

Learning regulatory programs that accurately predict differential expression with MEDUSA

Anshul Kundaje^a, David Quigley^b, Steve Lianoglou^a, Xuejing Li^c,
Marta Arias^d, Chris H. Wiggins^e, Li Zhang^f, Christina Leslie^{d*}

^aDepartment of Computer Science, ^bDepartment of Biomedical Informatics, ^cDepartment of Physics
^dCenter for Computational Learning Systems, ^eDepartment of Applied Physics and Applied Mathematics
^fDepartment of Environmental Health Sciences
Columbia University, New York, NY

*To whom correspondence should be addressed: cleslie@cs.columbia.edu

Abstract

Inferring gene regulatory networks from high-throughput genomic data is one of the central problems in computational biology and a principal focus of the DREAM initiative. In this paper, we describe a new predictive modeling approach for studying regulatory networks, based on a novel machine learning algorithm called MEDUSA. MEDUSA integrates promoter sequence, mRNA expression, and transcription factor occupancy data to learn gene regulatory programs that predict the differential expression of target genes. Instead of using clustering or correlation of expression profiles to infer regulatory relationships, MEDUSA determines condition-specific regulators and discovers regulatory motifs that mediate the regulation of target genes. In this way, MEDUSA meaningfully models biological mechanisms of transcriptional regulation. MEDUSA solves the problem of predicting the differential (up/down) expression of target genes by using boosting, a technique from statistical learning, which helps to avoid overfitting as the algorithm searches through the high dimensional space of potential regulators and sequence motifs. Experimental results demonstrate that MEDUSA achieves high prediction accuracy on held-out experiments (test data), i.e. data not seen in training.

The motivating problem behind the DREAM initiative is the difficulty of validating reverse engineered networks in the absence of a gold standard. Our approach of learning regulatory programs provides at least a partial solution for the problem: MEDUSA's prediction accuracy on held-out data gives a concrete and statistically sound way to validate how well the algorithm performs. With MEDUSA, statistical validation becomes a prerequisite for hypothesis generation and network building rather than a secondary consideration.

1 Introduction

We propose an approach to reverse engineering gene regulatory networks that is somewhat broader than the typical problem of learning network topology but, we believe, is still within the scope of the DREAM initiative. We present algorithmic methods for learning and interpreting predictive models of gene regulation, which we call "regulatory programs". In the context of this paper, a predictive model is one that accomplishes two goals. First, the model represents a regulatory program that predicts the differential expression of target genes in terms of biologically meaningful regulatory inputs, including the context-specific expression of transcriptional regulators and signal transducers and the presence of shared motifs in the regulatory sequences of target genes. Therefore, rather than directly learning a network or a set of clusters/modules, we are learning a prediction function, and we view the learning task as a prediction problem rather than a model selection problem. Second, the model should not only be able to make quantitative predictions, but it should make accurate predictions on data not seen in training (test data). The key issue is to use a learning strategy that avoids overfitting in the high dimensional feature space of potential regulators and sequence motifs and in the presence of noisy gene expression data. Our strategy is based on boosting [1], a technique from statistical learning theory that has empirically shown resistance to overfitting in noisy and high dimensional settings.

The core of our approach is a novel algorithm called MEDUSA [8] (= Motif Element Discrimination Using Sequence Agglomeration), which integrates mRNA expression and regulatory sequence data to discover motifs representing putative transcription factor binding sites and to build a global gene regulatory program. The inputs to MEDUSA are the gene-specific regulatory sequences and the experiment-specific expression levels of regulators, including those that do not bind DNA. If available, we can use hits of known motifs or transcription factor occupancy data from ChIP chip experiments to provide additional gene-specific features [9, 7]. We discretize target gene expression data into up, down, or baseline states in order to reduce noise,

so that rather than trying to predict a real-valued expression level, we predict only whether the gene is up- or down-regulated. This reduction also allows us to exploit modern and effective classification algorithms. MEDUSA use the Adaboost learning algorithm with a margin-based generalization of decision trees called alternating decision trees (ADTs). In experimental results (Section 3), we show that MEDUSA achieves high prediction accuracy on multiple yeast data sets.

