

A Context-Specific Network of Protein-DNA and Protein-Protein Interactions Reveals New Regulatory Motifs in Human B Cells

Celine Lefebvre¹, Wei Keat Lim¹, Katia Basso², Riccardo Dalla Favera²,
and Andrea Califano^{1,*}

¹ Center for Computational Biology and Bioinformatics, Columbia University
1130 St Nicholas Avenue, 9th Floor, New York NY 10032
califano@c2b2.columbia.edu

² Institute of Cancer Genetics, Columbia University
1130 St Nicholas Avenue, New York NY 10032

Abstract. Recent genome wide studies in yeast have started to unravel the complex, combinatorial nature of transcriptional regulation in eukaryotic cells, including the concerted regulation of proteins involved in complex formation. In this work, we use a Bayesian evidence integration framework to assemble an integrated network, including both protein-DNA and protein-protein interactions, in a specific cellular context (human B cells). We then use it to study common interaction motifs between protein complexes and regulatory programs, using an enrichment analysis approach. Specifically, we compare the frequency of mixed interaction motifs in the real network against random networks of equivalent connectivity. These motifs include sets of target genes regulated by multiple interacting transcription factors, and gene sets encoding same complex proteins regulated by different transcription factors.

Keywords: combinatorial regulation / evidence integration / human B cells / naïve Bayes / network motifs.

1 Introduction

Dissecting transcriptional regulation pathways in mammalian cells is an important step towards the elucidation of normal and disease-related cellular processes. Due to its simpler organization, yeast has so far provided an excellent model organism for the study of eukaryotic cellular networks, offering an initial basis to understand their dynamic complexity. Recently, motifs analyses in yeast networks combining Protein-DNA (P-D) and Protein-Protein (P-P) interactions revealed a trend towards co-regulation and complex formation in lower eukaryotes [1,2], showing that the integration of different interaction types helps elucidate the interface between transcriptional regulation and protein complex formation.

* Corresponding author.

Similar models have not yet become available for higher eukaryotes, including *Homo sapiens*, where transcriptional regulation, complex-formation, and transient protein-protein interactions networks have been studied in isolation and without cell context specificity, for instance by yeast two-hybrids [3,4]. Here, we propose a Bayesian evidence integration framework for network inference, which integrates a variety of generic and context specific experimental clues about P-P and P-D interactions - such as a large collection of B cell expression profiles - with inferences from different reverse engineering algorithms, such as GeneWays [5] and ARACNE [6]. This type of Bayesian learning was previously successful in inferring P-P interactions in yeast [7] and in human [8]. The resulting network is then used as a model to study the interface between regulatory programs and protein complexes.

We first analyzed the enrichment of simple three gene motifs involving both P-P and P-D interactions in our network and then combined them into larger composite motifs to identify combinatorial regulation mechanisms.

2 Material and Methods

2.1 Naïve Bayesian Evidence Integration

The Bayesian evidence integration model applies the Bayes theorem to compute the posterior odds that a specific interaction exists (O_{post}) as the product of the prior odds (O_{prior}) and of a likelihood ratio (LR) [7]:

$$O_{post} = LR \cdot O_{prior} \quad (1)$$

The prior odds are defined as the average odds that two random gene products are involved in an interaction and can be calculated as:

$$O_{prior} = \frac{P(I)}{P(\bar{I})} = \frac{P(I)}{1 - P(I)} \quad (2)$$

where $P(I)$ is the probability that two random gene products are involved in an interaction and $P(\bar{I})$ is the probability that they are not. The posterior odds of a specific interaction are defined as the ratio of the probabilities that two specific gene products, g_x and g_y , are respectively involved or not involved in an interaction, conditional to the presence of N different clues, $c_1 \dots c_N$:

$$O_{post} = \frac{P(I_{xy}|c_1 \dots c_N)}{P(\bar{I}_{xy}|c_1 \dots c_N)} \quad (3)$$

