Pacific Symposium on Biocomputing 14:264-275 (2009)

DISSECTING THE INTERFACE BETWEEN SIGNALING AND TRANSCRIPTIONAL REGULATION IN HUMAN B CELLS

KAI WANG[1,2,*], MARIANO J. ALVAREZ[2], BRYGIDA C. BISIKIRSKA[2], RUNE LINDING[3], KATIA BASSO[4], RICCARDO DALLA FAVERA[4], ANDREA CALIFANO[1,2,4,†]

1. Department of Biomedical Informatics, Columbia University, New York, NY, USA

2. Joint Centers for Systems Biology, Columbia University, New York, NY, USA

3. Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada

4. Institute of Cancer Genetics, Columbia University, New York, NY, USA

1. Abstract

A key role of signal transduction pathways is to control transcriptional programs in the nucleus as a function of signals received by the cell via complex post-translational modification cascades. This determines cell-context specific responses to environmental stimuli. Given the difficulty of quantitating protein concentration and post-translational modifications, signaling pathway studies are still for the most part conducted one interaction at the time. Thus, genome-wide, cell-context specific dissection of signaling pathways is still an open challenge in molecular systems biology.

In this manuscript we extend the MINDy algorithm for the identification of post-translational modulators of transcription factor activity, to produce a first genome-wide map of the interface between signaling and transcriptional regulatory programs in human B cells. We show that the serine-threonine kinase STK38 emerges as the most pleiotropic signaling protein in this cellular context and we biochemically validate this finding by shRNA-mediated silencing of this kinase, followed by gene expression profile analysis. We also extensively validate the inferred interactions using protein-protein interaction databases and the kinase-substrate interaction prediction algorithm NetworKIN.

2. Introduction

A key role of signal transduction pathways is to control transcriptional programs in the nucleus as a function of signals received by the cell via complex post-translational modification cascades, thus determining the cell's response to environmental stimuli (see Figure S1 for a schematic description). Their understanding is increasingly crucial in the dissection of human disease and in the identification of therapeutic intervention targets [1], because signaling molecules (e.g., GPCR receptors or tyrosine kinases) are much more effectively targeted by small molecules than transcription factors. Unfortunately, cell-context specific dissection of signaling pathways is still an open challenge

* Current affiliation: Rosetta Impharmatics (Merck & Co., Inc.), Seattle, WA 98109
† email: califano@c2b2.columbia.edu

because of the inherent difficulties in the high-throughput measurement of protein concentration and post-translational modification. As a result, the dissection of signaling pathway is still, for the most part, proceeding one protein-protein interaction at a time [2].

Conversely, availability of large collections of gene expression profiles (GEP) [3] has fostered significant progress in the genome-wide dissection of transcriptional programs [4, 5]. Until recently, GEPs have not been broadly used in the dissection of post-translational interactions. Several GEP-based studies of yeast signal transduction networks have been limited to the identification of gene modules regulated by a small number of regulators, including some signaling proteins [3], or to the reconstruction of signaling pathways using known protein-protein interactions as a topological backbone [6]. In general, however, a cell-context-specific map of the interface between signaling and transcriptional regulatory programs is still an elusive target both in yeast and in higher eukaryotes.

We recently introduced the MINDy algorithm (Modulator Inference by Network Dynamics) for the genome-wide identification of post-translational modulators of transcription factor (TF) activity [7]. MINDy tests whether the conditional mutual information (CMI), $I[TF; t \mid M]$, between a transcription factor TF and a target t, as a function of a modulator M is non-constant. In that case, M is inferred as a candidate post-translational modulator of the TF. Based on this analysis, MINDy can also determine whether the modulator protein will activate or repress the TF-target interaction, resulting in either a positive or negative *mode of action* (MoA). We have biochemically validated four inferred modulators of the transcription factor MYC, including a kinase (STK38), an histone deacetylase (HDAC1), and two transcription factors (BHLHB2 and MEF2B) by shRNA mediated silencing and other biochemical assays [8]. For full details on the MINDy algorithm, its applications as well as limitations, readers are referred to [7, 8] and the Methods section 5.2. In this manuscript, we extend the MINDy algorithm to the genome-wide exploration of the interface between signaling pathways and transcriptional networks in human B cells.

