Identifying Context-Specific Transcription Factor Targets from Prior Knowledge and Gene Expression Data

Elana J Fertig*, Alexander V Favorov*†‡ and Michael F Ochs*§
*Department of Oncology, SKCCC, School of Medicine, Johns Hopkins University, Baltimore, MD, USA
†Department of Computational System Biology, Vavilov Institute of General Genetics, Moscow, Russia
‡Laboratory of Bioinformatics, GosNIIGenetika, Moscow, Russia
§Department of Health Science Informatics, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

***Abstract*—Numerous methodologies, assays, and databases presently provide candidate targets of transcription factors (TFs). However, TFs rarely regulate their targets universally. The context of activation of a TF can change the transcriptional response of targets. Direct multiple regulation typical to mammalian genes complicates direct inference of TF targets from gene expression data. We present a novel statistic that infers context-specific TF regulation based upon the CoGAPS algorithm, which infers overlapping gene expression patterns resulting from coregulation. Numerical experiments with simulated data showed that this statistic correctly inferred targets that are common to multiple TFs, except in cases where the signal from a TF is negligible relative to noise level and signal from other TFs. The statistic is robust to moderate levels of error in the simulated gene sets, identifying fewer false positives than false negatives. Significantly, the regulatory statistic refines the number of transcription factor targets relevant to cell signaling in gastrointestinal stromal tumors (GIST) to genes consistent with the phosphorylation patterns of TFs identified in previous studies. As formulated, the proposed regulatory statistic has wide applicability to inferring set membership in integrated datasets. This statistic could be naturally extended to account for prior probabilities of set membership or to add candidate gene targets.**

***Keywords*-Bioinformatics; Genetic expression; Genomics**

I. Introduction

Transcriptional regulators play a key role in developmental processes and normal cellular homeostasis by controlling reprogramming of cells. Reprogramming of normal cells is typically driven by internal and external signals (e.g., metabolic changes, growth factors) with activation of transcription factors (TFs) leading to both repression and activation of transcription of target genes. If such transcriptional changes are driven at inappropriate times by aberrant activity, the system can switch to a diseased state. Therefore, identifying TF activity can implicate underlying biochemical processes in a disease system, such as cell signaling in tumorigenesis [1]. However, it is clear that TFs regulate different genes in different contexts, and the effect of TF activation therefore is cell-type and cell-context dependent. Determining TF targets activated in specific instances can help to refine understanding of disease by identifying the specific genes aberrantly activated in cells showing the disease phenotype. As a result, these targets may provide insight into treatment options or genomic biomarkers unidentifiable in single gene analyses or standard gene set analyses.

Inference of TF gene targets is an active area of research that involves numerous quantitative and experimental techniques. ChIP-chip and ChIP-seq assays have been used widely to detect TF targets (reviewed in [2]). However, these methods require laborious examination of each TF independently, leading to substantial costs for genome-wide investigation. Alternatively, the *in silico* techniques that predict candidate targets often detect all candidate targets regardless of actual activation by TFs in the specific biological background [3]. These limitations in measurement and prediction techniques make context-specific target identification difficult even in curated databases of TF targets.

Ideally, prior knowledge of TF candidate targets can be refined by integrating global gene expression measurements to infer context-specific TF targets from evidence of transcript generation. For example, integrating ChiP candidate targets in an analysis of expression response to Pou5f1, Sox2, and Nanog suppression refined context-specific knowledge of targets of these TFs [4]. However, such direct inference methods are intractable for most *in vivo* studies, in which TF activity can be neither directly manipulated nor measured. Recently, numerous techniques to refine genes in gene sets or identify patient specific pathways by integrating expression and other data have been developed [5], [6], extending previous algorithms [7]. These methods are generally based upon inference of common expression responses in candidate genes inferred with clustering, principal component analysis (PCA), or network-based analyses. However, previous studies have found that the techniques underlying these algorithms have difficulty accounting for the regulation of individual genes by multiple transcription factors or secondary transcriptional effects due to feedback [8].