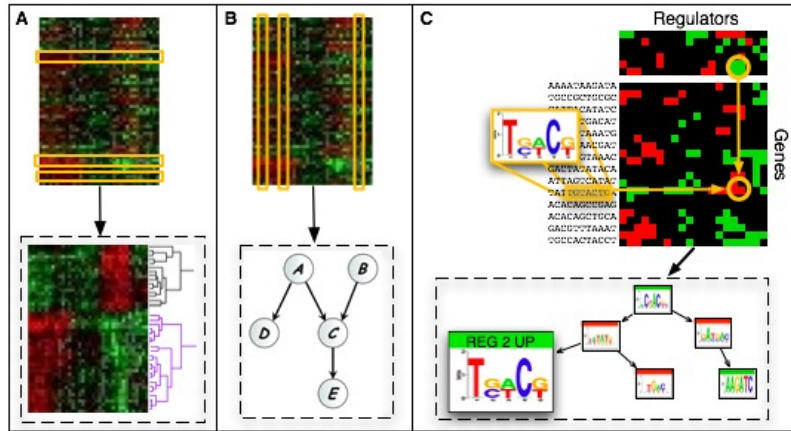


Figure 1: Use of expression data in clustering, Bayesian networks, and MEDUSA. (A) Clustering and many network inference algorithms consider rows of the expression matrix, representing expression profiles for genes, and compute pairwise similarities between rows. The pairwise similarities are used to produce a set of gene clusters or to determine edges in a network. (B) Bayesian networks treat each column of the expression matrix, representing the expression levels of all genes in a particular experiment, as the joint observation of thousands of gene random variables. Joint expression observations are modeled probabilistically to determine a model that maximizes the likelihood (or a related Bayesian objective function) of the data. (C) In MEDUSA, every differentially expressed target gene-experiment example is a separate training example, represented by sequence data (the gene’s promoter sequence) and regulation expression data (the expression states of regulators in the experiment). MEDUSA discovers sequence motifs and selects regulators that jointly help predict target expression across the entire training set. The regulatory program can then be used to predict up/down target gene expression in held-out data.

Our approach uses expression data in a significantly different way than previous approaches like clustering, correlation-based network inference algorithms, and Bayesian networks, as illustrated in Figure 1. Instead of computing correlations between rows of the expression matrix, as in clustering and many simple network learning approaches, or viewing every column of the matrix as a joint observation of thousands of gene variables, as in Bayesian networks, in the MEDUSA approach, every differentially expressed target gene example is a training example. We learn to predict the up/down expression of these training examples by using both sequence data (via motifs we discover in the promoters) and the expression of regulators. This way of using expression data and integrating sequence data allows us to learn from a very large training set, typically consisting of 10,000s of differentially expressed examples. We also avoid using information that is not apparently available to the transcriptional machinery of the cell, such as the identity of target genes or the cluster membership of genes. Finally, because we use biologically meaningful inputs for learning MEDUSA regulatory programs, we can also analyze the learned model to extract biologically meaningful information, derive specific hypotheses about gene regulation, and even produce network representations of the significant predictive relationships between regulators and targets, as shown below (Figure 5).

2 Methods

The core of our computational approach is a novel machine learning algorithm called MEDUSA, which integrates regulatory sequence and mRNA expression data to learn a regulatory program that can accurately predict the differential expression of target genes. MEDUSA differs from most previous studies by implementing a number of key algorithmic features: (1) it integrates promoter sequence and expression to learn a global regulatory program; (2) it avoids overfitting when training in a high dimensional feature space by use of a machine learning technique called boosting; (3) it learns functional contributions of both regulators and

motifs; (4) it learns binding site motifs directly from sequence without seeding the algorithm with known motifs; (5) it automatically learns the threshold for deciding the presence of motifs. These strengths of the algorithm are critical for achieving high predictive power across multiple data sets (see Section 3).

The inputs to the MEDUSA algorithm are a list of regulators, including those that do not bind DNA, the promoter sequences for all target genes, and gene expression training data that has been discretized into up, down, and baseline expression levels. MEDUSA learns sequence motifs whose presence in the promoters of target genes, together with the mRNA levels of regulators across experimental conditions, helps to predict the differential (up/down) expression of the targets. MEDUSA uses boosting [1], a general binary prediction algorithm from statistical learning theory, to build this prediction function or regulatory program. Empirically, boosting often learns to make large-margin (confident) predictions on the training set, which is theoretically linked to its ability to obtain good generalization on test data even when the feature space is very high dimensional (that is, it avoids overfitting the training data).