3. Results

3.1. *Network components*

In this work we define the signalome as the compendium of signaling proteins (SP) annotated as protein kinases, phosphatases or cell surface receptors in the Gene Ontology (GO) [9]. The term "transfactome" is borrowed from [10] and is

defined here as the compendium of proteins annotated as transcription factors (TF) in the GO. Only proteins expressed in a set of 254 GEPs from normal and tumor related human B cells were considered in the analysis (see 5.1). A total of 772 SPs and 595 TFs were selected based on these criteria, see Table 1.

Table 1. Selection of signalome and transfactome genes. # indicates the number of genes selected in each category. MF: molecular function; BP: biological process; CC: cellular component

Functional Category		#	GO Categories
Signalome	Kinases	421	Protein kinase activity (MF)
	Phosphatases	113	Phosphoprotein phosphatases (MF)
	Receptors	295	Receptor activity (MF) Cell surface receptor linked signal transduction (BP) Integral to plasma membrane (CC)
Transfactome	Transcription Factors	595	Transcription factor activity (MF)

3.2. *Signalome-transfactome interaction inference*

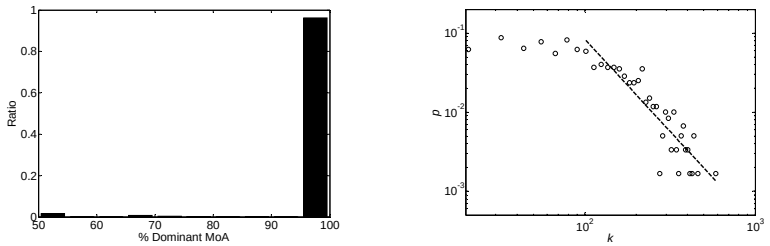
SP-TF interactions were inferred by assessing whether one or more TF-target interactions were modulated by the SP using the CMI test (see 5.2). The complete set of TF-targets modulated by a SP is called the SP's *regulon*, while the set of all TFs modulated by a SP is called the SP's *modulon* (Figure S3). At the genome-wide statistical significance level of 5% (see 5.2), MINDy inferred 44,349 SP-TF interactions. A summary of these results can be found in Table 2. Each SP modulates on average 29.6 TFs in human B cells, and a TF is controlled on average by 38.4 SPs. Some other interesting global properties of the signalome-transfactome interface are summarized in Figure 1 (and more are reported in the supplementary materials). These include: (a) The MoA is consistently inferred from each SP-TF-target triplet supporting a specific SP-TF interaction, even though they are independently tested. i.e., either the positive or the negative MoA is supported by the majority (95 - 100%) of the triplets. This

Signalome / Transfactome

	STK38	*CDK5*	*BCR*	*CSNK1G2*	*PRKDC*	*PTK2B*	*CCL2*	*CDK2*	*TGFBR2*	*CDC25A*	...	*GNRHR*	*HRH1*	*PRKG1*	*ADRB1*	*GLRA1*	*HUS1*	*MPL*	*BMX*	*TACR2*	*DNAJC6*
SMAD3	5	3	9	0	8	2	0	7	8	27	...	0	0	0	0	0	0	0	0	0	0
CREM	0	2	1	2	19	1	0	3	5	5	...	0	0	0	0	0	0	0	0	0	0
ZNF263	0	1	1	11	6	1	10	2	0	3	...	0	0	0	0	0	0	0	0	0	0
MEF2D	4	1	10	3	0	0	4	1	0	0	...	0	0	0	0	0	0	0	0	0	0
ETS1	0	16	1	5	4	4	63	0	0	1	...	0	0	0	0	0	0	0	0	0	0
PHTF1	3	22	0	7	32	0	0	27	8	0	...	0	0	0	0	0	0	0	0	0	0
NR4A1	0	2	0	8	3	0	2	16	0	28	...	0	0	0	0	0	0	0	0	0	0
ATF3	0	3	0	13	5	1	0	1	8	8	...	0	0	0	0	0	0	0	0	0	0
TAF7	12	0	1	0	3	0	7	0	0	3	...	0	0	0	0	0	0	0	0	0	0
ZNF85	0	20	0	9	26	1	0	46	11	15	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
IRF7	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
NFIC	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
SHOX	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
THRB	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
JARID1B	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
MDS1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
FOXI1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
HOXB1	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
NKX2-2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
FOXJ1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Table 2. Summary of MINDy results in the signalome-transfactome inference. Signaling proteins are shown on the columns and sorted in decreasing order with respect to the number of TFs each SP modulates. TFs were shown on the rows and are also sorted in decreasing order with respect to the number of SPs they are under control of. Each cell indicates the number of TF-target interactions modulated by the SP.

