Identification of Transcriptional Regulatory Elements by Functional Enrichment Analysis.

Amitava Karmaker[1]*, Stephen Kwek[2]

[1]University of Wisconsin-Stout, Menomonie, Wisconsin 54751, USA
[2]Microsoft Corporation, Redmond, Washington 98052, USA

Abstract - *Deciphering the complex interaction between transcriptional regulatory (both trans- and cis-) elements comprehensively and identifying these potential binding sites are fundamental problems in functional genomics. Therefore, determining the transcription factors that regulate a gene in different cell types and the cis-regulatory elements they are binding to will help lay the foundation for building gene regulatory networks. While many computational approaches have been developed for lower eukaryotes and prokaryotes, most of them often do not generalize to vertebrates. Here, we use gene ontological evidences to perform functional enrichment analysis among the TFs and genes, and group the functionally related genes to characterize their transcriptional association. We also analyze correlations between TFs and genes using their expression profiles. Thus, we search for putative transcriptional regulatory elements (transcription factor binding sites) along core promoter regions of the grouped genes. The performance of our search is highly satisfactory in term of binding site hit accuracy.*

Keywords: Transcriptional Regulatory Elements, Functional Enrichment, Gene Ontology, Gene Expression Profiles, Microarray Analysis.

1 Introduction

With the completion of draft sequencing of genomes of various species (a.k.a. human, mouse, rat, yeast etc.), one of the objectives of functional genomics is to interpret biological significance of the sequences, and to delineate the functional modules along the genomes. Although a large number of genes have been identified, their regulatory mechanism remains mostly unknown at the transcriptional level[1]. To understand the complex interaction of gene regulation comprehensively, we need to identify the regulatory elements in the human genome and comprehend how the genes regulate and interact with each other.

Simply put, the interaction between transcription factor (TF, a.k.a. *trans*-elements) and transcription factor binding sites (TFBS, a.k.a. *cis*-elements) plays a crucial role in controlling gene expression. To modulate transcription and consequently to control the expression of genes, transcription factor proteins bind to binding sites in the promoter regions and thus either facilitate or inhibit the gene expression. To some extent, the pattern of expression of each gene can be formulated as a function of specific transcription factors, and their binding to the *cis*-elements. So, transcription factors constitute one of the major components in constructing gene regulatory networks. Literally, *trans*-elements can be viewed as “keys” needed to unlock the *cis*-elements which act as “locks”. To comprehend gene transcription mechanism, it is not sufficient to know which keys (*trans*-elements) are needed to lock/unlock a specific gene, but we also need to identify their corresponding locks (*cis*-elements).

Since the human genome sequences are available, quite a number of computational approaches have been developed to discover functional elements in lower prokaryotes by combining genome sequence data and expression profiles[2]. But, due to more degenerate nature and complex interactions of TFs in the multi-cellular mammals (higher eukaryotes), most of the techniques are not able to generalize to mammal genomes. Moreover, these computational techniques are fallible to high false positive prediction rate[3]. In reality, this unusually high false prediction sometimes overwhelms the prospective techniques to deter finding regulatory regions accurately. On the other hand, comparative genome analysis, which is a biologically more relevant approach, provides a powerful way to search for similarities across the species at the sequence level and consequently to assign functional annotations[4]. Besides this, it is assumed that genes with similar functions are most likely to be regulated through the same mechanisms[5]. Thus, we can infer transcriptional sub-networks based on functional enrichment of genes.

In this paper, we propose a systematic technique to identify putative transcriptional regulatory elements in human genome by functional enrichment of genes using ontology. Our hypothesis is inspired by the axiomatic supposition that genes that are in the same functional complex and located in closer cellular proximity are often regulated by the same transcription factors[6]. In fact, two proteins, sharing same molecular function in alike biological process and residing in close physical location, are more likely to interact with each other[7]. Therefore, clustering the genes set using functional enrichment allows us search for *cis*-modules along the

*Corresponding author

promoter regions of the genes more efficiently. Initially, we analyze the correlations among the genes and corresponding TFs using microarray expression data. Besides this, we used the popular gene ontology to come up with the enrichment analysis of the genes. In fact, functional enrichment analysis complements the findings for correlations from expression profiles. To evaluate the efficacy of our approach, we validated our prediction for the transcription factor binding sites from functionally enriched gene clusters by comparing with TRANSFAC[8].