Sparse Markov chain Monte Carlo (MCMC) matrix factorization algorithms, such as CoGAPS [9], have been shown to infer patterns across samples that relate to transcription factor activity and account for multiple regulation of individual genes [1]. While previous extensions leveraged prior

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

knowledge of targets to more accurately estimate TF activity [10], no technique has been able to refine gene targets based on context. This paper extends our approach to include a gene-regulatory statistic that predicts context-specific TF targets. Analysis of simulations and time course data from gastrointestinal stromal tumor (GIST) cell lines demonstrate that the algorithm successfully integrates gene expression data and prior knowledge to accurately determine context-specific targets.

II. Algorithm to identify context-specific transcription factor targets

We sought a regulatory statistic that computes the probability that a gene g is a member of gene set $\mathcal{G}$ based upon prior knowledge of gene set membership, $\Pr(g \in \mathcal{G})$, and gene expression data. We based the proposed regulatory statistic for membership of g in $\mathcal{G}$ upon comparisons of the expression profile of g to the expression pattern CoGAPS infers for members of $\mathcal{G}$. In the formulation below, we assumed that the priors on $\Pr(g \in \mathcal{G})$ were binary, but we note that the algorithm could be modified easily to include non-binary probabilities, such as arise in ChIP-seq. Due to the binary prior, we set the symbol $\mathcal{G}$ to represent candidate set members, $\mathcal{G} \equiv \{g|\Pr(g \in \mathcal{G}) = 1\}$, in addition to representing the gene set itself.

The CoGAPS algorithm was developed to infer expression patterns across samples that are shared by multiple genes. The algorithm thus modeled gene expression by factoring the n gene $\times$ m sample data $\mathbf{D}$ into an $n \times p$ amplitude matrix $\mathbf{A}$ and a $p \times m$ pattern matrix $\mathbf{P}$, so that

$$D_{ij} \sim \mathcal{N}\left((\mathbf{AP})_{ij}, \Sigma_{ij}\right), \tag{1}$$

where $\mathcal{N}(\mu, \sigma)$ is a normal distribution with mean μ and standard deviation σ, and Σ_{ij} is a standard deviation representing the uncertainty in D_{ij}. Using an atomic prior [11], CoGAPS inferred $\mathbf{A}$ and $\mathbf{P}$ by Markov chain Monte Carlo (MCMC). CoGAPS then calculated a Z-score statistic to infer the amplitude of the gene set $\mathcal{G}$ in each of the patterns as

$$Z_{\mathcal{G}p} = \frac{1}{G} \sum_{g \in \mathcal{G}} \frac{\langle A_{gp} \rangle}{\mathrm{sd}(A_{gp})}, \tag{2}$$

where G is the number of elements in $\mathcal{G}$ and where $\langle A_{gp} \rangle$ and $\mathrm{sd}(A_{gp})$ are respectively the posterior mean and standard deviation of A_{gp} estimated by MCMC. The probability that members of $\mathcal{G}$ are upregulated in each pattern ($\Pr_{\mathcal{G}p}$) is then inferred using a permutation test, comparing $Z_{\mathcal{G}p}$ to the comparable Z-score statistic resulting from applying eq. 2 to random sets drawn from the appropriate column of $\mathbf{A}$ containing G members.

We hypothesized that a gene g was regulated similarly to other members of $\mathcal{G}$ *in this context*, if its expression profile was similar to other members of the set. This hypothesis was quantified by comparing the activity of gene g to that of $\mathcal{G}$ across all patterns, using

$$S_{g,\mathcal{G}} = \frac{\sum_p -\log(\Pr_{\mathcal{G}p}) \langle A_{gp} \rangle / \mathrm{sd}(A_{gp})}{\sum_p -\log(\Pr_{\mathcal{G}p})}. \tag{3}$$