is highly biologically consistent, since a given SP is expected to either activate or repress a TF's activity but not to do both at the same time. (b) MINDy-inferred TF-target interactions exhibits the previously observed scale-free like degree distribution [11]. However, while previously inferred regulatory networks included only static (i.e. not SP-modulated) TF-target interactions, MINDy inferred network includes conditional interactions as well.



(a) **(b)**

Figure 1. Properties of the signalome-transfactome network inferred by MINDy. (a) Histogram of the proportion of dominant MoA for all SP-TF interactions. Plotted on the x-axis is the percent of MINDy inferred SP-TF-*t* triplets supporting the dominant MoA for a SP-TF interaction. If all triplets support a positive MoA, for instance, the fraction would be 1; if 50% of the triples support a positive MoA and 50% a negative one, the fraction is 0.5. (b) Degree distribution of the MINDy inferred TF-targets network.

3.3. *In-silico validation*

To benchmark the accuracy of MINDy inferences, a set of gold standard SP-TF interactions in human B cells would be required. Compiling such a reference set is relatively difficult because: (a) experimentally validated protein-protein interactions in databases are still very sparse, especially for transient interactions (e.g., kinase-substrate); (b) MINDy-inferred modulators can be either direct (i.e. a physical SP-TF interaction), or pathway-mediated (i.e. SP upstream of the TF in a signaling pathway). While the former could be represented in existing databases, the latter are poorly characterized; and (c) while MINDy inferred SP-TF interactions are highly cell–context specific (to human B cells), human protein-protein interactions have been validated in highly heterogeneous or even artificial (e.g., yeast two-hybrid, Y2H) cellular contexts. To address (a) and (b) we benchmarked MINDy with the following datasets:

Protein-protein interactions (PPIDB): We collected all known human PPIs from high-quality public databases including HPRD [12], BIND [13], DIP [14] and IntAct [15]. These have been experimentally assessed, either in single biochemical assays, or by high throughput techniques such as Y2H. Compared

to the 772 SPs and 595 TFs analyzed by MINDy, these datasets cover 428 SPs and 141 TFs (i.e., 13.1% of the MINDy search space).

Kinase-substrate interactions (KSIDB): We also included kinase-substrate interactions inferred by the experimentally-validated algorithm NetworKIN [16]. This algorithm utilizes information from consensus motifs on the kinase catalytic sites, substrate phosphorylation sites assessed by Mass Spectrometry, as well as cellular context and curated pathways. Such information is completely orthogonal to that used by MINDy (i.e. strictly GEP). Therefore enrichment of their common predictions can be used to assess MINDy's validity, as false positives from the two methods should not be correlated if either of them makes random predictions. Due to the limited number of kinase families for which consensus motifs are known, NeworKIN covers only 74 of 772 MINDy SPs and 240 of 595 MINDy TFs (i.e., 3.9% of the MINDy search space).

Table 3 offers a comparative view of MINDy inferred SP-TF interactions as well as those from PPIDB and KSIDB. Due to the higher coverage of GEP, MINDy covers a much larger space (~7-fold) than the other two data sources combined and may thus provide important information for SP-TF interactions that cannot be studied using other methods.

Table 3. Summary of predictions made by MINDy, PPIDB interactions and KSIDB predictions.