2 Related works

In silico discovery[9] of binding sites is quite effective for prokaryotes, like *Escherichia coli[10]*, where genomes are more compact with many genes being regulated by a single operon, is relatively easy to locate. Similar successes have been reported for simple unicellular eukaryotes, like *Saccharomyces cerevisiae*[2]. The main approach for finding *cis*-elements of such simple organisms is to find overrepresented motifs modeled by known background profiles, such as position weighted matrices (PWMs)[11], position specific score matrices (PSSMs)[12], while some use clustering to demarcate *cis*-regulatory modules[13, 14].

For higher multi-cellular eukaryotes, model-based approaches[1, 15] that discover patterns among co-expressed genes with respect to regulating transcription factors have been proposed. The idea behind these techniques involves the proximity of common *cis*-regulatory modules among the co-expressed genes. Among other common model-based (a.k.a. machine learning) techniques, artificial neural networks[16], greedy algorithm[17], Gibbs Sampling[18], Markov chains[19], Expectation Maximization (EM) algorithm[20] are widely used for eukaryotes. However, it has been reported that these model-prediction techniques are susceptible to high false positive prediction rate and majority of predicted TFBS generated with predictive models (*in silico*) have no functional role *in vivo* [21].

Jin et al.[22] analyzed conserved human-mouse orthologous gene pairs to find core promoter elements and Bussemaker et al.[23] addressed the issue of detecting regulatory elements using correlation of expressions. A recent paper by Kim et al.[24] dealt with predicting transcriptional regulatory elements of human promoters using gene expression and promoter analysis data, which compare two pools of genes using z-scores.

3 Methods and materials

3.1 Data preprocessing

We collected publicly available microarray data of normal human tissues[25], which provide us with 26,260 unique genes from 35 different organs. In total, the data set consists of 115 tissue specimens. For each experimental tissue sample, Cy5- and Cy3- labeled samples were co-hybridized to a cDNA microarray containing 39,711 human cDNA's, representing 26,260 different genes [26]. Expression ratios were globally normalized by mean-centering each gene across all arrays.

3.2 Calculation of correlation co-efficient

If a transcription factor does regulate a gene, according to reported results[15] in the literature, it is expected that they are linearly correlated. However, we observed that very often there seems to be a saturation point where the effect on the expression level of the gene diminishes as the level of transcription factor continues to increase and may reach a plateau or even decrease in some cases. Thus, instead of using simple linear correlation, we measure the correlation using Equation 1 as our regression curve.

$$y' = ye^{\alpha y}, \qquad (1)$$

Where α is an exponential constant

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \qquad (2)$$

Where, ***n*** is the sample size, and ***x_i*** and ***y_i*** are the sum of **X** and **Y** from ***i = 1 to n***

In Equation (1), y is the original expression level, and it is multiplied by some exponential constant to generate new values. The value of parameter α was set to 0.25. This correlation coefficient is more general than simple linear correlation coefficient. By setting α = 0.0, we end up with the simple linear correlation coefficient. We calculated Pearson's Correlation Coefficient (Equation (2)) of all pairs of gene and TF. The correlation coefficients indicate how tightly genes are up-regulated and down-regulated with respect to transcription factors. The values of Pearson's Correlation Coefficient range from -1 to +1. Any value in positive scale indicates increasing correlationship, with +1 being perfectly linear correlated and negative values denote the case of a negative correlationship. Any value in between in all other cases represents the degree of dependency between the variables (i.e. gene and TF pair).

3.3 Gene Ontology

Genome-wide comparison has revealed that a large fraction of genes encoding the core biological processes and molecular functions are shared by all the eukaryotes, with a few exceptions[27]. In fact, comprehensive knowledge about biological roles of common gene products in diverse species can obviously explain, and often provide strong implication of, its function in the like genomes. However, due to divergent nomenclatures and interpretations of biological elements, it has been difficult for the researchers to talk in common language. To address this issue, the Gene Ontology (GO) Consortium[28] has been formed. Basically, Gene Ontology (GO) provides a great resource for describing gene products by standardizing biological concepts and by consolidating gene annotation information from heterogeneous data sources in a consistent manner. As a mainstay standard for facilitating annotation of gene products, it has been successfully used in unraveling protein-protein interactions and classifications in genomes, such as *Homo sapiens*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, *Arabidopsis thaliana*.

