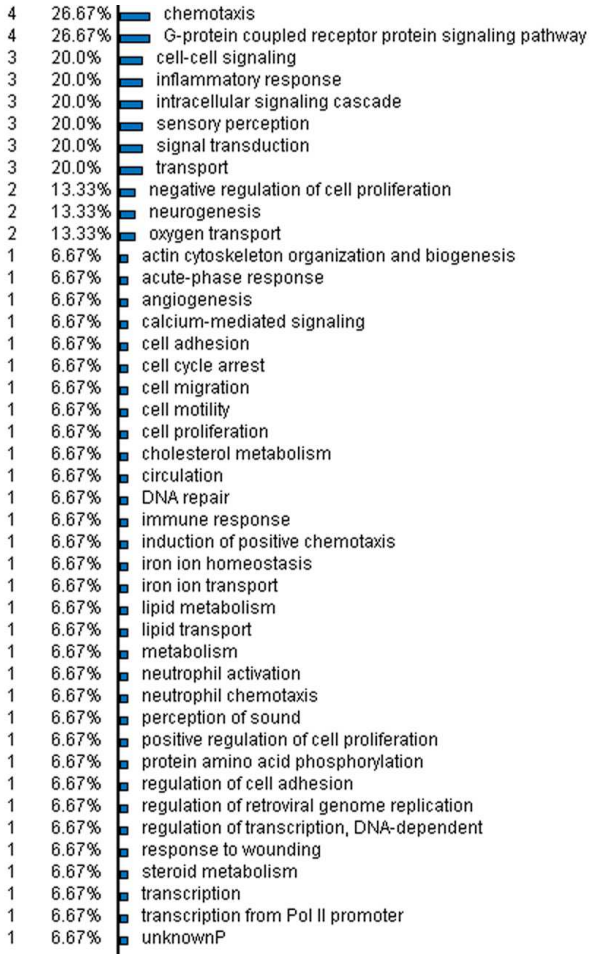

Fig. 2. Functional profiling of the 15 genes found in TRANSFAC with known TFBS. The results showed that the top 7 biological processes (*chemotaxis, G-protein coupled receptor protein signalling pathway, transport, cell-cell signalling, intracellular signalling cascade, inflammatory response*) were clearly related to immune response similar to *cytokine activity*, and involved 12 out of 15 input genes.

We performed the search using the module sequence from gene CXCL2, including the intermediate base pairs between the elements. Using MEGABLAST [43], [44] from NCBI, we found that the sequence is found only on chromosome 4 at 3 different locations as shown in Figure 3. These locations are found immediately upstream of genes CXCL1, CXCL2, and CXCL3. As expected, the module was found upstream of genes CXCL1 and CXCL2 with 93% and 100% identity, respectively. The module found upstream of gene CXCL3 had an 89% identity.

Further analysis of the BLAST alignment results revealed that there were a number of mismatches in CXCL3 compared to CXCL1 and CXCL2 that contain known TFBSs for a number of transcription factors including high mobility group protein I(Y) (HMG I(Y)) as shown in Table I and Figure 3. Thanos et al. showed that both NF-kappaB and HMG I(Y) are required for induction of the human interferon-beta (IFN-beta) gene where HMG I(Y) stimulates the binding of NF-kappaB [45]. This experimental finding combined with our analysis results strongly suggests that CXCL3 is not differentially regulated because of the mismatch in HMG I(Y) binding site. Furthermore, the mismatches in CXCL3 in the binding sites of other transcription factors suggests a possible role played by those factors (see Table I) in the regulation of interferon genes during immortalization that needs to be verified through further experiments. It is very likely that CXCL1 and CXCL2 are regulated by more than one transcription factors, and hence, have more than one TFBSs. However, our analysis indicate that the module discussed here is active in the condition under study. There may be other TFBSs that may be active under different conditions. We are currently analyzing a number of data sets obtained under different biological conditions that will be reported in the future.