	MINDy	**PPIDB**	**KSIDB**	**PPIDB + DSIDB**
Interactions	9017	434	1105	1506
No. SPs	772	428	74	439
No. TFs	595	141	240	291
% Coverage	100%	13.1%	3.9%	15.8%
% Prediction	2.0%	0.7%	6.2%	1.2%

We first tested the hypothesis that modulator sets affecting the same TF should be more physically inter-connected than random genes (as they are more likely to cluster within signaling pathways). This was done by counting the number of PPIDB and KSIDB interactions among MINDy-inferred SPs that modulate the same TF, compared to that among the same number of SPs chosen at random. Of 595 tested TFs, 400 (67.2%) show significant enrichment among their inferred modulator SPs by Fisher's Exact Test (FET) (i.e. p-value < 0.05/595). This suggests that SPs modulating the same TF tend to cluster in physical pathways.

Next, we measured the overlap between MINDy-inferred interactions and those in PPIDB and KSIDB. Since interactions predicted by MINDy include both direct and pathway-mediated interactions, we expanded the interactions in

our reference databases with additional ones for which a linear chain of PPIDB (undirected) and/or KSIDB (directed) interactions exists between a SP and a TF, as first proposed in [6]. We plotted the precision of MINDy (i.e. percent of MINDy predictions in the extended database) as a function of the MINDy *p*-value threshold (i.e. the threshold that controls the MINDy false positive rate). As expected (Figure 2), MINDy precision and number of predictions respectively increase and decrease as the *p*-value threshold becomes more stringent. Based on this plot, we selected the inflection point, corresponding to the MINDy *p*-value of 7.5×10^{-7}, as an optimal p-value cutoff, resulting in 9,017 inferred interactions. Of these, 3,739 (41.5%) are supported by either direct or pathway-mediated PPI interactions. The probability that such an overlap occurs by chance is $< 2.9\times10^{-89}$ by FET. A similar analysis using KEGG [17] and GenMAPP [18] annotated pathways also yielded statistically significant enrichment (results reported in the supplementary materials). These highly significant enrichments suggest that MINDy is able to recapitulate known interactions at the signalome-transfactome interface.

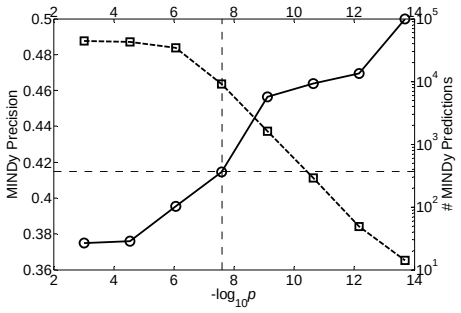


Figure 2. MINDy benchmarking using PPIDB and KSIDB interactions. X-axis shows the $-\log_{10}$ of MINDy *p*-value. Precision of MINDy as a function of *p*-value cutoffs is plotted using the solid line on the left y-axis. The dashed line plots the number of MINDy predictions at each *p*-value cutoff on the right y-axis. Dotted lines indicate the optimal *p*-value cutoff selected in the text.

3.4. *Experimental validation*

Given the exceedingly large set of MINDy prediction space, a systematic experimental validation plan is clearly impractical. Instead, we decided to validate the candidate modulator controlling the largest number of TFs. This is a serine-threonine kinase STK38 [19] that is poorly characterized in the literature, a rather surprising fact since MINDy infers it as a modulator of 303 TFs. We silenced STK38 by lentiviral-vector mediated shRNA expression in the ST486 Burkitt's line, and verified that the STK38 protein level significantly decreased at 60h after transduction (Figure 3a).

MINDy regulon analysis: We first tested whether the 1219 TF-targets that were inferred as STK38 modulated via the 303 TFs (the STK38 *regulon*) were indeed

affected by STK38 silencing. Gene Set Enrichment Analysis (GSEA) [20] confirmed a very significant enrichment ($p < 10^{-4}$) of the STK38 regulon in genes that were differentially expressed after silencing (Figure 3b). Moreover, the set of 303 STK38 modulated TFs (the STK38 *modulon*) was not enriched in differentially expressed genes (Figure 3c), suggesting that the differential regulation of the STK38-regulon is not mediated by transcriptional activation/repression of the TFs in its modulon, consistent with the MINDy model which is designed to identify post-translational interactions.