MEDUSA models the control logic of transcriptional regulation in the form of an alternating decision tree (ADT). An ADT is a generalization of a decision tree that consists of alternating layers of decision nodes, which ask yes/no questions based on particular features, and prediction nodes, which contain a real-valued score associated with the yes or no answer. Given the promoter sequence of a gene and the expression level of the regulators in an experiment, the MEDUSA regulatory program asks yes/no questions of the form, “Is motif X present in the upstream region of the gene and is the state of regulator Y up (or down) in that experiment?”, in the ADT decision nodes. If the answer is “yes”, we add the real value contained in the prediction node to the overall prediction score for the example, and we continue down to the next decision node; if the answer is “no”, there is no score contribution. To compute the prediction score for a gene-experiment example, we start at the root node and recursively check which decision nodes we can pass through by answering “yes” to the condition, working from the top to the bottom of the ADT; the prediction score is the sum of all the prediction node scores in all paths in the ADT that we visit in this process.

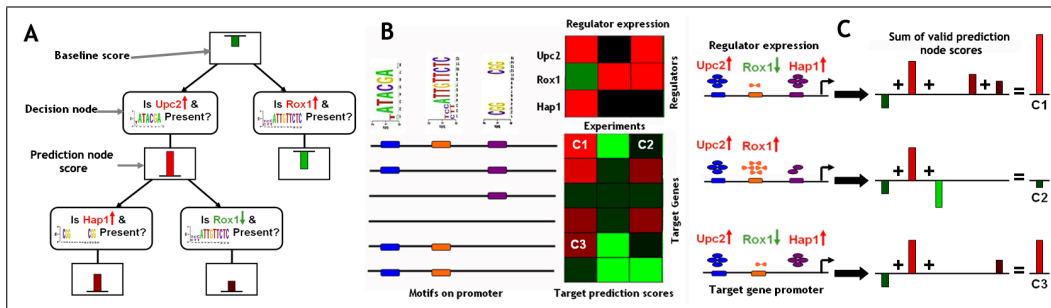


Figure 2: MEDUSA learns genome-wide, context-specific regulation programs. A simplified example to show how MEDUSA regulatory programs predict differential target gene expression. (A) The MEDUSA regulatory program is described by an alternating decision tree that asks questions about the expression level of regulators in the experimental condition and the presence of motifs in the gene’s promoter. (B) A heat map showing predicted scores as computed by the regulatory program, based on motif content in the promoters and condition-specific expression of regulators. (C) Interpretation of the prediction scores. For example, in gene-experiment C1, all three motifs are present in the gene’s promoter, while regulators Upc2 and Hap1 are up and Rox1 is down; therefore, both Upc2 and Hap1 are likely to bind to the gene’s promoter, while Rox1 is less likely to bind. Reading off contributions from active nodes in the regulatory program and summing the contributions, we obtain a confident prediction of upregulation of the target gene in this condition.

Figure 2 gives a simplified example of a MEDUSA regulatory program and shows how it makes genome-wide, context-specific predictions of differential target gene expression. This example is motivated by the significant regulators and motifs we found in our analysis of the yeast hypoxia data set (unpublished results, [6]), but the small regulatory program shown is just a pedagogical example.

3 Results

We report experimental results for MEDUSA on three expression data sets of different sizes and involving very different biological processes in the yeast *S. cerevisiae*: a large data set measuring yeast stress response

to diverse environmental stresses (ESR) [3], a smaller data set for response to DNA-damaging agents (DNA damage) [2], and a small unpublished data set from our collaborator Dr. Li Zhang studying the role of oxygen, heme, Hap1, and Co²⁺ in oxygen sensing and regulation (Hypoxia). The Leslie and Zhang labs have collaborated on a detailed MEDUSA-based study of the yeast oxygen sensing and regulatory network using this last data set [6].