The measure $S_{g,\mathcal{G}}$ will be large when g has amplitude in patterns that are upregulated in $\mathcal{G}$, and zero otherwise. The probability of set membership was then computed by comparing the value of $S_{g,\mathcal{G}}$ for each gene to the distribution of values of the statistic in eq. 3 for genes outside of the set (namely, $\mathcal{G}^C = \{g|\Pr(g \in \mathcal{G}) = 0\}$). Using the logarithm in eq. 3 insured that our statistic had greater sensitivity to patterns with small p-values of upregulation than would result from scaling the Z-scores by $1 - \Pr_{\mathcal{G}p}$. Because eq. 3 did not not include a penalty term for activity in sets outside of $\mathcal{G}$, $S_{g,\mathcal{G}}$ can have large values for multiple gene sets $\mathcal{G}$ with correspondingly low p-values inferred from the permutation test. It is therefore likely that the statistic can further facilitate inference of multiple regulation of a gene g within a CoGAPS analysis.

III. Methods

A. Implementation

All analyses reported in this manuscript were performed using the Bioconductor package for CoGAPS [9], and the set membership statistic of eq. 3 is implemented in the function `computeGeneGSProb`. The inferred, normalized pattern for each gene set was computed as

$$\tilde{P}_{\mathcal{G}} = \sum_p -\log(\Pr_{\mathcal{G},p}) \mathbf{P}_{p\bullet}, \tag{4}$$

where $\mathbf{P}_{p\bullet}$ represents the p^{th} row of the pattern matrix and $\tilde{P}_{\mathcal{G}}$ is normalized to have a sum of 1 for plotting.

B. Simulated data

The simulated dataset depicted in Figure 1a consisted of four TFs, with targets represented in Figure 1b and activity represented by patterns in Figure 1c. Genes were pre-selected as being regulated by one to four of the TFs. The specific TFs that regulated each of these genes were selected at random to avoid biasing the analysis toward any subset of the four TFs. The maximum magnitude by which a TF regulates each gene target when fully active (i.e., a value of 1 for thecorresponding element in the $\mathbf{P}$ matrix) were drawn from an exponential distribution with parameter $\frac{1}{3}$. Typical multiplicative noise of 10% was added to the simulated data.

C. GIST data and transcription factor targets

The GIST expression data and TF targets from TRANSFAC [12] were preprocessed [1] and deposited as data files in the CoGAPS package [9]. Raw data is available in GEO (GSE17018). The analysis was performed as described in the CoGAPS User's Manual.

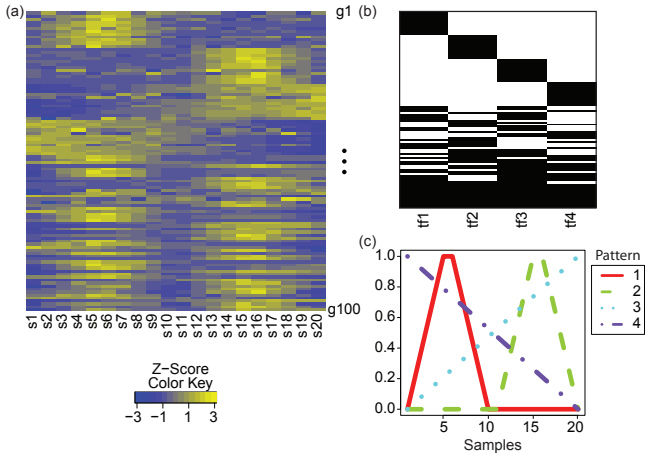

Figure 1. Overview of simulated data. (a) Heatmap of the relative $\log_2$ expression of each gene, colored according to the color key. (b) Gene targets of each simulated TF in black shading in rows. (c) Simulated patterns across samples. Patterns are numbered according to the TF to which they are assigned (e.g., 1 for tf1).