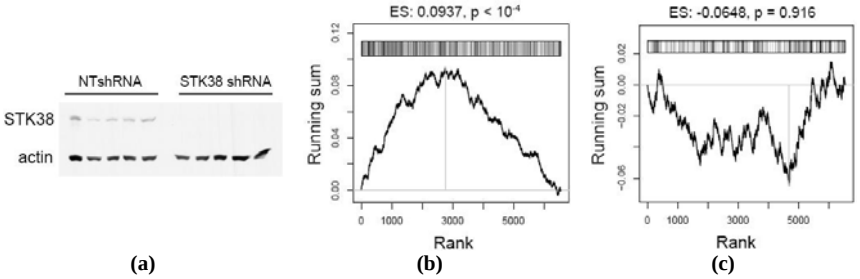


Figure 3. (a) STK38 and β-actin immunoblot on total cell lysates from ST486, at 60h after lentiviral vector-mediated non-target control (NTshRNA) and STK38 shRNA expression. (b) GSEA enrichment of STK38 regulon and (c) STK38 modulon among differentially expressed genes, following STK38 silencing. GSEA test is performed as in [20] using default parameters. Genes were ranked along the x-axis based on their differential expression *p*-value based on two-sample t-test. The "bar code" on top indicates the position of STK38 regulon (for panel b) and modulon (for panel c) genes in the ranked list. The black intensity of each bar is proportional to the local density of surrounding bars. False positive rate is estimated by permutation test, in which 10^4 null scores were obtained by selecting at random the same number of genes as STK38 regulon (for panel b) and modulon (for panel c).

Modulator comparison: When compared to the GSEA enrichment of the other SP regulons, the STK38 regulon scored 8th out of 772. Additionally, the most significantly enriched MINDy regulon was that of the SP having the highest overlap with the STK38 regulon (CDC2L5, with 506 common regulon genes out of 1039), further suggesting that STK38 silencing affects the MINDy-inferred STK38 regulon in a highly specific way. Indeed, based on this high overlap, we hypothesize that CDC2L5 is directly downstream of STK38 so that the higher enrichment is justified by a more specific regulon (1039 for CDC2L5 vs. 1219 for STK38). Further experimental data are required to confirm this hypothesis.

STK38-TF interaction validation: To test MINDy's ability to infer individual SP-TF interactions, we selected 257 TFs in the STK38 modulon with more than 150 STK38 modulated targets. The 150-target threshold was chosen to ensure sufficient statistical power of the GSEA test. We then tested whether the

MINDy inferred targets of each TF were enriched in differentially expressed genes after STK38 silencing. The analysis shows that 78 out of 257 target-sets (30%) are significantly enriched at 5% false positive rate, corresponding to a false discovery rate of 16% (i.e. 257×0.05/78). This is a very high percent because (a) STK38 is so pleiotropic that individual TFs may not be affected at all due to combinatorial regulation effects, and (b) modulation was not inferred by MINDy in the ST486 cell line but rather from a collection of 17 distinct B cell phenotypes (see 5.1). Hence, some TF's targets may not be affected because key co-factors, signals, or effectors are missing.

3.5. *Signalome-transfactome interaction network*

Table 4 and Figure S8 summarize the first genome-wide *in silico* map of direct and pathway mediated SP-TF interactions. The SPs are clustered based on their modulon overlaps (see 5.4). As expected, the interactions are sparse in general, but tightly clustered into modules representing functionally coherent SP-sets associated with known biological processes. This further validates MINDy's ability to characterize the modulon of arbitrary signaling genes and to annotate their functions.