Such clues could include, for instance, the correlation of the two genes' expression profiles, the results of specific experimental assays, the functional categorization of the gene pair, etc. Similarly, the LR is defined as:

$$LR(c_1 \dots c_N) = \frac{P(c_1 \dots c_N|I_{xy})}{P(c_1 \dots c_N|\bar{I}_{xy})} \quad (4)$$

In the Naïve Bayes Classifier (NBC) model, the clues are assumed to be statistically independent. Then, the LR can be computed as the product of individual LR from the respective datasets:

$$LR(c_1 \cdots c_N) = \prod_{i=1}^{i=N} LR(c_i) = \prod_{i=1}^{i=N} \frac{P(c_i|I_{xy})}{P(c_i|\bar{I}_{xy})} \quad (5)$$

A useful property of the NBC model is that performance does not significantly deteriorate if weak dependencies among the clues exist. Under this assumption, the posterior odds of a specific interaction can be calculated as:

$$O_{post} = \prod_{i=1}^{i=N} \frac{P(c_i|I_{xy})}{P(c_i|\bar{I}_{xy})} \cdot O_{prior} \quad (6)$$

O_{prior} can be estimated from prior knowledge on the number of expected P-P interactions or P-D interactions in a cellular context, while the LR s are estimated by counting how many times a specific clue is observed in a positive and negative *gold standard* set. A positive gold standard set should include only gene product pairs that are known to interact, while a negative gold standard set should include only gene product pairs that are known not to interact. O_{post} , computed as the product of these two values, is related to the probability of an interaction to be true as $P_{post} = O_{post}/(O_{post} + 1)$, then achieving a posterior probability of at least 50% is equivalent to achieve $O_{post} \geq 1$ or $LR \geq 1/O_{prior}$.

2.2 Gold-Standard Sets for P-P Interactions

To generate a positive gold standard set (GSP) for P-P interactions, we extracted 25,642 human P-P interactions from HPRD [9], 7,862 from IntAct [10], 4,812 from BIND [11], and 868 from DIP [12], originating from low-throughput, high quality experiments. This resulted in a GSP set of 34,842 unique P-P interactions involving 7,323 genes (28,542 interactions for 6,953 genes after homodimers removal). Based on an estimate for the total number of P-P interactions of 300,000 in a human cell, among 22,000 proteins [3], the prior odds for an interaction is approximately 1 in 800 ($300,000/(22,000^2/2 - 300,000)$). This implies that any protein pair with a $LR \geq 800$ has at least a 50% probability of being involved in an interaction. Generating a negative gold standard set (GSN) is somewhat more complicated because negative interaction examples are not easily identified from the literature. Thus, based on a previous analysis [13], we classified the Gene Ontology (GO) terms into four subcellular compartments (cell periphery and exocytic pathway, cytoplasm, mitochondria and nucleus), and mapped human genes into those compartments. Then, for each compartment pair, we computed the enrichment of protein pairs known to interact (from the GSP) using a Fisher exact test (FET). This revealed compartment pairs that are more likely to host proteins involved in a P-P interaction. Obviously, all pairs where the two compartments were identical (e.g., nucleus-nucleus) showed enrichment.

However, interestingly, the pair cytoplasm-mitochondrion also showed enrichment. The GSN was then defined by using all proteins from non-enriched cell compartment pairs. Note that we also excluded nucleus-mitochondrion protein pairs in the GSN, as the FET was borderline. This resulted in a GSN with 18,359,948 candidate non interacting gene pairs. As could be expected, the GSN had a small overlap with the GSP (1,890 pairs) reflecting the heuristic nature of the approach used to identify negative interactions. However, the overlap is much smaller than expected by chance thus validating that the method provides a relatively good first order approximation of non interacting protein pairs GSN. For our subsequent analysis, we removed from the GSN all the pairs that were also present in the GSP.