| Data set | #expts | cross-validation set-up | algorithm, #rounds | sequence features | error rate |
|------------|--------|--|--------------------|----------------------|------------|
| ESR | 173 | 10-fold c.v., held-out expts, replicate expts grouped | ADTs, 700 rounds | promoters only | 13.4% |
| | | | | AlignAce motifs only | 16.1% |
| | | | | TRANSFAC motifs only | 20.8% |
| | | 10% random held-out examples | stumps, 100 rounds | promoters only | 17.6% |
| DNA damage | 52 | 10% random held-out examples | stumps, 100 rounds | promoters only | 25.0% |
| Hypoxia | 18 | 10-fold c.v., held-out examples | ADTs, 450 rounds | promoters + ChIP | 8.0% |
| | | | | ChIP only | 26.0% |
| | | 10-fold c.v., held-out examples replicate gene-expts grouped | ADTs, 450 rounds | promoters + ChIP | 23.9% |

Figure 3: Prediction performance for MEDUSA across multiple yeast data sets.

| TFNAME | TFPSSM | MEDUSA PSSM | RoleID | ALLR | Pvalue | TFNAME | TFPSSM | MEDUSA PSSM | RoleID | ALLR | Pvalue |
|--------|--------|---------------------|--------|---------|-----------|--------|--------|------------------------|--------|---------|-----------|
| RAP1 | | CCATACA | 19 | 13.4732 | 1.431e-11 | ZAP1 | | TAAAGGG | 146 | 10.4110 | 6.867e-09 |
| TBP | | TTTATAT | 76 | 10.1249 | 1.023e-08 | XBP1 | | TCGAGAA | 24 | 9.6252 | 2.265e-08 |
| PRX3 | | AGCGGAC | 118 | 9.6252 | 2.588e-08 | ZAP1 | | ACCCCTC | 125 | 9.6252 | 3.05e-08 |
| HAC1 | | CCGACGC | 112 | 9.6252 | 3.328e-08 | ZAP1 | | CGCCTGA | 32 | 9.6252 | 5.5e-08 |
| GCN4 | | CTCATCG | 129 | 9.6252 | 8.735e-08 | MSN2 | | AGGGGT | 21 | 8.5612 | 1.801e-07 |
| GCN4 | | CTCATA | 40 | 8.6595 | 5.654e-07 | HAC1 | | CACAGC | 159 | 8.3255 | 5.899e-07 |
| STE12 | | GTTGAAA | 63 | 7.8624 | 1.167e-06 | STE12 | | CGCAAA | 14 | 7.7001 | 1.274e-06 |
| ZAP1 | | CCC ₂ AT | 30 | 7.9777 | 1.28e-06 | MSN2 | | TT ₂ AGGG | 10 | 7.6848 | 1.316e-06 |
| ROX1 | | ACAACGC | 110 | 7.7001 | 1.638e-06 | MATa1 | | TGTACGG | 7 | 7.7001 | 1.82e-06 |
| STE12 | | GAATTTTTC | 13 | 7.7001 | 2.002e-06 | REB1 | | ACCC ₂ TTGG | 169 | 7.7001 | 2.34e-06 |

Figure 4: Significant TF binding site motifs found by MEDUSA for the ESR dataset. The table shows some of the PSSMs found by MEDUSA that most significantly match experimentally verified TF binding sites represented in TRANSFAC. The significance of the match is reported by the p -value associated with the averaged log likelihood score for the pair of PSSMs, computed by MatAlign.

Results showing that MEDUSA achieves low prediction error rates across all three data sets and in different cross-validation experiments are summarized in Figure 3. In all experiments, we used a comprehensive list of 475 candidate regulators, including known and putative transcription factors (TFs), kinases, phosphatases, and receptors. We used 1000bp 5' sequences as promoter sequences. In all cases, we report the error rate of MEDUSA's up/down predictions on the differentially expressed test examples, i.e. baseline examples are not used for training nor are they included in this evaluation.