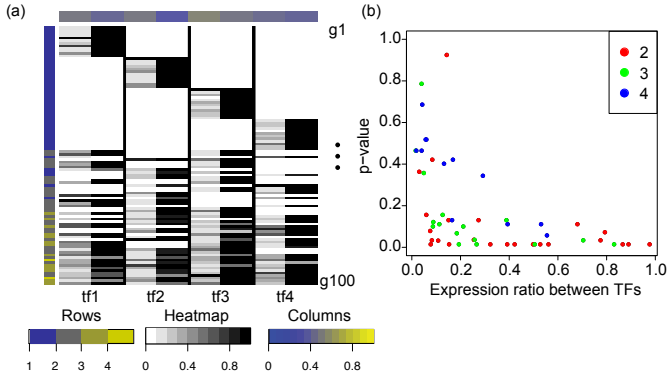

Figure 2. Results of regulatory statistics for the simulated dataset. (a) Heatmap comparing the maximum amount of activity in a gene resulting from TF activity (left) to the estimated probability of set membership (right) for each TF. The true magnitude of gene regulation is rescaled to 0.1 at minimum and 1 at maximum for each TF. Similarly, estimated p-values are converted to probabilities of activity for plotting $(1 - p)$. Color shading in rows and columns estimate the number of gene targets inferred at a p-value threshold of 0.1, relative to the total number of simulated genes in the case of column shading. (b) For each gene, for the transcription factor with the smallest regulation of expression, the p-value inferred with the regulatory-statistic is plotted against the ratio between the amount that the selected transcription factor regulates expression and the maximum expression across all TFs. Colors indicate the number of TFs that regulate each of the genes plotted, according to the color key

IV. Simulation Results

In order to test the regulatory statistic resulting from the permutation test on the summary statistic of eq. 3, we simulated a dataset containing four TFs as described in Methods (Figure 1a). Each of these TFs (tf1 through tf4) had unique targets and targets shared with the remaining TFs (Figure 1b). The expression for the targets were set at a variety of levels when the TF was active (activity indicated in the rows $\mathbf{P}$, Figure 1c). The analyses described in the following subsections assessed the accuracy of the gene-regulation statistic of eq. 3 when applied to the ground truth simulated TF target-sets with varying degrees of error.

A. *CoGAPS analysis recovers simulated TF activity*

We first applied the GAPS matrix factorization algorithm of eq. 1 to the simulated dataset. Three simulations using the GAPS algorithm successfully recovered the true $\mathbf{A}$ and $\mathbf{P}$ matrices. Moreover, each simulation had χ^2 fit values (1843.7, 1844.0, and 1843.5 respectively) comparable to the true χ^2 value (2010.0). The CoGAPS gene set statistic resulting from a permutation test on the Z-score statistic of eq. 2 likewise uncovered the patterns in which the simulated transcription factors were upregulated.

B. *The regulatory statistic inferred set membership when the signal was above noise levels*

We initially validated the proposed regulatory statistic in the ideal scenario for which the true members of the gene set were input as the elements of $\mathcal{G}$ in eq. 3. The regulatory statistic recovered most targets of each of the simulated TFs (Figure 2a). At a p-value threshold of 0.1, the statistic identified 85% of the targets for tf1, 73% for tf2, 89% for tf3, and 93% for tf4.

In most cases, the statistic also correctly assigned a gene to multiple TFs when that gene was multiply regulated. The statistic only missed multiply regulated genes for TFs that weakly regulated the gene, as shown in Figure 2b where weak regulation is toward the left. We hypothesized that in these cases, the algorithm could not distinguish the weak signals induced by activation of these TFs from noise in the simulated data. Further confirming this hypothesis, the genes excluded by the statistic from a TF target set had significantly different expression patterns than the pattern estimated by CoGAPS for that TF (Figure 3). Here, the pattern from the weakly regulated TF was lost due to weak signal relative to noise and to the regulation by the other TFs. In contrast genes assigned to the TF had expression patterns that at least partially matched the estimated transcription factor pattern. In this case, the regulatory statistic inferred genes that were not purely correlated to one pattern, facilitating the successful inference of multiple regulation.