2.3 Gold-Standard Sets for P-D Interactions

Defining a realistic GSP for P-D interactions is much more difficult, as the amount of biochemically validated data is orders of magnitude smaller. We thus decided to focus on a very well-studied transcription factor (TF), MYC for which extensive binding data was collected. The GSP was thus defined as a set of 1,041 B cell specific MYC target genes collected from the MYC database [14] and the GSN as its genomic complement. This allowed us to estimate the prior odds for a MYC P-D interaction to be 1 in 21. This causes two problems: First this is likely an underestimate of the total number of MYC targets in a B cell, thus resulting in a corresponding underestimate of the *LR* for MYC interactions. Second, this *LR* should not be used for other TFs that may have a smaller or larger number of targets. However, since this data is not available, we used this value as a first order approximation from all TFs. The *LR* can be iteratively corrected on a TF by TF basis either by estimating the number of actual targets (e.g. by using general properties of the network connectivity [2,15]) or as additional biochemical evidence emerges, such as from ChIP-Chip data.

2.4 Gene Expression Profiles

A collection of 254 gene expression profiles was used, representing 27 distinct cellular phenotypes derived from populations of normal and neoplastic human B lymphocytes [16]. Gene expression profiles were collected using the Affymetrix HG-U95A GeneChip®System (approximately 12,600 probe sets). Expression measurements were normalized using MAS5.0, and probe sets with absolute expression mean < 50 and coefficient of variation < 0.3 , were considered non-informative and were excluded a-priori from the analysis, leaving 7,476 probe sets [15]. We computed the mutual information (MI) between the 7,896 probes (6,083 genes) passing this threshold. Mutual Information [6] is an optimal measure of statistical dependence in a non linear setting. After applying a threshold ($MI \geq 0.069$), corresponding to a p-value of 10^{-7} , we identified 4,711,682 statistically significant MIs between the 6,083 genes. The highest MI among all the probe-set pairs corresponding to a gene pair was used when multiple probes were present in the set.

2.5 Information Content

As previously described [17], we computed the information content of a Gene Ontology (GO) term [18] as follow:

$$I(go_n) = \log_2 \frac{k(go_n)}{\bigcup_{i=1}^m k(go_i)} \quad (7)$$

where go_n represent a GO term, $k(go_n)$ the gene set annotated by go_n and m the number of annotations in the biological process ontology.

2.6 Transcription Factor Classification

To identify human transcription factors, we selected the human genes annotated as "transcription factor activity" in Gene Ontology and the list of transcription factors (TFs) from Transfac Professional [19]. From this list, we removed general TFs (e.g. stable complexes like polymerases or TATA-box-binding proteins), and added some TFs not annotated by GO, producing a final list of 1,722 TFs, from which 632 were on the filtered microarray gene set.

2.7 GeneWays

GeneWays is a computer system designed for automatic analysis of literature data to extract knowledge about molecular interactions [5]. It provides a list of gene pairs associated with a keyword (action), defining the interaction type, and a score between -1 and 1.

2.8 ARACNE

ARACNE is an information-theoretic method for identifying transcriptional interactions between gene products using microarray expression profile data [6]. ARACNE has proven effective in identifying transcriptional targets in complex mammalian networks [15]. We used the bootstrapping version of ARACNE with a list of 632 transcription factors.

2.9 Motifs Enrichment

The combined P-D/P-P interaction network is represented as two independent graphs where the nodes are genes products and a directed or undirected edge represents respectively a P-D or a P-P interaction. Directed edges point from a transcription factor to its target. Randomized networks were built to have the same connectivity properties as the real network. Specifically, the randomized networks have identical distribution for the following properties: (a) P-P interaction degree (number of P-P interactions) per node, (b) P-D interaction in-degree (number of edges pointing to the node) and (c) P-D interaction out-degree (number of edges originating at the node). Randomized networks with this connectivity constraint were built with the *igraph* library of the statistical

software R. Statistical significance of motif enrichment in the real network was obtained by computing the zscore:

$$zscore = \frac{N_{real} - mean(N_{random})}{sd(N_{random})} \quad (8)$$

where N_{real} is the number of motifs in the real network, and $mean(N_{random})$ and $sd(N_{random})$ are the mean and the standard deviation of the number of motifs in 1,000 randomized networks.