Gene Ontology (GO) Consortium has developed a database consisting of standardized, structured, dynamically controlled vocabularies (ontological) to encode various aspects of gene products in organisms[28]. The Gene Ontology (GO) is categorized into three orthogonal entities: (1) molecular function (MF) describes the role of a gene product in molecular level; (2) biological process (BP) outlines the processes (objectives) the gene products partake in; (3) cellular component (CC) refers to the cellular localization of the proteins where they are active. Each GO is represented as a directed acyclic graph (DAG), in which each term is either a child of one or multiple parents ("is-a" relationship) or a constituent instance ("part-of" relationship) of the parent terms. In the graph, the nodes correspond to the GO terms, while edges denote the relationships among the terms. Depending on the depth (level) of a node, we can determine the specificity of the term. The closer to the root a term is, the more general the term is. Conversely, if it is located in the leaf levels, the term is the most specific with respect to that particular ontology.

3.4 Functional enrichment measure

Although semantic similarity based methods are popular in assessing functional similarity among the gene products, there are a number of drawbacks we need to consider. First of all, different methods treat the commonality (a.k.a. generality and specificity) of nearest common ancestors in different ways. Secondly, in the GO graph, the depth of terms does not actually signify the specificity of the corresponding concepts. Different terms in the same rank (depth) are necessarily not equally specific. Finally, as the GO is a continuing project where new vocabularies are constantly added (updated), therefore very often the similarity measures are subject to change.

Regarding all these issues, we attempt to define a similarity metric based on assigned GO terms to a gene product instead of concerning much about frequently changing GO semantic structure. Again, as we are interested in clustering functionally related genes on the basis of their GO terms, our distance measure provides straightforward approach to group them together. The idea behind our metric definition is that the more genes have common (general) GO terms, and the less they have specific GO terms, the more likely they tend to be functionally related. Our distance measure is based on the Czekanowski-Dice formula (see Equation 3).

Let two sets of GO terms of annotated genes $G1$ and $G2$ be $GO_1 = \{go_{11}, go_{12}, go_{13}, ..., go_{1m}\}$ and $GO_2 = \{go_{21}, go_{22}, go_{23}, ..., go_{2n}\}$ in order. According to our algorithm, the distance measure between G1 and G2:

$$D(G1, G2) = \frac{|GO_1 - GO_2| + |GO_2 - GO_1|}{|GO_1 \cup GO_2| + |GO_1 \cap GO_2|} \quad (3)$$

The closer the genes are in respect with biological function, the lesser the distance measure ($D(G1, G2)$) is. In our analysis, we label this distance measure as functional enrichment score. This distance formula weighs more on the significance of the common GO terms by giving more emphasis to similarities than to dissimilarities. Thus, if two gene products do not share any GO terms, the distance value would be one (1), the highest possible value, while for two gene products sharing exactly the identical set of GO terms, the distance value is zero (0), which is the lowest possible value.

3.5 Finding *cis*-regulatory elements

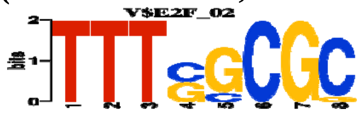
To determine the (putative) *cis*-regulatory elements, we identify associated genes with certain TF with correlation co-efficient greater that a threshold (>0.5). Using functional enrichment analysis, we construct cluster of genes that are functionally related to certain transcription factor. After calculating the distance measures (functional enrichment scores) of the respective TF against rest of the genes, we sort them by enrichment score in ascending order (genes with less score at the top). For further analysis, we selected top ten genes from this list, which include genes that are functionally enriched with corresponding TF (enrichment score < 1.0) with moderately high correlation coefficient (~>0.60).

Transcriptional regulatory elements are found either upstream or downstream of genes, scattered all along thousands of bps in both intergenic and intragenic regions. However, most TFBS predictors tend to focus in the proximal promoter region[3] because the difficulty of TFBS prediction tends to increase with the size of the region of interest. Besides, increasing the region of interest upstream of the transcription start site to more than a few thousand base pairs increases the chances of falsely identifying common repeat elements. This, we focus on the core promoter regions from 1500 bps upstream to 500 bps downstream (-1500 to +500, total 2000 bps) and extracted the nucleotide sequences for the genes as FASTA format.