IV. Conclusions

In spite of a number of advances in both biotechnology and bioinformatics, gene transcription regulation is only a partially understood phenomenon. In this paper, we propose a bioinformatics technique to identify TFBSs. Our results suggest that the proposed technique can potentially be used for identification of condition-specific transcription factor binding sites for co-regulated genes involved in related biological processes.

Acknowledgment

This work has been supported by the following grants: NIH (NHGRI) 1R01HG003491, NSF DBI-0234806, DOD DAMD 17-03-02-0035, NIH(NCRR) 1S10 RR017857-01, MLSC MEDC-538 and MEDC GR-352, NIH 1R21 CA10074001, 1R21 EB00990-01 and 1R01 NS045207-01.

References

[1] P. Khatri and S. Draghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, vol. 21, no. 18, pp. 3587–3595, 2005.

[2] P. Khatri, S. Drăghici, G. C. Ostermeier, and S. A. Krawetz, "Profiling gene expression using Onto-Express," *Genomics*, vol. 79, no. 2, pp. 266–270, February 2002.

[3] C. J. Krieger, P. Zhang, L. A. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Y. Rhee, and P. D. Karp, "MetaCyc: a multiorganism database of metabolic pathways and enzymes," *Nucleic Acids Research*, vol. 32, no. 1, pp. D438–42, January 2004.

[4] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, *et al.*, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.

[5] K. Dahlquist, N. Salomonis, K. Vranizan, S. Lawlor, and B. Conklin, "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways," *Nature Genetics*, vol. 31, no. 1, pp. 19–20, May 2002.

[6] P. Khatri, S. Sellamuthu, P. Malhotra, K. Amin, A. Done, and S. Drăghici, "Recent additions and improvements to the Onto-Tools," *Nucleic Acids Research*, vol. 33, no. Web server issue, Jul 2005.

[7] M. L. Bulyk, "Computational prediction of transcription-factor binding site locations," *Genome Biology*, vol. 5, no. 1, December 2003.

[8] N. V. Taverner, J. C. Smith, and F. C. Wardle, "Identifying transcriptional targets," *Genome Biology*, vol. 5, no. 3, February 2004.

[9] R. C. Périer, T. Junier, and P. Bucher, "The Eukaryotic Promoter Database EPD," *Nucleic Acids Research*, vol. 26, no. 1, pp. 353–357, 1997.

[10] J. Schug and G. C. Overton, "TESS: Transcription Element Search Software on WWW," CBIL-TR-1997-1001-v0.0 of the Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania, Tech. Rep., 1997.

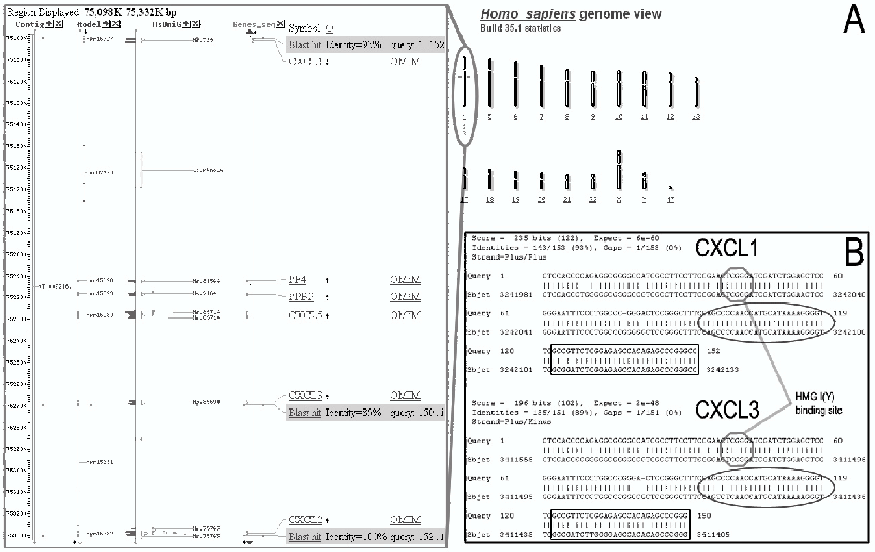

Fig. 3. Genome-wide search of CXCL1 and CXCL2 transcription module sequence. BLAST alignment revealed that the module sequence was only found on chromosome 4 (see panel A). In addition, the alignment also revealed the differences in binding sites of HMG I(Y) and other transcription factors (panel B).