3 Results and Discussion

We used separate Naïve Bayes classifiers to predict P-P and P-D interactions. This requires positive and negative Gold Standard datasets (GSP and GSN) for both P-P and P-D interactions to evaluate the likelihood ratio (LR) of each evidence source. Construction of these datasets is described in the previous section. Note that, as for similar approaches [7,8], we consider fixed P-P and P-D priors in this paper. This is only a first order approximation and it will need to be adjusted in a Protein and TF-specific way, as additional evidence is collected, since cellular network connectivity appears to be approximated by a power-law [2,15].

3.1 P-P Interactions

To infer P-P interactions, we integrated the following P-P interaction evidence: (a) non human interactions for four eukaryotic organisms, (b) two yeast two-hybrid (Y2H) datasets, (c) the GeneWays literature datamining algorithm [5], (d) the Gene Ontology biological process annotations [18], and (e) gene co-expression data from a large collection of 254 B cell expression profiles [15]. Each evidence source was represented as categorical data (continuous values were binned as necessary) and used to compute a LR based on the GSP and GSN data as further described (Table 1). Note that we only considered LR greater than 1 for the different evidences.

Non human interactions for four eukaryotic organisms and human Yeast two-hybrid (Y2H): We extracted putative P-P interactions clues from IntAct and BIND for the three model organisms *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus* and from IntAct, BIND and MIPS [20] for *Saccharomyces cerevisiae*. We defined four different groups of predicted P-P interactions, one for each organism, by mapping model organisms' genes to human genes using the Inparanoid database that describes eukaryotic orthologous clusters [21]. As these four sources contain redundant information, we chose to combine them, together with human interactions extracted from the two Y2H experiments, in one non-redundant source. In this final group, interactions were classified according to the number of evidence sources supporting them (from 1 to 5) for computing a LR . As expected, the LR distribution shows that interactions between genes that are supported by more than one data source are

likely to predict P-P interactions between the corresponding orthologous genes in *Homo sapiens*.

GeneWays literature datamining algorithm: By studying the action keyword enrichment for 6,904 P-P interactions in the GSP (from the HPRD), which were also reported by GeneWays, we identified 19 action keywords associated with P-P interactions. These include the following: *assemble, associate, bind, coexpress, coimmunoprecipitate, colocalize, connect, coprecipitate, copurify, dephosphorylate, dissociate from, form, form a complex, immunoprecipitate, interact, recruit, required for, synergize* and *ubiquitinate*. Enrichment was computed with a fisher exact test (p -value $< 10^{-3}$). This list allowed us to extract 25,985 putative GeneWays P-P interactions among 5,797 genes. These were further classified in two groups according to their score ($s \leq 0$ and $s > 0$, respectively). The LR was computed independently for the two groups.

Gene Ontology biological process annotation: It was also observed that interacting proteins tend to share the same biological process [22]. Thus, GO

Table 1. P-P and P-D interaction evidence and Likelihood Ratio (LR)

	Evidence	Bins	LR
P-P Interaction Integration	Protein-Protein Interactions	1	33
		≥ 2	848
	GeneWays	≤ 0	165
		> 0	404
	Gene Ontology	< 6	13
		6-7	29
		7-8	39
		8-9	95
		9-10	174
		10-11	203
		11-12	321
	> 12	496	
	Mutual Information	0.22-0.27	2
0.27-0.32		4	
0.32-0.37		8	
0.37-0.42		22	
0.42-0.47		37	
0.47-0.52		83	
0.52-0.57		127	
0.57-0.62		326	
> 0.62	1713		
P-D Interaction Integration	Mouse Protein-DNA Interactions		42
	GeneWays	≤ 0	3
		> 0	10
	ARACNE	< 0.27	3
≥ 0.27		24	

annotations provide additional clues about a P-P interaction. We assembled a list of 4,510,212 human gene pairs sharing a biological process annotation. They were classified using the information content of each GO term, retaining the highest value in case of multiple annotations. This information was binned in 8 groups to compute the LR . The LR distribution shows that GO categories with higher information content, reflecting very precise functional similarity, are more likely to support a P-P interaction than those with smaller values.

