

# Comparisons of Predicted Genetic Modules: Identification of Co-Expressed Genes through Module Gene Flow

**Boris E Shakhnovich**  
borya@bu.edu

**Timothy E Reddy**  
treddy@bu.edu

**Kevin Galinsky**  
galinsky@bu.edu

**Joseph Mellor**  
mellor@bu.edu

**Charles Delisi**  
delisi@bu.edu

Boston University Bioinformatics Program, Boston University, Boston, MA 02215,  
USA

## Abstract

A question of fundamental importance is the definition and identification of modules from microarray experiments. A wide variety of techniques have been used to gain insight into the elucidation of such modules. One problem, however, is the inability to directly compare results between the different data sets produced due to the inherent parameterizations of their approaches. We first aim to provide a mechanism by which different approaches to module finding can be directly compared. Moreover, the same approach can be used to internally compare the modules predicted by the same technique, but at different parameterizations. We apply this approach to analyze the flow of genes through modules at different module thresholds of the Barkai Signature method, thereby further resolving the modules into sets of co-expressed genes.

**Keywords:** genetic module, gene expression, module gene flow, module identification

## 1 Introduction

The identification of genes that belong to co-expressed modules has become the cornerstone of microarray research [4, 6]. The identification of modules has enabled researchers to infer important insight into the internal “wiring” of genetic networks [7, 9]. The inclusion characteristics used to assemble genes into modules are diverse and based on many “distance” metrics and algorithms that are often not readily comparable. Most authors compare their published results without consideration of the parameters inherent in either their algorithms or parameters inherent in the algorithms to which the comparison is made. Consequently, there are two issues that are seldom addressed by the authors of algorithms aimed at identification of these modules. First, most authors publish modules that are based on some threshold parameter, however most often they only publish data for only the “best one” of these thresholds. There is little discussion of how modules change with respect to the parameters inherent in the algorithm. Second, the majority of the algorithms that deal with microarray data identify only the sets of genes that belong to the same module. These algorithms do not provide any insight of whether these modules can be further broken down into parts. Consequently, this formulation prevents definition of an “expression window” where a particular gene or set of genes is most prevalent, or decision of whether the module is really a combination of two or more subsets of modules. In this paper we attempt to solve both of these problems by coherent evaluation and comparison of module algorithms between each other and within themselves with respect to internal parameters.

## 2 How Algorithms Find Modules

We start by comparing three popular module detection algorithms – GRAM proposed by Bar-Joseph *et al.* [2], GeneXpress [10] proposed by Segal *et al.* and the Signature algorithm proposed by Barkai *et al.* [3, 5]. The GRAM algorithm is based on identification of all possible combination of transcriptional regulators indicated by DNA-binding data [7] with a stringent criterion. Once a set of genes is identified, the algorithm identifies a subset of genes with highly correlated expression profiles using Pearson correlation. The algorithm then revisits the DNA-binding data and adds other genes with a correlated expression profile and similar set of binding transcription factors.

GeneXpress uses a set of commonly known transcription factors to partition the genes into modules with a similar expression profile. The algorithm searches for sets of regulatory genes that explain the behavior of the module in the most coherent manner. This algorithm as well as GRAM depends on prior knowledge of transcription factors. The two also depend on probability cutoffs such as Pearson correlation for GRAM and Bayesian maximum likelihood threshold for GeneXpress. Unfortunately, at the time of writing the authors were able to obtain only a single set of modules from each algorithm corresponding to a best-case scenario.

Finally, the Signature algorithm from Barkai lab uses an iterative procedure where sets of genes are used as seeds that identify sets of experiments where that set of genes are expressed higher than threshold  $T_g$ . That set of experiments is then used to identify a set of genes that are expressed higher than threshold  $T_c$ . The algorithm runs until a stable point is found such that the set of genes contained in the module does not change. This algorithm creates different modules depending on the choice of  $T_g$  that can be intuitively understood by imagining some genes have expression “regimes” where an increase or a decrease in their expression indicates a change in regulation. For example, in amino-acid biosynthesis the HIS genes are controlled by the GCN4 transcription factor which yields a certain expression level. However, another transcription factor BAS1 [8, 11] can act as an enhancer and change the expression level of the HIS genes under certain conditions. Thus, there are two regimes for the HIS genes, the low expression regime where they are controlled by only GCN4 and the high expression regime where they are controlled by both GCN4 and BAS1 [1].