To ensure that our putative TF binding sites are of high quality, we validated them with TRANSFAC database[15], which is the largest repository for experimentally derived (validated) TFBS. We also performed further corroboration of our putative sites using P-Match[29]-public (which is a TRANSFAC subsidiary) and ConSite[30], which combines pattern matching and weight matrix approaches thus providing higher accuracy of recognition than each of the methods alone. To reduce false-positive validation using P-match, we chose "high quality vertebrate matrices only" as our default option. We obtained the report for all pre-selected genes, setting cut-off selection for matrices to minimize (1) false-positive, (2) false-negative, and (3) the sum of both error rates. Moreover, ConSite[30] is a user-friendly, web-based tool for finding cis-regulatory elements in genomic sequences using high-quality transcription factor models and cross-species comparison filtering.

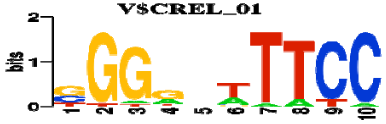
Table 1: The list of identified binding sites for *E2F5* and *RELB* TFs. Results were validated using both TRANSFAC and ConSite.

***E2F5* (TRANSFAC: E2F, ConSite: E2F)**



Genes	Correlation Coefficient	Functional enrichment score	Position in sequence (strand)	Consensus sequence
MBD4	0.82118	0.76	942 (+)	**TTTGCcgc**
DCK	0.79218	0.904	1496 (-)	**gcgCCAAA**
MCM6	0.78034	0.629	1347 (+)	**TTTGGcgc**
MYBL1	0.76635	0.538	N/A	**N/A**
DR1	0.76331	0.578	N/A	**N/A**
LSM6	0.75098	0.739	1755 (-)	**ccgCGAAA**
EZH2	0.74767	0.583	1533 (+)	**TTTGGcgc**
PCNA	0.73964	0.769	1442 (-)	**gcgGGAAA**
HMGB2	0.69681	0.75	336 (+)	**TTTGGcgc**
NMI	0.61465	0.733	1553 (+)	**TTTCGcgg**

***RELB* (TRANSFAC: c-REL, ConSite: c-Rel)**



Genes	Correlation Coefficient	Functional enrichment score	Position in sequence (strand)	Consensus sequence
PSMB9	0.91903	0.833	1107 (-)	**GGAAAgtccc**
COX7B	0.80250	0.76	N/A	**N/A**
ZFP106	0.76718	0.913	1343 (-)	**GGAATcctca**
ARHGAP5	0.76682	0.909	1884 (+)	**gggtgCTTTC**
NFE2L1	0.74318	0.619	641 (-)	**GAAACatccc**
MAPKAPK3	0.73573	0.904	197 (-)	**TGTAGcaccc**
RYR2	0.72470	0.8	549 (-)	**GGAATgctcg**
DNAJB6	0.71556	0.809	137 (+)	**gggatTTTTC**
ARF1	0.71359	0.933	256 (+)	**ggggcTTTCC**
IRF2	0.70548	0.474	1468 (+)	**ggggaTTTCC**

4 Results and Discussion

As a case study, we selected *E2F5* and *RELB* for our candidate TF. We screened out genes that are functionally enriched with these TFs. In order to quantify the regulatory elements along these gene sequences, the core promoter regions (see Methods) were fed to P-Match[29] using all three available options for handling false discoveries. Basically, the output with option "minimizing false negative" considers merely minimal number of base pairs match and calls it a hit. Thus it improves its recall numbers (maximize loose-bound relevance at the cost of precision), with a huge list of *cis*-element candidates. We expect the false-positive rate to be extremely high for the predictions to be meaningful. Therefore, we did not discard this option. Among the other options, "minimize false positive" tries to find exact (~100%) PWM match and accounts for the most precise TF hits. The other option "minimize sum of both error rates" seems to take advantage from the best of both worlds (keeping balance on both recall and precision) and evens out high false discovery rates. To ensure better quality of our analysis, we considered only the option "minimize false positive", which maximizes the precision values without compromising too much with recall values. We summarize the sample results for *E2F5* and *RELB* genes in Table 1. The results for consulting ConSite are furnished as well. The consensus sequences (Logo-plots[31]) for respective TFBS were extracted from TFM-Explorer[11].