Gene co-expression data: It has been established that some interacting proteins, especially those in stable complexes, tend to be co-expressed [23]. Thus co-expression in a large expression profile dataset can provide clues about P-P interactions. We computed mutual information (MI) between 6,083 human genes using their mRNA expression levels measured by the Affymetrix chip HG-U95Av2 in 254 normal, tumor related, and experimentally manipulated B cell populations [15]. We binned the MI into 9 categories to classify the gene pairs. As expected, the LR significantly increases with the MI, reflecting the fact that interacting proteins tend to be co-expressed.

3.2 P-D Interactions

We combined information on P-D interactions from different sources including (a) mouse data from Transfac Professional [19] and BIND and (b) human P-D interactions inferred by the GeneWays and ARACNE algorithms. The data from each clue was binned and tested against the GSP and GSN to compute the LR , reflecting the ability of individual clues to predict MYC targets and, by generalization, other transcriptional interactions (Table 1).

Mouse data from Transfac Professional: We extracted mouse P-D interactions from the Transfac Professional and BIND databases and used the Inparanoid database to predict human P-D interactions, selecting the genes associated to a cluster with a score of 1 only. We defined 551 potential interactions involving 431 genes.

GeneWays: GeneWays interactions contained 250 interactions from Transfac Professional, revealing enrichment for 12 actions: *activate*, *depend on*, *include*, *independence*, *influence*, *mediate*, *regulate*, *repress*, *transactivate* and *up-regulate*. In the case where we found enrichment for an action in both P-D interaction and P-P interaction groups (e.g. bind) we retained the action for the group that showed the most significant enrichment for that action. This list allowed us to extract 4,141 potential human P-D interactions, involving 1,754 genes, further classified into two groups according to their score. The LR was computed for the two groups.

ARACNE: ARACNE was successful in predicting MYC targets that were experimentally validated, allowing us to use these results to compute the reliability of ARACNE predictions. We classified the predicted MYC targets according to their MI for computing the LR , revealing that a MYC target with a MI greater than 0.27 has a $p > 50\%$ probability to be a true interaction. We also assumed

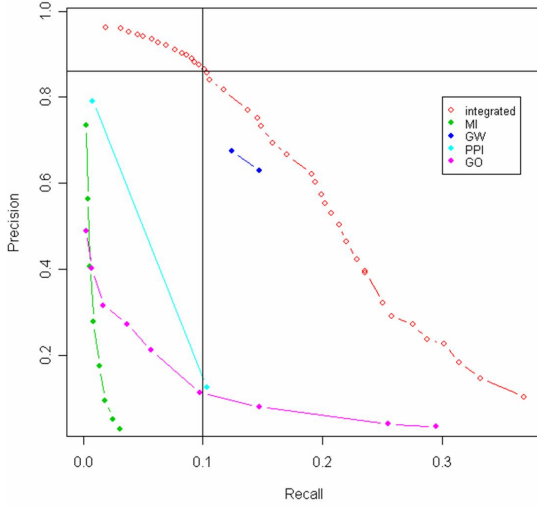


Fig. 1. 10-fold cross-validation: Precision (TP/TP+FP) vs. Recall (TP/TP+FN) curve for the individual and integrated sets (GO = Gene Ontology, GW = GeneWays, MI = Mutual Information, PPI = interactions in model organisms and human Y2H data). TP (True Positive), FP (False Positive) and FN (False Negative) were calculated as GSP and GSN interactions.

that this threshold would produce result similar to those for MYC for other TFs. This was biochemically validated using the BCL6 transcription factor (data not reported). Here, ARACNE predicted 76,251 P-D interactions in human B cells. Note that the MI used for categorizing the LR was computed using the same version as for the new bootstrapping version of ARACNE.