Our largest scale experiments were performed on the ESR data set. Here we divided experiments into 10 folds and grouped replicate experiments together within folds for 10-fold cross-validation. This procedure ensured that replicates of an experimental condition would never be in both training and test sets, making the prediction task more difficult. In this setting, learning motifs directly from promoter sequences, MEDUSA achieves an impressively low error rate of 13.4%. We also compared this result against using MEDUSA's underlying boosting algorithm with a fixed set of database motifs as sequence features; using either TRANSFAC [13] or AlignAce [10] resulted in a significantly higher prediction error rate, suggesting that MEDUSA can extract sequence information that is not represented in these databases. Conversely, to compare MEDUSA motifs to known TF binding sites as represented in TRANSFAC and show that we do retrieve true motifs,

we performed a statistical analysis. For each of the MEDUSA PSSMs generated in the first 200 rounds of boosting, we compared the MEDUSA motif against all TRANSFAC PSSMs using p -values based on an averaged log likelihood ratio (ALLR), as reported by the MatAlign program (personal communication, Dr. Gary Stormo). Figure 4 shows the top 20 matches of MEDUSA matches to experimentally verified TF binding sites from TRANSFAC and their ALLR p -value. Of these 20 most significant matches, MEDUSA found 10 in the first 40 rounds of boosting. The most frequent regulators to appear in nodes of the MEDUSA ADT included Tpk1, a key kinase in the PKA signaling pathway that operates in stress response; Usv1, an important TF also identified in a previous computational study [12]; and Xbp1, a universal stress repressor.

For the DNA damage data set, we did a quicker evaluation of MEDUSA performance, using 10% randomly held-out gene-experiment examples as a test set. For faster results, we also used boosting with stumps instead of ADTs (i.e. all nodes are placed directly under the root node) and ran for only 100 rounds. DNA damage is both smaller and more diverse than ESR, with many of the experiments involving gene knockout strains and mutants, so that the reference conditions vary across the data set; our noise analysis also revealed that the data were far noisier than in ESR. For these reasons, the error rate of 25%, while not as good as ESR, still represents significant generalization performance. Among the 21 regulators that MEDUSA selects for the regulatory program, we find CDC5 (a key transcription factor for inactivation of the Rad53 checkpoint), several other cell cycle regulators (CLB2, CLN2, KAR4) and oxidative damage pathway regulators (MSN4, YAP1, TPK1, and TPK2). We also found several members of the GLC7 phosphatase complex: SDS22, GLC8, PPZ2 and SHP1. An analysis of the GO terms of the regulators discovered by MEDUSA reveals enrichment of response to oxidative stress, believed to be related to DNA damage pathways [11].

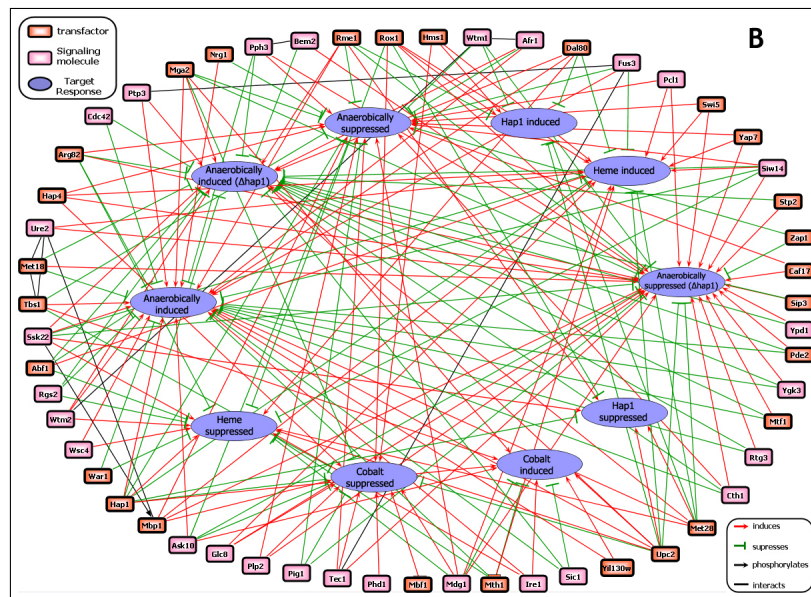


Figure 5: Oxygen sensing and regulatory network assembled through analysis of MEDUSA regulatory program. A network showing the putative regulatory effect of the statistically significant regulators on sets of target genes that are differentially expressed in various conditions. A red arrow indicates that the regulator induces target genes in the set, while a green arrow indicates that the regulator suppresses the target gene set. Known phosphorylation interactions are shown by black arrows and known protein-protein interactions by black edges without arrows.