C. *The regulatory statistic is robust to moderate error in the prior for set membership*

We also tested the regulatory statistic in the more realistic scenario in which the prior estimate of membership in $\mathcal{G}$ was error-prone. In this case, we performed simulations in which a fixed number of genes from the true TF regulatory set were replaced with a random set of genes that were not regulated by that TF. Fifty simulations were performed for errors from

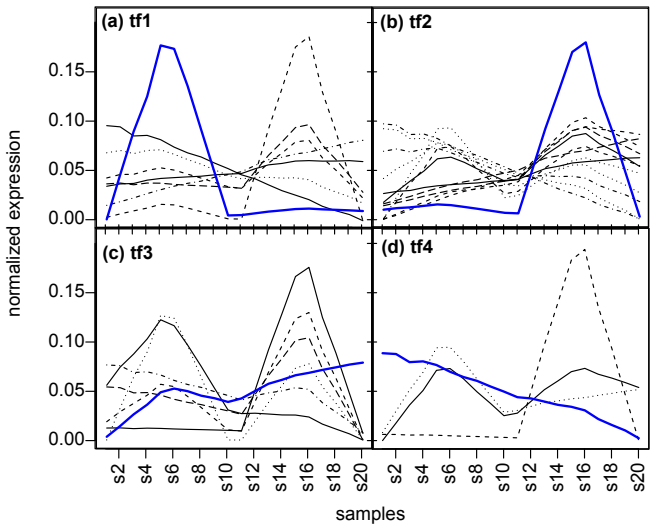

Figure 3. Comparison of the inferred gene set activity to excluded genes for the simulated dataset. The normalized, de-noised expression profile inferred from the GAPS approximation of the gene expression data ($\mathbf{AP}$) for each gene with a p-value greater than 0.1 from the regulatory statistic (black lines) is compared to the inferred, normalized pattern for the gene set (blue line). Results are plotted for each of the four TFs in panels (a)-(d) as labeled in the figure. In each case, the black line represents a gene that prior information declared to be regulated by the TF but which the algorithm determined not to be.

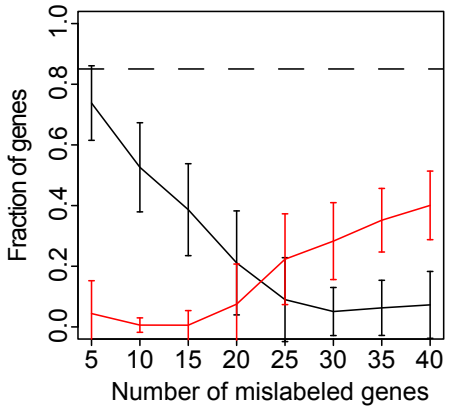

Figure 4. Inference of set-membership with errors in each of the simulated transcription factor sets. Mean fraction of true gene set members inferred with a p-value threshold of 0.1 for the regulatory statistic (black) and false gene set members (red) for fifty simulations with the number of genes misassigned to the gene set on the x-axis. Error bars for both curves represent the corresponding standard deviation estimated over each of the sets of fifty simulations. The dashed black line represents the percentage of true genes recovered in simulations containing no incorrect assignements in the gene set for reference.

5 genes to 40 genes, representing a range of errors from 10% to 83% for tf1, 10% to 83% for tf2, 9% to 75% for tf3, and 12% to 93% for tf4. We compared the number of genes correctly inferred to the number of genes incorrectly inferred summarized across all of the TF sets (Figure 4). The statistic rapidly dropped genes from set membership (black line) as the misassignment in $\mathcal{G}$ increased. However, the statistic only rarely incorrectly assigned genes to a set (red line) even for a significant error rate based on 15 genes (31% of the targets of tf1 and tf2, 28% of tf3, and 35% of tf4). Once the prior on the gene sets was dominated by incorrect genes, the statistic inferred a larger number of false gene set members than true gene set members. In contrast to the set of true genes, the number of false gene set members inferred never reached the majority of set members. Similar results were observed when analyzing the effects of gene set errors in individual TFs.