3.3 Bayesian Integration

The Naïve Bayes classifications allowed integrating different sources in a final set of 15,278 P-P interactions (4,373 genes) and 16,640 P-D interactions (462 TFs and 2,026 putative targets) with a posterior probability $p > 50\%$ of being true interactions. We called this set a *mixed interaction* network. The P-P interaction LR distribution (see Table 1) shows that each individual clue is not sufficient to predict interactions, except for clues from strong gene co-expression and from model organisms and Y2H. This last group was expected to be a good predictor as it already intrinsically combines different information sources. Considering P-D interactions, except for 551 interactions predicted from mouse data, only P-D interactions that are ARACNE positive could achieve sufficient LR to exceed the significance threshold ($LR_0 = 21$) determined by the prior (see Table 1). Since ARACNE inferences depend on expression profile data (which is cell-context specific), we claim that the transcriptional part of the network is B cell specific.

To evaluate the performance of the P-P interaction classifier, we computed precision and recall for each evidence source and for the final integrated set,



using a ten-fold cross-validation process (Figure 1). For an LR_0 threshold of 800, our network achieves recall of up to 10% with precision always greater than 86%. These measures also illustrate that the interaction clues, when combined together, are a much better predictor than each one taken independently.

In a previous Naïve Bayes classification of human P-P interactions [8], the authors defined 38,379 P-P interactions involving 5,791 genes, using a LR threshold of 381. With this threshold, we identify 40,161 putative P-P interactions among 7,603 genes. Of these, 3,995 are common to the two studies supporting 19,072 P-P interactions in the Rhodes set and 19,226 in our set respectively. Of these, 3,201 were common to both studies, corresponding to 17% of each of the predicted sets. This small, yet highly statistically significant overlap can be explained by the use of our highly context-specific gene expression profile dataset, which is likely to identify interactions that are specific to B cells. We thus consider our P-P interaction interactome to be at least partially indicative of interactions that are B cell specific. Similarly, since the most significant contribution to the total LR comes from the ARACNE algorithm, we also consider P-D interactions to be B cell specific.

3.4 Mixed Interaction Network and Motifs Analysis

To build the final mixed interaction network, we included all missing interactions in the GSPs as well as transcriptional interactions for the TFs reported in Transfac Professional and BIND (respectively 25,473 P-P and 2,798 P-D interactions). The final network contains 40,751 P-P (7,888 genes) and 19,370 P-D (3,768 genes) interactions. Respectively, 12,209 and 16,445 of these were new (i.e., not previously in the GSP or Transfac and BIND). We searched this network for three gene motifs that were highly statistically enriched with respect to the null hypothesis (1,000 randomized networks of identical connectivity). We were particularly interested in two highly enriched regulatory motifs (referred to as R1 and R2) combining both interaction types (Table 2).

Table 2. Regulatory motifs involving P-D and P-P interactions

Motif		#motifs		
Name	Representation	Real Network	Random Network (mean \pm SD)	Z-score
R1		23,056	3,496 \pm 109	179
R2		3,735	801 \pm 153	19

R1 motifs (z-score $Z_{R1} = 179$) describe the regulation of two proteins in a complex by the same TF, suggesting that genes encoding proteins that interact in a complex tend to have a common regulatory program. Among the 6,107 P-P interactions in R1 motifs, 2,037 are jointly regulated by more than one TF (see Supplementary Table S1). These combinatorial regulation events were highly statistically significant in the real network compared to the null hypothesis, highlighting regulation of protein complexes of higher complexity such as for example the ribosome or the collagen. As an illustration, we reported the common targets of CEBPD and MAF that regulate 48 genes encoding proteins organized in complexes (Fig. 2a).

Table 3. Enriched motifs with at least 30% new interactions. P-value was computed with a fisher exact test and reported non-corrected and corrected for multiple testing (bonferroni correction).