Our predictions for *cis*-elements for these two TFs are highly accurate. Out of the ten human genes that are associated with *E2F5* (E2F transcription factor 5), a member of *E2F* TF family, eight genes (80% hit rate) carry the supposed binding sites precisely (negative strands are give as reverse complemented. Comparing the sequence patterns of binding sites, we can say that almost all of them share the consensus sequence **'TTTSSCGC'** where S could be a C or G. Likewise, for the ten human genes functionally correlated with TF *RELB*, we have found nine genes have the consensus sequence for *RELB* binding sites, which achieves a hit rate of ~90%. Here, we found "**TTTCC"** as sense (+), or "**GGAAA"** as anti-sense (-) complementary, to be common motif with a number of out of pattern nucleotides around.

.

5 Conclusions

In short, we propose a computational method to identify putative transcriptional regulatory elements by analyzing functional enrichment using gene ontology. Although there are a lot of computational techniques for this purpose, it is not possible to extend those from motif finding in lower prokaryotes to that in mammals. These techniques also tend to show higher false discovery rates. We demonstrated that the use of our similarity (distance) metric can group genes based on enrichment score and it strengthens the findings from gene expression profile analysis. In each group genes are functionally related to the corresponding TFs; so searching for functional modules along the promoters of genes is more appropriate for capturing possible regulatory relationship. Finally, we validate our prediction for *cis*-regulatory motifs in both genomes using TRANSFAC. As a possible further step to confirm the regulatory relationships, the TF-gene pairs and their functional enrichment constructed here may serve as a reference of additional evidence for ChIP-chip results.

6 References

[1] J. W. Fickett and W. W. Wasserman, "Discovery and modeling of transcriptional regulatory regions," *Curr Opin Biotechnol,* vol. 11, pp. 19-24, Feb 2000.

[2] M. Tompa*, et al.*, "Assessing computational tools for the discovery of transcription factor binding sites," *Nat Biotechnol,* vol. 23, pp. 137-44, Jan 2005.

[3] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nat Rev Genet,* vol. 5, pp. 276-87, Apr 2004.

[4] R. Stevens*, et al.*, "Ontology-based knowledge representation for bioinformatics," *Brief Bioinform,* vol. 1, pp. 398-414, Nov 2000.

[5] D. J. Allocco*, et al.*, "Quantifying the relationship between co-expression, co-regulation and gene function," *BMC Bioinformatics,* vol. 5, p. 18, Feb 25 2004.

[6] W. K. Huh*, et al.*, "Global analysis of protein localization in budding yeast," *Nature,* vol. 425, pp. 686-91, Oct 16 2003.

[7] X. Wu*, et al.*, "Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations," *Nucleic Acids Res,* vol. 34, pp. 2137-50, 2006.

[8] V. Matys*, et al.*, "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res,* vol. 31, pp. 374-8, Jan 1 2003.

[9] G. Pavesi*, et al.*, "In silico representation and discovery of transcription factor binding sites," *Brief Bioinform,* vol. 5, pp. 217-36, Sep 2004.

[10] L. A. McCue*, et al.*, "Factors influencing the identification of transcription factor binding sites by cross-species comparison," *Genome Res,* vol. 12, pp. 1523-32, Oct 2002.

[11] M. Defrance and H. Touzet, "Predicting transcription factor binding sites using local over-representation and comparative genomics," *BMC Bioinformatics,* vol. 7, p. 396, 2006.

[12] P. E. Boardman*, et al.*, "SiteSeer: Visualisation and analysis of transcription factor binding sites in nucleotide sequences," *Nucleic Acids Res,* vol. 31, pp. 3572-5, Jul 1 2003.

[13] M. C. Frith*, et al.*, "Cluster-Buster: Finding dense clusters of motifs in DNA sequences," *Nucleic Acids Res,* vol. 31, pp. 3666-8, Jul 1 2003.

[14] N. Rajewsky*, et al.*, "Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo," *BMC Bioinformatics,* vol. 3, p. 30, Oct 24 2002.