#	*N*	*Annotation*	*p*
1	26	Protein biosynthesis	2.6×10^{-3}
		Cell homeostasis	2.6×10^{-3}
2	23	Cell cycle	3.9×10^{-5}
		Apoptosis	2.1×10^{-2}
3	21	Cell surface receptor linked signal transduction	2.8×10^{-4}
4	20	G1 to S cell cycle reactome	1.9×10^{-3}
5	14	Cell-cell adhesion	9.0×10^{-4}
		Cell motility	1.3×10^{-3}
6	12	G-protein coupled receptor class B	8.0×10^{-3}
		G-protein coupled receptor class C	8.0×10^{-3}
7	11	Gap junction	6.6×10^{-3}
		Integrin-mediated cell adhesion	1.4×10^{-2}
8	10	G-protein coupled receptor class A	7.3×10^{-5}

Table 4. Modules of SPs identified by clustering the signalome. *N*: module size; *p*: fisher's exact test *p*-value. Only top 8 modules with size equal to or greater than 10 are listed. Significant pathways were selected using *p*-value cutoffs specified in 5.4.

Finally, the most likely topology of the interface between SPs and TFs in human B cells, supported by significant physical evidence from both literature and novel biochemical assays, was obtained by mapping the MINDy predictions onto direct interactions in PPIDB and KSIDB, producing a hybrid network

depicted in Figure 4. The connectivity of nodes in this network follows a power-law distribution (see Figure 4 inset), suggesting that the network is scale-free. MINDy-inferred signaling interactions supported by KSIDB appear to be more pleiotropic than those supported by PPIDB, perhaps due to the nature of the two evidence sources: NetworKIN makes predictions on well-studied kinases (with known consensus motif), whereas high-throughput PPI measurements in the database, e.g. Y2H, tend to have less selection bias and a high false negative rate.

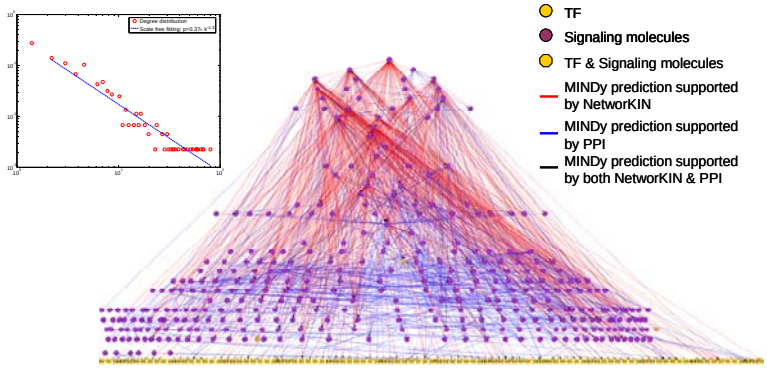


Figure 4. Visualization of the signalome-transfactome network by integrating MINDy predictions with PPIDB and KSIDB interactions. Two types of interactions are represented in the network: 1) MINDy predicted SP-TF interactions supported by PPIDB or KSIDB (i.e. modulatory interactions predicted by MINDy that have physical interaction evidence); 2) MINDy predicted SP-SP interactions supported by PPIDB or KSIDB (i.e. between modulators predicted by MINDy of the same TF, and are supported by physical interaction evidence). Depending on the source of evidence, these interactions can be either un-directed (supported by known PPIs), or directed (supported by NetworKIN, i.e. kinase→substrate). The inset figure shows the degree distribution of this network, with a linear fitting in the log-log space indicating a scale-free topology.

4. Discussion

In this manuscript we have provided both computational and experimental evidence suggesting that the MINDy algorithm can be effectively used to map physical and pathway-mediated post-translational interactions between signaling proteins and transcription factors, using only large GEP datasets. Due to lack of appropriate high-throughput technologies, such as microarray expression profiles and ChIP-Chip/Seq assays, dissection of post-translational interactions is lagging significantly behind that of their transcriptional counterpart. As a result, algorithms providing high-accuracy, cell-context specific hypotheses for biochemical validation may significantly improve our ability to elucidate post-translational processes and their effect on transcriptional networks. Specifically,

we have shown that MINDy predictions are highly enriched in experimentally validated interactions and that silencing the most pleiotropic modulator, STK38, produces expression profiles that are highly consistent with the inferred STK38 modulon and regulon.