Our above descriptions of various module-prediction methods are only intended to provide a most basic overview of the methods involved; we therefore encourage readers to refer to the original reports for complete descriptions.

## 3 Comparing Modules

The first thing we do is compare the partitioning of the set of genes into modules as identified by each algorithm. We proceed to turn each partitioning of genes into modules into a square matrix where  $i, j$  is 1 if gene  $i$  and gene  $j$  are in the same module and 0 otherwise Eq 1. We build one matrix for the output of each algorithm at each cutoff that we consider. Our goal is then to compare these matrices and quantify the degree of correlation between them.

$$N_T = \begin{pmatrix} a_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{nn} \end{pmatrix} a_{i,j} \rightarrow \begin{cases} 1 & \text{iff domain } i, j \text{ belong to the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

Equation 1: Here  $N_T$  is the matrix built from the partitioning of genes into modules at threshold  $T$ .

The problem of comparison of these matrices is analogous to one of comparing two binary strings (or in our case binary matrices) where each digit represents the existence of the pair of genes in the same module. Let us remind ourselves that if the two domains  $i$  and  $j$  exist in the same cluster the

value of the  $i, j$  element in the matrix is 1 else it is 0 (Equation 1). We can use this notation to define a few important quantities that will enable us to compare two partitionings. A true positive (TP) is when both graphs place the two domains  $i$  and  $j$  into the same cluster i.e. the value  $M_{ij}$  in both matrices is 1. A false positive (FP) and false negative (FN) is when one graph puts the two domains into the same cluster, while the other doesn't. True negative (TN) is when both graphs have the two domains in different clusters. This is summarized in Table 1 and can be calculated using Eq 2 for each pair of partitionings  $N$  and  $M$  at thresholds  $T_1$  and  $T_2$  respectively. In order to calculate the distance between two partitionings we have to compare the number of TP versus FP and FN that the comparison between the two yields.

$$\begin{aligned} TP_{N,M}^{T_1,T_2} &= \sum_{i,j \in \{1,\dots,n\}} N_{ij}^{T_1} \wedge M_{ij}^{T_2} \\ FN_{N,M}^{T_1,T_2} + FP_{N,M}^{T_1,T_2} &= \sum_{i,j \in \{1,\dots,n\}} \left( N_{ij}^{T_1} \vee M_{ij}^{T_2} - N_{ij}^{T_1} \wedge M_{ij}^{T_2} \right) \\ TN_{N,M}^{T_1,T_2} &= |N|^2 - \sum_{i,j \in \{1,\dots,n\}} N_{ij}^{T_1} \vee M_{ij}^{T_2} \end{aligned}$$

Equation 2: Here  $N$  and  $M$  represent the two matrices describing partitioning of two different sets (Equation 1) at  $T_1$  and  $T_2$  that are the two thresholds.

Table 1: A sample truth table that is built for every gene pair. A “1” symbolizes that the domains are in the same cluster while a 0 indicates that they are in different clusters. For every gene pair in common between two graphs we calculate whether the annotation is a TP, FP, FN or TN.

Name	Graph 1	Graph 2
TP	1	1
FP	1	0
FN	0	1
TN	0	0

After the above quantities have been defined, the distance measure between two partitionings is merely a calculation of how many true positives the two share with respect to false negatives and false positives. This measure is meant to calculate the level of agreement between the two sets with respect to how many gene pairs they classify in the same module. Of course this is only the first approximation of distance. A true measure of distance between two partitionings would also search for all combinations of similarly annotated triples, quadruples . . . n-tuples of genes. Since that is too computationally expensive to be viable, we satisfy ourselves with only the first order approximation. We pick the simplest measure that compares the TP to FP and FN e.g. Jaccard defined as

$$J = \frac{TP}{FP + FN + TP}$$

Equation 3: The Jaccard Distance.

While there are many measures that can perform a similar task of measuring sensitivity and specificity based distance between two partitionings, such as ROC curves, we have checked that the choice of the particular measure does not affect the conclusions of this paper (data not shown). It is important, however, to note that in contrast to measures used in other comparable studies, this measure is reflexive i.e. it doesn't depend on the direction of comparison. On the other hand, since it counts both true positives and true negatives, the actual quantity behaves as though less than those generated by one-sided comparison measures such as sensitivity and specificity.