V. Analysis of GIST cell line gene expression data

We applied the gene regulatory statistic from eq. 3 to the time course gene expression data of GIST cell lines treated with imatinib (IM) and TF targets from TRANSFAC for TFs downstream of the c-KIT activating mutation [1]. The CoGAPS algorithm recovered the response of TFs to therapy previously validated [1]. The matrix factorization identified patterns of gene expression that decreased in time

Table I
Relative number of transcription factor targets retained from the regulatory statistic with a p-value threshold of 0.1 for the GIST cell line data.

TF	# genes	# significant genes
Ap1	99	10
cJun	71	0
cMyc	13	4
CREB	91	14
E2F-1	24	4
Elk-1	13	6
FOXO	9	4
NF-kappaB	51	6
p53	27	10
Smad4	19	2
Sp1	398	42
STAT3	22	5

with IM treatment, increased with IM treatment, and had a transitional increase with IM treatment. The CoGAPS gene set statistic likewise inferred significant upregulation of Elk-1 in the rising pattern, upregulation of cJun, cMyc, and Elk-1 in the falling pattern, and significant upregulation of cJun and p53 in the transient pattern.

The gene regulatory statistic of eq. 3 significantly decreased the number of targets of all TFs, regardless of the patterns to which TFs were assigned (Table I), and decreased the number of genes predicted to be multiply regulated (Table II). Due to the significant reduction, no genes were inferred to be regulated by more than two TFs. Based upon the simulated data results, we hypothesized that the reduction in genes inferred as multiply regulated arises from the general, significant decrease in context-specific

Table II
SUMMARY OF THE STATISTICS FOR CANDIDATE TARGET GENES THAT THE TRANSFAC DATABASE NOTES AS REGULATED BY $1-6$ OF THE TFS CONSIDERED IN TABLE I (ROWS). COLUMNS INDICATE THE NUMBER OF CANDIDATE TARGETS FROM TRANSFAC, NUMBER OF GENES NOT REGULATED WITH A p-VALUE THRESHOLD OF 0.1, AND NUMBER OF GENES REGULATED BY 1 OR 2 OF THE TRANSCRIPTION FACTORS WITH A p-VALUE THRESHOLD OF 0.1.

# TFs	# genes	0	1	2
1	438	380	58	0
2	115	91	16	8
3	33	26	3	4
4	9	7	0	2
5	2	1	0	1
6	4	4	0	0

targets rather than an inability of the statistic to account for multiple regulation.

To explore this hypothesis, we then compared the CoGAPS inferred expression patterns for each transcription factor to the expression patterns of each of their target genes. In contrast to the simulated data, genes assigned to the prior regulatory set closely follow the inferred expression pattern, while genes removed were not correlated to the inferred patterns. Often, the genes that were eliminated tended to show little expression response across the time course, suggesting that they were not regulated by these TFs in GIST cells. We further explored the expression profile of genes for which the regulatory statistic of eq. 3 was below the 0.1 threshold in the TFs p53, STAT, c-Myc, and Elk-1, whose activity was validated in [1] (Figure 5).

In the case of p53 (Figure 5a), the inferred expression pattern reflects the transitory activity of the TF, decreasing at 9 hours, although the Western blots showed the strongest p53 phosphorylation, indicative of p53 TF activity, at 9-18 hours [1]. While the timing in the CoGAPS inferred pattern differed from the measured transcription factor activity, the genes that the regulatory statistic inferred to be regulated by these TFs tended to have increased expression at 9-18 hours. Similarly, the expression in regulatory targets selected for both STAT3 (Figure 5b) and Elk-1 (Figure 5d) increased more strongly at 6 hours than the CoGAPS inferred pattern, consistent with the Western phosphorylation patterns. Unlike Elk-1 (Figure 5d), the regulatory statistic selected targets for c-Myc (Figure 5c) whose expression consistently decreased in time, agreeing with a lack of recovery of c-Myc activity in later time points.