[15] E. M. Conlon, *et al.*, "Integrating regulatory motif discovery and genome-wide expression analysis," *Proc Natl Acad Sci U S A,* vol. 100, pp. 3339-44, Mar 18 2003.

[16] C. T. Workman and G. D. Stormo, "ANN-Spec: a method for discovering transcription factor binding sites with improved specificity," *Pac Symp Biocomput,* pp. 467-78, 2000.

[17] G. Z. Hertz and G. D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences," *Bioinformatics,* vol. 15, pp. 563-77, Jul-Aug 1999.

[18] M. C. Frith, *et al.*, "Finding functional sequence elements by multiple local alignment," *Nucleic Acids Res,* vol. 32, pp. 189-200, 2004.

[19] A. V. Favorov, *et al.*, "A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length," *Bioinformatics,* vol. 21, pp. 2240-5, May 15 2005.

[20] W. Ao, *et al.*, "Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR," *Science,* vol. 305, pp. 1743-6, Sep 17 2004.

[21] W. B. Alkema, *et al.*, "MSCAN: identification of functional clusters of transcription factor binding sites," *Nucleic Acids Res,* vol. 32, pp. W195-8, Jul 1 2004.

[22] V. X. Jin, *et al.*, "Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs," *BMC Bioinformatics,* vol. 7, p. 114, 2006.

[23] H. J. Bussemaker, *et al.*, "Regulatory element detection using correlation with expression," *Nat Genet,* vol. 27, pp. 167-71, Feb 2001.

[24] S. Y. Kim and Y. Kim, "Genome-wide prediction of transcriptional regulatory elements of human promoters using gene expression and promoter analysis data," *BMC Bioinformatics,* vol. 7, p. 330, 2006.

[25] R. Shyamsundar, *et al.*, "A DNA microarray survey of gene expression in normal human tissues," *Genome Biol,* vol. 6, p. R22, 2005.

[26] D. J. Maron, *et al.*, "Gene therapy of metastatic disease: progress and prospects," *Surg Oncol Clin N Am,* vol. 10, pp. 449-60, xi, Apr 2001.

[27] D. Devos and A. Valencia, "Intrinsic errors in genome annotation," *Trends Genet,* vol. 17, pp. 429-31, Aug 2001.

[28] M. Ashburner, *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet,* vol. 25, pp. 25-9, May 2000.

[29] D. S. Chekmenev, *et al.*, "P-Match: transcription factor binding site search by combining patterns and weight matrices," *Nucleic Acids Res,* vol. 33, pp. W432-7, Jul 1 2005.

[30] A. Sandelin, *et al.*, "ConSite: web-based prediction of regulatory elements using cross-species comparison," *Nucleic Acids Res,* vol. 32, pp. W249-52, Jul 1 2004.

[31] J. L. DeRisi, *et al.*, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science,* vol. 278, pp. 680-6, Oct 24 1997.

[32] T. L. Ferea, *et al.*, "Systematic changes in gene expression patterns following adaptive evolution in yeast," *Proc Natl Acad Sci U S A,* vol. 96, pp. 9721-6, Aug 17 1999.

[33] N. Ogawa, *et al.*, "New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis," *Mol Biol Cell,* vol. 11, pp. 4309-21, Dec 2000.

[34] P. T. Spellman, *et al.*, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Mol Biol Cell,* vol. 9, pp. 3273-97, Dec 1998.

[35] A. P. Dempster, *et al.*, "Maximum-likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society,* vol. B39, pp. 1-38, 1977.

[36] M. Ouyang, *et al.*, "Gaussian mixture clustering and imputation of microarray data," *Bioinformatics,* vol. 20, pp. 917-23, Apr 12 2004.

[37] E. Frank, *et al.*, "Data mining in bioinformatics using Weka," *Bioinformatics,* vol. 20, pp. 2479-81, Oct 12 2004.

[38] R. Jornsten, *et al.*, "DNA microarray data imputation and significance analysis of differential expression," *Bioinformatics,* vol. 21, pp. 4155-61, Nov 15 2005.

[39] R. Jornsten, *et al.*, "A meta-data based method for DNA microarray imputation," *BMC Bioinformatics,* vol. 8, p. 109, 2007.