[11] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhuser, M. Pr, F. Schacherer, S. Thiele, and S. Urbach, "The TRANSFAC system on gene expression regulation," *Nucleic Acid Research*, vol. 29, no. 1, pp. 281–283, 2001.

[12] K. Cartharius, K. Frech, K. Grote, B. Klocke, M. Haltmeier, A. Klingenhoff, M. Frisch, M. Bayerlein, and T. Werner, "MatInspector and beyond: promoter analysis based on transcription factor binding sites," *Bioinformatics*, vol. 21, pp. 2933–42, 2005.

[13] R. Elkon, C. Linhart, R. Sharan, R. Shamir, and Y. Shiloh, "Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells," *Genome Research*, vol. 13, no. 5, pp. 773–780, May 2003.

[14] K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner, "MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data," *Nucleic Acids Research*, vol. 23, pp. 4878–84, 1995.

[15] C. Workman and G. Stormo, "ANN-spec: A method for discovering transcription factor binding sites with improved specificity," *Bioinformatics*, vol. 16, no. 1, pp. 16–23, 2000, available online at http://psb.stanford.edu.

[16] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. Menlo Park, California: AAAI press, 1994, pp. 28–36.

[17] E. Segal, M. Shapira, A. Regev, D. Pe'er, S. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nature Genetics*, vol. 34, no. 2, pp. 166–76, June 2003.

[18] E. Segal, R. Yelensky, and D. Koller, "Genome-wide discovery of transcriptional modules from DNA sequence and gene expression," *Bioinformatics*, vol. Suppl 1, pp. i273–82, 2003.

[19] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. D. Moor, P. Rouzé, and Y. Moreau, "A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling," *Bioinformatics*, vol. 17, no. 12, pp. 1113–1122, 2001.

[20] X. S. Liu, D. L. Brutlag, and J. S. Liu, "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes," *Proc. Pacific Symposium on Biocomputing*, pp. 127–38, 2001.

[21] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church, "Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation," *Nature Biotechnology*, vol. 16, no. 10, pp. 939–45, October 1998.

[22] G. Thijs, K. Marchal, M. Lescot, S. Rombauts, B. D. Moor, P. Rouz, and Y. Moreau, "A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes," *Journal of Computational Biology*, vol. 9, no. 2, pp. 447–464, April 2002.

[23] R. Harrison and C. DeLisi, "Condition specific transcription factor binding site characterization in *Saccharomyces cerevisiae*," *Bioinformatics*, vol. 18, no. 10, pp. 1289–1296, 2002.

[24] H. Motohashia, K. Igarashi, K. Onodera, S. Takahashi, H. Ohtani, M. Nakafuku, M. Nishizawa, J. D. Engel, and M. Yamamoto, "Mesodermal- vs. neuronal-specific expression of MafK is elicited by different promoters," *Genes to Cells*, vol. 1, no. 2, pp. 223–38, 1996.

[25] S. Fessele, H. Maier, C. Zischek, P. J. Nelson, and T. Werner, "Regulatory context is a crucial part of gene function," *TRENDS in Genetics*, vol. 18, no. 2, pp. 60–63, February 2002.

[26] A. E. Kel, O. V. Kel-Margoulis, P. J. Farnham, S. M. Bartley, E. Wingender, and M. Q. Zhang, "Computer-assisted identification of cell cycle-related genes: New targets for E2F transcription factors," *Journal of Molecular Biology*, vol. 309, no. 1, pp. 99–120, May 25 2001.