VI. CONCLUSION

Context-specific transcriptional response plays an important role in biological systems, since evolution has driven reuse of genes and proteins in multiple contexts [13]. One key point where evolution plays this role is in the context-specific targets of transcription factors (TFs). Here, we have introduced a novel statistic, eq. 3, and demonstrated the recovery of context-specific TF targets in simulations and

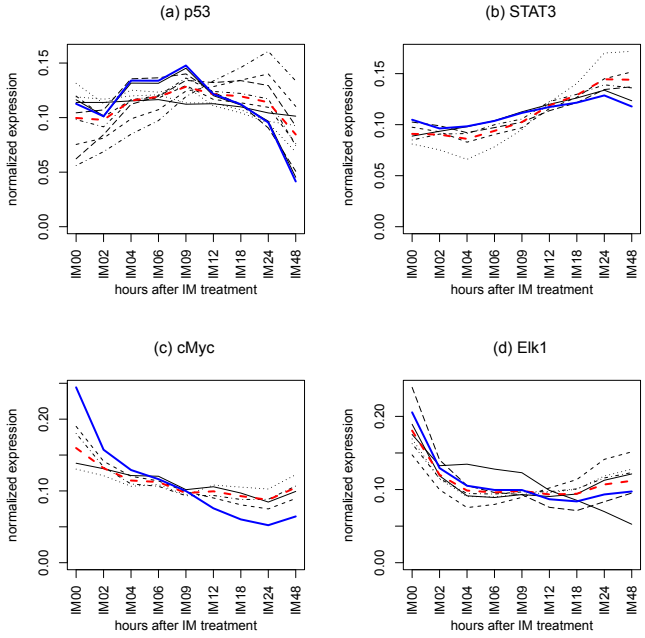

Figure 5. Summary of inferred transcription-factor patterns in GIST data. Normalized, de-noised GAPS inferred expression patterns for each gene with a gene regulatory p-value less than 0.1 (black lines) and their average expression pattern (red), compared to the inferred, normalized pattern for the gene set (blue line). Results are plotted for inferred targets of (a) p53, (b) STAT3, (c) cMyc, and (d) Elk-1.

in imatinib-treated GIST cell lines. Although the summary statistic in eq. 3 could be applied to inferences from any matrix factorization or differential expression algorithm, the sparse atomic prior encoded in CoGAPS facilitates the removal of candidate genes. This is consistent with the fact that the retained TF targets are more robustly tied to the CoGAPS patterns than the removed targets and explains the observed tendency in the statistic to false negatives rather than false positives. Although the statistic formulated here was developed for the CoGAPS algorithm, the regulatory statistic of eq. 3 is general. For example, this statistic may allow similar improvement in the inference of set membership when applied to a matrix factorization that uses the candidate targets in the prior distribution (e.g., [10]).

Because of the MCMC structure, CoGAPS generates uncertainty estimates for the association of each pattern with samples. This identifies patterns with significant strength ($< P_{i\bullet} > / sd(P_{i\bullet}) > 0$) in subtypes of samples (e.g., experimental conditions [14]). As a result, the statistic for set membership of eq. 3 could be extended to infer targets within sub-contexts of the expression data by averaging over only those patterns that are present in the sub-context.

While formally similar to the mSD algorithm [15], CoGAPS naturally accounts for the error distribution of gene expression data, unlike the sparse component analysis. As

demonstrated previously [16], accounting for this uncertainty improves inference of functional gene relationships. Similarly, several databases and binding assays, such as ChiP-seq, commonly provide continuous prior probabilities for TF target membership, rather than the binary membership assumed in eq. 3. The probability information can be incorporated in our statistic by naïve assignment of genes to each of the sets $\mathcal{G}$ according to a selected threshold. However, a more robust estimate for the summary statistic would be obtained by weights in eq. 3 based on prior probabilities.

Although this statistic was implemented on gene expression data measured with microarrays, the algorithm can be extended naturally for expression measured by RNA-seq or even to broader measures of common activity across samples. Moreover, the statistic can be extended easily to handle varying probabilities, as from ChIP-seq measurements, rather than binary probabilities from databases. Computing the summary statistic resulting from permuting the prior probabilities would likely provide a robust null distribution for the modified regulatory statistic. This extension will be particularly valuable when ChIP-seq and expression profiling are performed simultaneously in the desired experimental context. A similar extension of the gene set statistic using small, but non-zero priors for genes that the database does not include in the target set could likewise be used to add genes as candidate targets of the transcription factor. In this case, it may also be necessary to add a penalty term to eq. 3 to avoid over-estimating candidate targets.