Using the above algorithm we compare the partitioning obtained from the Barkai algorithm to the ones obtained by GeneXpress and GRAM. Jaccard distances were calculated between Barkai modules over a range of thresholds and the GRAM and GeneXpress modules (Figure 1, A and B, respectively). Both the GeneXpress and GRAM modules were most highly correlated with the modules from the least stringent Barkai thresholds, and showed a considerable decrease in correlation at increasing Barkai thresholds. By comparing the Jaccard values, we can naively postulate that the GRAM method gives stronger correlation with Signature (.4) than GeneXpress (.2). Comparison of GRAM with GeneXpress yields Jaccard = .14. Thus, even though the maximum overlap between either GRAM and GeneXpress and Signature is large, the correlation between the two algorithms is significantly smaller. However, to be sure, we choose to compare the real Jaccard values obtained in Figure 1 with ones that would be expected at random.

It is important to note that our comparisons do not attempt to evaluate the “correctness” of any given module-finding method - the definition of a correct genetic module is subject to interpretation, and is highly dependent upon the definitions imposed by the researcher. This study is concerned primarily with correspondence of the gene sets inside modules defined by different methods. As such, no claim of accuracy of the individual method is made. That said, the module comparison method we describe could be used by a researcher to determine the module-prediction algorithm and parameters thereof most coherent with a known or ideal gene module indicative of the types of modules important to the researcher. That evaluation would be an essential step in understanding which algorithm and parameterizations are most appropriate to a given line of investigation, allowing the investigator to make an informed choice of module-prediction algorithms for downstream research.

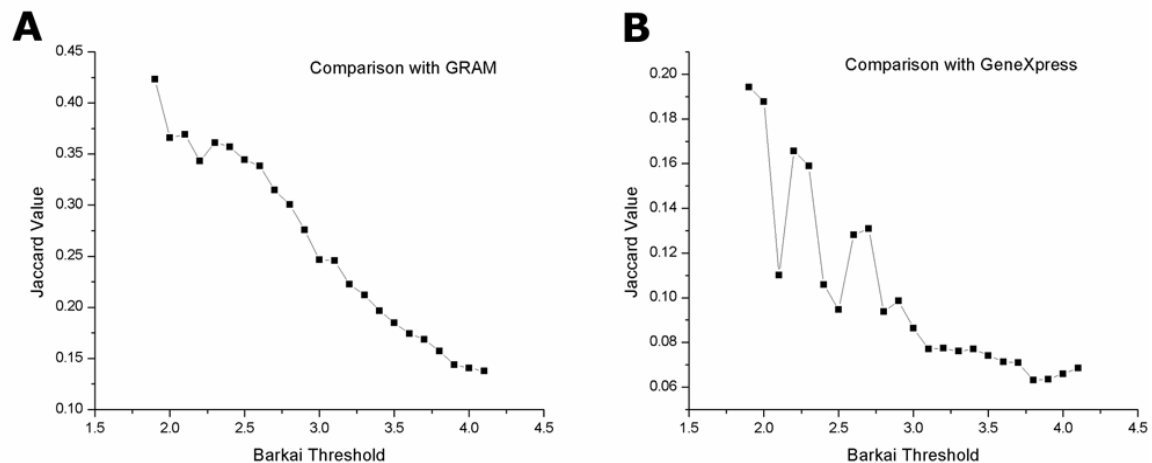


Figure 1: Comparison of modular analysis from the Barkai Signature method at different thresholds to GeneXpress and GRAM. It seems that both GeneXpress and GRAM identify modules most similar to those identified by the Signature method at low expression thresholds.

## 4 Comparison with Random Data

In order to better understand the significance and robustness of the calculated Jaccard distances between the module sets (Figure 1), all distances were normalized against a set of random permutations of the GRAM and GeneXpress modules. Genes were randomly reshuffled among the existing modules, keeping the original modules sizes the same (Figure 2). We calculated Jaccard distances between the each of the reshuffled modules and the original Barkai modules, and obtained a mean and standard deviation for the Jaccard distance between the GRAM and GeneXpress modules and the Barkai modules at each Barkai threshold (Figure 3). Upon comparison with the random permutations it

becomes clear that the naïve observation was incorrect. With respect to expectation of random correlation, the GeneXpress modules are better correlated than the GRAM modules with the Signature data set at low threshold (Figure 3).