Finally, we performed 10-fold cross-validation experiments on held-out gene-experiment examples for the Hypoxia data set [6]. Here, we achieved a very low error rate of 8.0%, but the prediction task was made easier by the presence of replicates in the data set: each of 6 experimental conditions was represented by 3 replicates, though one was a different yeast strain. Still, using TF occupancy features from ChIP data (from [5]) without promoter sequence data led to a much higher error rate (26.0%), suggesting that the problem is not trivial (and that the conditions under which the ChIP chip experiments were performed are not relevant to hypoxia). We repeated the 10-fold cross-validation but grouped replicate measurements (replicate expression values for the same gene and in the same condition) within folds, so that we would never see the same gene

in the same condition in both training and testing. In this more difficult setting, MEDUSA still achieved significant prediction performance, with an error rate of 23.9%.

We carried out a detailed analysis of the MEDUSA regulatory program for the Hypoxia data set (full results are described in [6]). By restricting the regulatory program to sets of induced or suppressed target genes in specific conditions and ranking the importance of the regulators using a margin-based score, we identified 54 significant context-specific regulators. In Figure 5, we show a network representation of our analysis, where an edge between a regulator and a target gene set indicates that the regulator makes a significant contribution to the correct prediction of differential expression for the targets. We have also indicated known phosphorylation interactions [14] and protein-protein interactions [4] between regulators.

4 Discussion

Learning regulatory programs with MEDUSA provides an alternate approach to the problem of validating reverse engineering algorithms without a gold standard. MEDUSA uses boosting to avoid overfitting in the high dimensional feature space of potential regulators and motifs, and its prediction accuracy on held-out data gives a concrete statistic for validation. MEDUSA's regulatory programs can be used to extract significant context-specific regulators and to build regulatory networks. However, we should also keep in mind, at this early stage in the reverse engineering enterprise, that networks are not the only way to conceptualize the interconnectivity of interactions in the cell. Perhaps the more general notion of a regulatory program will emerge as a useful abstraction for future work in reverse engineering.

References

- [1] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
- [2] A. P. Gasch, M. Huang, S. Metzner, D. Botstein, S. J. Elledge, and P. O. Brown. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol Biol Cell.*, 12(10):2987–3003., Oct 2001.
- [3] A. P. Gasch, P. T. Spellman, C. M. Kao, Orna Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11:4241–4257, 2000.
- [4] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature.*, 440(7084):631–6. Epub 2006 Jan 22., Mar 30 2006.
- [5] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature.*, 431(7004):99–104., Sep 2 2004.
- [6] A. Kundaje, C. Lan, M. Zhou, C. Leslie, and L. Zhang. A predictive model of the oxygen sensing and regulatory network in yeast. Submitted for publication.
- [7] A. Kundaje, M. Middendorf, M. Shah, C. H. Wiggins, Y. Freund, and C. Leslie. classification-based framework for predicting and analyzing gene regulatory response. *BMC Bioinformatics.*, 7 Suppl 1:S1–5., Mar 20 2006.
- [8] M. Middendorf, A. Kundaje, M. Shah, Y. Freund, C. Wiggins, and C. Leslie. Motif discovery through predictive modeling of gene regulation. *Proceedings of the Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2005)*, 2005.
- [9] M. Middendorf, A. Kundaje, C. Wiggins, Y. Freund, and C. Leslie. Predicting genetic regulatory response using classification. *Proceedings of the Twelfth International Conference on Intelligent Systems for Molecular Biology (ISMB 2004)*, 2004.
- [10] Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, 2:153–159, 2001.
- [11] T. B. Salmon, B. A. Evert, B. Song, and P. W. Doetsch. Biological consequences of oxidative stress-induced DNA damage in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, 32(12):3712–23. Print 2004., Jul 14 2004.
- [12] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: Identifying regulatory modules and their condition specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, 2003.
- [13] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüss, I. Reuter, and F. Schacherer. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Research*, 28:316–319, 2000.
- [14] H. Zhu, J. F. Klemic, S. Chang, P. Bertone, A. Casamayor, K. G. Klemic, D. Smith, M. Gerstein, M. A. Reed, and M. Snyder. Analysis of yeast protein kinases using protein chips. *Nat Genet.*, 26(3):283–9., Nov 2000.