ACKNOWLEDGMENT

This work was supported by grants NLM009382 to MFO, CA141053 to EJF, Russian Foundation for Basic Research 11-04-02016-a and JHU Framework for the Future to AVF. We thank Ludmila Danilova for helpful discussions. Code for results presented are available upon request to EJF (ejfertig@jhmi.edu).

REFERENCES

[1] M. F. Ochs, L. Rink, C. Tarn, S. Mburu, T. Taguchi, B. Eisenberg, and A. K. Godwin, "Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data," *Cancer Res*, vol. 69, no. 23, pp. 9125–32, Dec 2009.

[2] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology," *Nat Rev Genet*, vol. 10, no. 10, pp. 669–80, Oct 2009.

[3] H. S, "Eukaryotic transcription factor binding sites–modeling and integrative search methods," *Bioinformatics*, vol. 24, pp. 1325–31, 2008.

[4] A. A. Sharov, S. Masui, L. V. Sharova, Y. Piao, K. Aiba, R. Matoba, L. Xin, H. Niwa, and M. S. H. Ko, "Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data," *BMC Genomics*, vol. 9, p. 269, 2008.

[5] H. M. S. Asif and G. Sanguinetti, "Large-scale learning of combinatorial transcriptional dynamics from gene expression," *Bioinformatics*, vol. 27, no. 9, pp. 1277–83, May 2011.

[6] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM," *Bioinformatics*, vol. 26, no. 12, pp. i237–45, Jun 2010.

[7] G. E. Gonye, P. Chakravarthula, J. S. Schwaber, and R. Vadigepalli, "From promoter analysis to transcriptional regulatory network prediction using PAINT," *Methods Mol Biol*, vol. 408, pp. 49–68, 2007.

[8] A. V. Kossenkov and M. F. Ochs, "Matrix factorisation methods applied in microarray data analysis," *Int J Data Min Bioinform*, vol. 4, no. 1, pp. 72–90, 2010.

[9] E. J. Fertig, J. Ding, A. V. Favorov, G. Parmigiani, and M. F. Ochs, "CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data," *Bioinformatics*, vol. 26, no. 21, pp. 2792–3, Nov 2010.

[10] A. V. Kossenkov, A. J. Peterson, and M. F. Ochs, "Determining transcription factor activity from microarray data using Bayesian Markov chain Monte Carlo sampling," *Stud Health Technol Inform*, vol. 129, no. Pt 2, pp. 1250–4, 2007.

[11] S. Sibisi and J. Skilling, "Prior distributions on measure space," *Journal of the Royal Statistical Society, B*, vol. 59, no. 1, pp. 217–235, 1997.

[12] V. Matys, E. Fricke, R. Geffers, E. Gössling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res*, vol. 31, no. 1, pp. 374–8, Jan 2003.

[13] J. Preston, L. Hileman, and P. Cubas, "Reduce, reuse, and recycle: Developmental evolution of trait diversification," *Am J Bot.*, vol. 98, pp. 397 – 403, 2011.

[14] E. J. Fertig, Q. Ren, H. Cheng, H. Hatakeyama, A. Dicker, U. Rodeck, M. Considine, M. F. Ochs, and C. H. Chung, "Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma," *BMC Genomics*, vol. 13, p. 160, 2012.

[15] T. Gong, J. Xuan, L. Chen, R. Riggans, H. Li, E. Hoffman, R. Clarke, and Y. Wang, "Motif-guided sparse decomposition of gene expression data for regulatory module identification," *BMC Bioinformatics*, vol. 11, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/82

[16] W. G, K. A. V, and O. M. F, "LS-NMF: a modified non-negative matrix factorization algorithm utilizing uncertainty estimates," *BMC Bioinformatics*, vol. 28, p. 175, 2006.