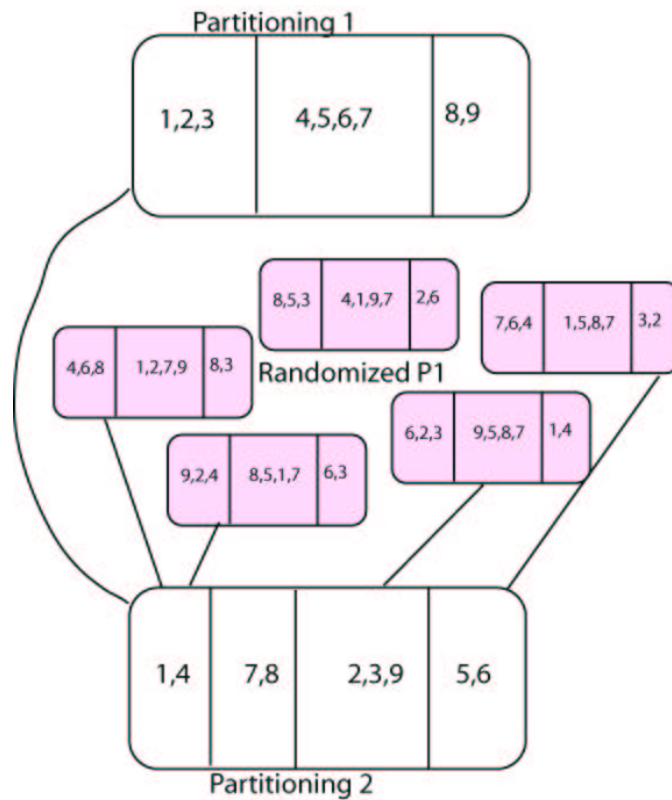


Figure 2: Schematic picture outlining the randomization procedure. We start with two sets of modules Partitioning 1 and Partitioning 2. We then generate a set of random reshufflings P1 by taking Partitioning 1 and redistributing the genes inside randomly, keeping the sizes of the partitions constant. We then compare Partitioning 2 to the original Partitioning 1 and to all the random variants in P1. We calculate the mean and the standard deviation of the Jaccard values by comparing Partitioning 2 with the random variants P1 and compare to the real Jaccard value from comparison of Partitioning 2 with the original Partitioning 1.

From our analysis, we can conclude the GRAM and GeneXpress data sets are most strongly correlated with the lowest Barkai thresholds. It also appears that the modules predicted by the Signature method at increasing thresholds gradually shift away from the modules predicted by the GRAM and GeneXpress methods. The shifting is probably due to the insensitivity of either GRAM or GeneXpress to sub-modularization. We can hypothesize that these algorithms find the most “coherent” modules. These modules often include internal regulation machinery that can further subdivide them into parts.

## 5 Examining Sub-Modularization

In order to better understand the changes in Barkai modules over changing thresholds, we looked at the correlation of the Barkai Signature data set with itself at different thresholds (Figure 4). To do this, we examine the Jaccard overlap between partitionings at different Barkai thresholds  $T_G$ . As expected, we observed a uniformly very strong correlation between identical thresholds, as indicated by the sharp peak along the diagonal in Figure 4. As the difference between the thresholds increases,

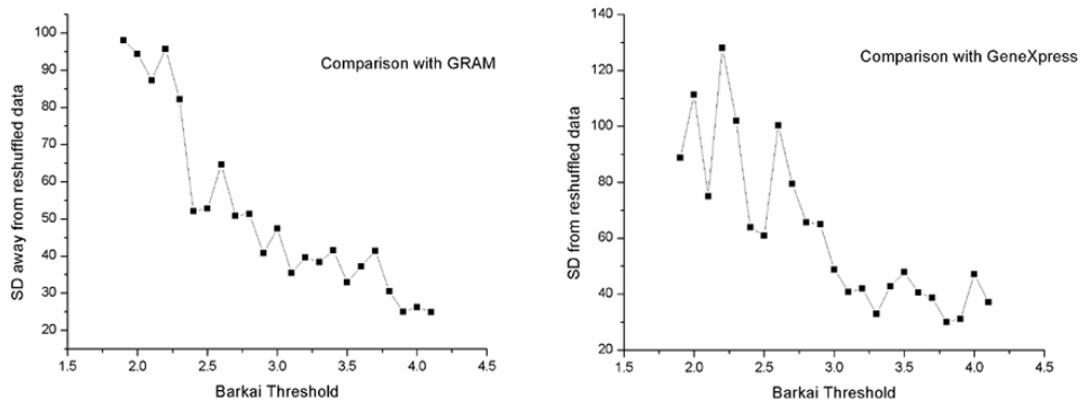


Figure 3: Normalized comparison of GeneXpress and GRAM modules against Barkai Signature method at different Barkai thresholds. In contrast to Figure 1, the GeneXpress modules appear to be much more significantly correlated with modules from low Barkai thresholds than the GRAM data.

however, the modules appear to drift apart from each other, indicating a steady change in overall module composition across threshold values; there appears to be an overall flux of genes through the Barkai modules over increasing threshold values.