Such a network topology can be interrogated to address specific biological questions, such as (a) what are the signaling proteins that control a specific transcriptional program (i.e. TF)? (b) What are the shortest paths through which a signaling protein may affect the activity of a TF? And (c) what signaling proteins are upstream/downstream of other signaling proteins. Taken together, this represents the first genome-wide computational analysis of the interface between signaling and transcriptional networks. The combination of these results and those from *in vivo* experiments may significantly improve our understanding of the role of cellular signaling in the regulation of transcriptional programs and provide new targets for therapeutic intervention.

Lastly, another insight that can be gleaned from the signalome-transfactome network reconstructed by MINDy is the specificity of the signaling genes in terms of their ability to regulate the transcriptional response of a cell. In Table 2 shown earlier, signaling genes on columns to the left are more pleiotropic, whereas those to the right are more specific, with respect to the number of TFs (i.e. distinct transcriptional programs) they control. Similarly, TFs on top rows are controlled by broader signaling pathways than those on the bottom rows. Since signaling proteins are often selected as drug target (e.g. by small molecule compound), these results can provide guidance to the selection of intervention point that has the least side effect. Specifically, one may want to target a kinase that is very specific, so that it causes less cross-talk with other transcriptional programs not intended to be affected by the drug molecule.

5. Methods

5.1. *Gene expression profile dataset*

254 GEP were generated using the Affymetrix HG-U95Av2 GeneChip® System (~12,600 probe sets) from a collection of normal and tumor related B cell samples. Probe sets with expression mean $\mu < 50$ and standard deviation $\sigma < 0.3\mu$, were excluded as non-informative, leaving 8680 probe sets. Further details on the GEP dataset can be found in the supplementary materials.

5.2. *MINDy Analysis*

Given a triplet (TF, M, t), with ($t \neq TF$ and $t \neq M$), MINDy assesses whether the CMI, $I[TF; t \mid M]$, is constant as a function of M. Assuming that the CMI is a monotonic function of M,

this can be efficiently tested by measuring $\Delta I = I\left[TF;t \mid M \in L_m^+\right] - I\left[TF;t \mid M \in L_m^-\right] \neq 0$, where L_m^+ and L_m^- represent two subsets including the 35% of the samples where M is respectively most and least expressed. The *p*-value corresponding to a specific ΔI is obtained by permutation tests, and Bonferroni corrected for the total number of tested modulator-target pairs. Significant triplets are further pruned if there exists a third gene, *x*, such that $I\left[TF;x\right] \geq I\left[TF;t\right]$ and $I\left[t;x\right] \geq I\left[TF;t\right]$ in both $L_m^{\pm}$, indicating an indirect relationship between TF and t, mediated by *x*, as suggested by the Data Processing Inequality [11]. Readers are encouraged to refer to the supporting online materials and [7, 8] for more details on MINDy.

For each TF in the transfactome, MINDy first identifies the set of candidate modulators among all SPs whose expression profiles are independent of that of the TF. All other genes are then tested as candidate targets *t* of the TF. Each TF-SP-*t* triplet is then analyzed by the MINDy test at a 5% statistical significance level, after asymptotic Bonferroni correction for multiple testing (i.e. using a *p*-value cutoff $= 0.05/\left(N_m \times N_t\right)$, where N_m is the number of candidate modulators and N_t the number of candidate target genes). The analysis was run at the Affymetrix probe set level, and duplicated probe sets mapping to same gene were merged *a posteriori*.

5.3. *Lentiviral mediated STK38 knock-down*

Human embryonic kidney 293T and Burkitt's lymphoma cell line ST486 were maintained in DMEM and IMDM, respectively. All cell culture media were supplemented with 10% FBS (Invitrogen) and antibiotics. Supernatants for the lentiviral vector containing the STK38 shRNA (TRCN0000010216, Sigma) and non-target control shRNA (SHC002, Sigma) were produced in 293T cells. 5 independent samples of ST486 cells ($2x10^6$ cells/ml) were transduced with viral supernatants for either STK38 shRNA or non-target control shRNA. Transduction was performed by centrifugation at 450xg for 2h with supernatants supplemented with 8 μg/ml polybrene. Total RNA was extracted 60h after transduction and prepared for gene expression profiling according to Affymetrix's protocol.