[27] F. Long, H. Liu, C. Hahn, P. Sumazin, M. Q. Zhang, and A. Zilberstein, "Genome-wide prediction and analysis of function-specific transcription factor binding sites," *In Silico Biology*, vol. 4, no. 0033, 2004.

[28] P. Sudarsanam, Y. Pilpel, and G. M. Church, "Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in saccharomyces cerevisiae," *Genome Research*, vol. 12, no. 11, pp. 1723–31, November 2002.

[29] M. Ashburner, "On the representation of gene function in genetic databases," in *Proc. of ISMB, Montreal*, 1998.

[30] S. T. Smale, "Core promoters: active contributors to combinatorial gene expression." *Genes and Development*, vol. 15, pp. 2503–2508, 2001, a good paper to start with.

[31] L. Snyder and W. Champness, *Molecular Genetics of Bacteria*, 2nd ed. American Society for Microbiology Press, 2003.

[32] T. Werner, S. Fessele, H. Maier, and P. J. Nelson, "Computer modeling of promoter organization as a tool to study transcriptional coregulation," *The FASEB Journal*, vol. 17, no. 10, pp. 1228–37, 2003.

[33] O. I. Kulaeva, S. Drăghici, L. Tang, J. M. Kraniak, S. J. Land, and M. A. Tainsky, "Epigenetic silencing of multiple interferon pathway genes after cellular immortalization," *Oncogene*, vol. 22, no. 26, pp. 4118–4127, June 2003.

[34] S. Drăghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz, "Global functional profiling of gene expression," *Genomics*, vol. 81, no. 2, pp. 98–104, February 2003.

[35] S. Drăghici, P. Khatri, P. Bhavsar, A. Shah, S. A. Krawetz, and M. A. Tainsky, "Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3775–81, July 2003.

[36] P. Khatri, P. Bhavsar, G. Bawa, and S. Drăghici, "Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments," *Nucleic Acids Research*, vol. 32, pp. W449–56, Jul 2004.

[37] M. Ashburner *et al.*, "Gene Ontology: Tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, May 2000.

[38] N. Mukaida, Y. Mahe, and K. Matsushima, "Cooperative interaction of nuclear factor-kappaB and cis-regulatory enhancer binding protein-like factor binding elements in activating the interleukin-8 gene by pro-inflammatory cytokines," *The Journal of Biological Chemistry*, vol. 265, no. 34, pp. 21 128–33, December 1990.

[39] B. Stein and A. S. Baldwin, Jr., "Distinct mechanisms for regulation of the interleukin-8 gene involve synergism and cooperativity between C/EBP and NF-kappa B," *Molecular and Cellular Biology*, vol. 13, no. 11, pp. 7191–98, November 1993.

[40] A. Anisowicz, M. Messineo, S. W. Lee, and R. Sager, "An NF-kappa B-like transcription factor mediates IL-1/TNF-alpha induction of gro in human fibroblasts," *The Journal of Immunology*, vol. 147, no. 2, pp. 520–7, July 15 1991.

[41] A. Smit, R. Hubley, and P. Green, "Repeatmasker open-3.0," http://www.repeatmasker.org, 1996-2005.

[42] J. Jurka, P. Klonowski, V. Dagman, and P. Pelton, "CENSOR-a program for identification and elimination of repetitive elements from DNA sequences," *Computers and Chemistry*, vol. 20, no. 1, pp. 119–121, March 1996.

[43] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–10, 1990.

[44] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, "A greedy algorithm for aligning DNA sequences," *Journal of Computational Biology*, vol. 7, no. 1-2, pp. 203–14, 2000.

[45] D. Thanos and T. Maniatis, "The high mobility group protein HMG I(Y) is required for NF-kappa B-dependent virus induction of the human IFN-beta gene," *Cell*, vol. 71, no. 5, pp. 777–789, 1992.