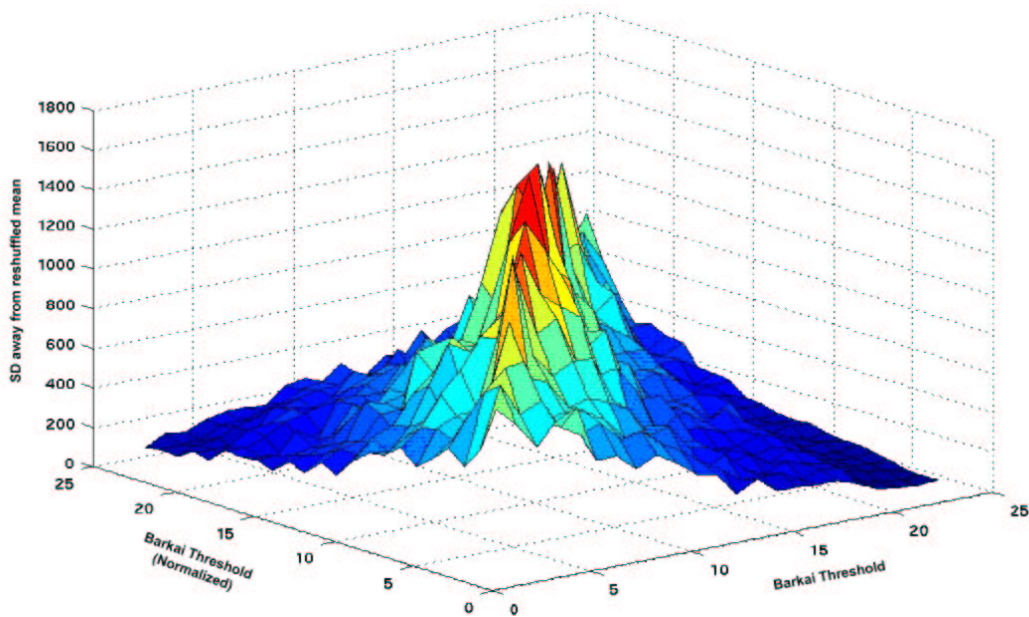


Figure 4: Comparison of Barkai Signature data set with itself at different thresholds using a normalized Jaccard distance metric. For very close thresholds Barkai modules are highly correlated with each other. However, as thresholds begin to move apart, the Barkai modules slowly drift apart until little correlation is observed between the different modules.

## 6 Future Work

Understanding of the gene flux across modules with respect to the thresholds may enable us to divide the global modules into sub-modules. This can be done through careful observation of genes that have different “expression windows” from the main module and identifying sets of genes that spend a large

amount of expression value time with each other. Thus, we have to identify the beginning of the gene expression value which is just the threshold  $T_G$  at which we first observe the gene. The end expression value is the upper bound of  $T_G$  where we can observe the gene “exiting” the module. We also have to calculate the total “expression time” that any two genes spend in the same module. Understanding of these values will enable us to not only define subsets of the genes in each module that correspond to sub-modules subject to different regulation akin to HIS genes but also define the expression “window” where we can most likely observe the expression of the genes in the sub-modules.

As a general trend, a gene  $G$  first enter a module at a given threshold  $T_G^-$ , continues to belong to that module through some number of increasing threshold iterations (i.e. the gene’s expression time), and eventually separates from the module at some higher threshold value  $T_G^+$ . In some instances, a gene’s expression time will to some extent overlap the expression time of another gene in the same module. In this case, we can define a value  $T_{G_i, G_j}$  to be the total amount of expression time shared by gene  $i$  and gene  $j$ . These values can be formally defined as:

$$\begin{aligned}
T_G^- &= N * \delta(B_{N, N+1}^+); N \in \{1, \dots, k\}; \delta(B_{N, N+1}^+) \rightarrow \left\{ \begin{array}{l} 1 \text{ iff } G \in \{B_{N, N+1}^+\} \\ 0 \text{ Otherwise} \end{array} \right\} \\
T_G^+ &= (N + 1) * \delta(B_{N+1, N}^-); N \in \{1, \dots, k\}; \delta(B_{N+1, N}^-) \rightarrow \left\{ \begin{array}{l} 1 \text{ iff } G \in \{B_{N+1, N}^-\} \\ 0 \text{ Otherwise} \end{array} \right\} \\
\sum_G \delta(B_{N, N+1}^+) - \sum_G \delta(B_{N+1, N}^-) &= 0; N \in \{1, \dots, k\} \\
T_{G_i, G_j} &= \sum_{M=1}^i \sum_{N=1}^k (B_N^M) * \delta(G_i^N G_j^N); \delta(G_i^N G_j^N) \rightarrow \left\{ \begin{array}{l} 1 \text{ iff } (G_i, G_j) \in \{B_N^M\} \\ 0 \text{ Otherwise} \end{array} \right\}
\end{aligned}$$

Equation 4,5,6, and 7: Definition of module gene flow, and associated times of co-expression.

where  $B_N^M$  is the module  $M$  at threshold  $N$ , and  $B_{N+1, N}^-$  is the difference in genes between threshold  $N$  and  $N + 1$ .

By definition, the thresholds at which genes flow into and out of a given module represent the activation thresholds at which those genes become and cease to become co-expressed participants of the module. We hypothesize that genes which share a relatively long window of co-expression within the module are co-regulated with respect to the module, and represent functional groups within the individual modules. Further analysis along these lines will allow us to obtain finer resolution of the individual modules, as well as a better understanding of the mechanisms of co-regulation common to these functional groups.

Also of interest is the overall rate of gene flow into and out of a given module across all thresholds. Identifying threshold regions of greatest gene influx into the module and greatest gene efflux out of the module, we hypothesize will delineate threshold windows in which the individual sub-modules are most active. The existence and approximation of these windows would provide new and interesting insights into the global behavior of the genetic network.

## 7 Conclusion

We developed a method based on a normalized Jaccard distance to directly compare predictions between different module prediction methods. As a result, we found that both GRAM and the GeneXpress module prediction methods correspond to the Signature method at low threshold parameters. Furthermore, we applied our method to directly compare sets of modules predicted at different threshold of the Signature method. In this comparison, we observed an overall gene flow through the Signature modules at different thresholds. In our future work, we plan to better analyze the module

gene flow inherent to the signature method as a means of gaining finer resolution into the regulatory mechanisms within the modules, as well as a better understanding of the global relationships between the modules.

## References

- [1] Arndt, K.T., Styles, C., and Fink, G.R., Multiple global regulators control HIS4 transcription in yeast, *Science*, 237(4817):874–880, 1987.
- [2] Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A., and Gifford, D.K., Computational discovery of gene modules and regulatory networks, *Nat. Biotechnol.*, 21(11):1337–1342, 2003.
- [3] Bergmann, S., Ihmels, J., and Barkai, N., Iterative signature algorithm for the analysis of large-scale gene expression data, *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, 67(3 Pt 1):031902, 2003.
- [4] Gershon, D., Microarray technology: An array of opportunities, *Nature*, 416(6883):885–891, 2002.
- [5] Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N., Revealing modular organization in the yeast transcriptional network, *Nat. Genet.*, 31(4):370–377, 2002.
- [6] Murray, A.W., Whither genomics?, *Genome Biol.*, 1(1):COMMENT003, 2000.
- [7] Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., and Young, R.A., Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science*, 298(5594):799–804, 2002.
- [8] Pinson, B., Sagot, I., Borne, F., Gabrielsen, O.S., Daignan-Fornier, B., Mutations in the yeast Myb-like protein Bas1p resulting in discrimination between promoters *in vivo* but not *in vitro*, *Nucleic Acids Res.*, 26(17):3977–3985, 1998.
- [9] Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E.D., Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo, *BMC Bioinformatics*, 3(1):30, 2002.
- [10] Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N., Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data, *Nat. Genet.*, 34(2):166–176, 2003.
- [11] Tice-Baldwin, K., Fink, G.R., and Arndt, K.T., BAS1 has a Myb motif and activates HIS4 transcription only in combination with BAS2, *Science*, 246(4932):931–935, 1989.