5.4. *Signalome clustering*

Signaling proteins are clustered based on the similarity of their modulon inferred by MINDy. Specifically, each signaling protein is associated with a vector $T_j = \{t_1, t_2, ..., t_N\}$ where t_i, $i = 1, 2, ..., N$ is the number of targets of TF i modulated by the signaling protein j, and N is the total number of TFs in the transfactome. We then performed hierarchical clustering of the signaling genes using average linkage method and Pearson correlation between these vectors as the distance metric. Signaling genes were assigned into modules when linkage stopped at 70% of the maximal linkage score. Modules consisting of more than 10 genes are subsequently queried for over-represented pathways against the background of all SPs. Only pathways with ≥5 SPs are searched, including 286 GO biological process categories and 44 KEGG/GenMAPP pathways. Enrichment is calculated using FET with *p*-value cutoff set to 1/286 for GO and 1/44 for KEGG/GenMAPP.

5.5. *Supporting online materials*

Supplementary materials, including lists of PPIDB and KSIDB interactions, MINDy predictions and MINDy software are available at: http://wiki.c2b2.columbia.edu/califanolab/PSB2009/.

Acknowledgements

This work is supported by the NCI (R01CA109755), the NIAID (R01AI066116), and the National Centers for Biomedical Computing NIH Roadmap Initiative (U54CA121852).

References

1. Gough, N.R., Signal Transduction Pathways as Targets for Therapeutics. 2001. p. pe1-.
2. Zhang, Y., et al., Time-resolved Mass Spectrometry of Tyrosine Phosphorylation Sites in the Epidermal Growth Factor Receptor Signaling Network Reveals Dynamic Modules. Mol Cell Proteomics, 2005. **4**(9): p. 1240-1250.
3. Roberts, C.J., et al., Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science, 2000. **287**: p. 873-880.
4. Gardner, T.S., et al., Inferring genetic networks and identifying compound mode of action via expression profiling. Science, 2003. **301**(5629): p. 102-5.
5. Basso, K., et al., Reverse engineering of regulatory networks in human B cells. Nat Genet, 2005. **37**(4): p. 382-390.
6. Steffen, M., et al., Automated modelling of signal transduction networks. BMC Bioinformatics, 2002. **3**(34).
7. Wang, K., et al., Genome-Wide Discovery of Modulators of Transcriptional Interactions in Human B Lymphocytes. Lecture Notes in Computer Science, 2006. **3909**: p. 348 - 362.
8. Wang, K., et al., Genome-wide Identification of Post-translational Modulators of Transcription Factor Activity in Human B Cells. Submitted, 2008.
9. Ashburner, M., et al., Gene Ontology: tool for the unification of biology. Nat Genet, 2000. **25**(1): p. 25.
10. Foat, B.C., R.G. Tepper, and H.J. Bussemaker, TransfactomeDB: a resource for exploring the nucleotide sequence specificity and condition-specific regulatory activity of trans-acting factors. Nucl. Acids Res., 2007: p. gkm828.
11. Margolin, A., et al., ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics, 2006. **7**(Suppl 1): p. S7.
12. Peri, S., et al., Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res, 2003. **13**(10): p. 2363-71.
13. Bader, G.D., D. Betel, and C.W. Hogue, BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res, 2003. **31**(1): p. 248-50.
14. Xenarios, I., et al., DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res, 2002. **30**(1): p. 303-5.
15. Hermjakob, H., et al., IntAct: an open source molecular interaction database. Nucleic Acids Res, 2004. **32**(Database issue): p. D452-5.
16. Linding, R., et al., Systematic Discovery of In Vivo Phosphorylation Networks. Cell, 2007. **129**(7): p. 1415-1426.
17. Kanehisa, M., et al., From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res, 2006. **34**(Database issue): p. D354-7.
18. Dahlquist, K.D., et al., GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet, 2002. **31**(1): p. 19-20.
19. Tamaskovic, R., S.J. Bichsel, and B.A. Hemmings, NDR family of AGC kinases--essential regulators of the cell cycle and morphogenesis. FEBS Lett, 2003. **546**(1): p. 73-80.
20. Subramanian, A., et al., From the Cover: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS, 2005. **102**(43): p. 15545-15550.