TF1	TF2	common targets z-score			Gene Ontology		
		#targets	%new		BP Annotation	P-value (corrected)	
CEBPB	CEBPG	3	67	9	–		–
ELK1	ELK3	2	50	14	–		–
IRF8	IRF1	2	50	8	–		–
IRF1	SPI1	3	50	7	antimicrobial response	humoral	9.10^{-4} (3.10^{-1})
SMAD4	TFE3	2	50	11	–		–
RB1	RBL1	2	50	13	–		–
SRF	YY1	4	50	3	muscle development		2.10^{-4} (7.10^{-2})
CEBPB	SPI1	4	38	6	antimicrobial response	humoral	2.10^{-3} (6.10^{-1})
FOS	SRF	4	38	6	positive regulation of cell proliferation		2.10^{-3} (8.10^{-1})
IRF1	NFKB1	4	38	7	natural killer cell activation		1.10^{-4} (4.10^{-2})
FOS	NFKB1	6	33	9	natural killer cell activation		3.10^{-4} (1.10^{-1})
GATA1	GATA2	3	33	9	–		–

R2 motifs (z-score $Z_{R2} = 19$), on the other hand, show that TFs in a complex tend to regulate the same target genes. Only 597 TF pairs are represented in the 3,735 R2 motifs, indicating that many TF pairs regulate more than one gene. Specifically, 66 TF pairs - with statistically independent expression profiles - were found to regulate two or more common targets (see Supplementary Table S2). We report 12 motifs containing at least 30% new interactions (Table 3). This list shows several TF complexes involving proteins from the same family, such as CEBPB and CEBPG (Fig. 2b), as well as non-related proteins, such as YY1 and SRF, known to bind to the same DNA sites (CArG boxes) on target gene promoters (Fig. 2c). Additionally, some TF complexes included proteins known

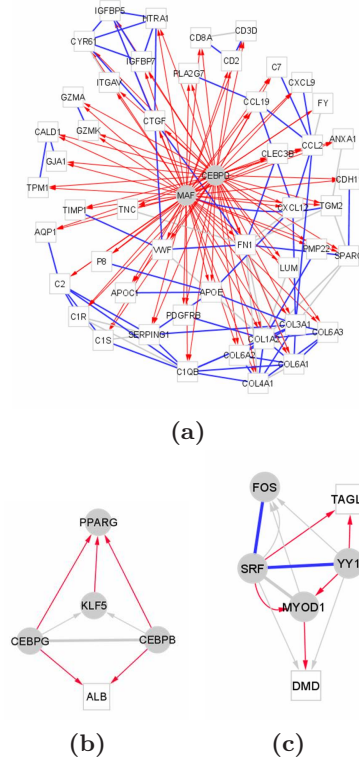


Fig. 2. Examples of enriched regulatory motifs. Undirected and directed edges represent P-P and P-D interactions respectively colored in blue and red if new and in grey if in the GSP. (a) Protein complex regulated by CEBPD and MAF. (b) CEBPB-CEBPG motif. (c) SRF-YY1 motif.

to bind to distinct binding sites. For example, SMAD4 and TFE3 bind respectively to a 4 base pair Smad element and to an E-box. These sites were found to be adjacent in the promoter of known target genes [24]. These differences may help distinguish among different cooperative regulation modes: two modes are associated with either a TF complex binding to a single binding site or two adjacent sites in the promoter of the target genes, while the third mode is associated with two TFs independently binding to different sites on the promoter.

4 Conclusions

The proposed framework constitutes the first example of a mammalian mixed interaction network. Transcriptional cell context specificity was achieved by constraining the inferred P-D interactions on clues dependent on expression profile data. P-P interactions are more likely affected by protein availability than by changes in P-P affinity in different cell types.

5 Supplementary Materials

Supplementary materials are available at:

<http://wiki.c2b2.columbia.edu/califanolab/index.php/Publications>

Acknowledgments

We thank Andrey Rzhetsky and Raul Rodriguez-Esteban (Columbia University) for providing us GeneWays interactions and for their helpful comments. This work was supported by the National Cancer Institute (R01CA109755), the National Institute of Allergy and Infectious Diseases (R01AI066116), and the National Centers for Biomedical Computing NIH Roadmap Initiative (U54CA121852).

References

1. Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R.Y., Alon, U., Margalit, H.: Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. In: Proc Natl Acad Sci. USA 101(16), 5934–5939 (2004)
2. Yu, H., Xia, Y., Trifonov, V., Gerstein, M.: Design principles of molecular networks revealed by global comparisons and composite motifs. *Genome Biol.* R7(7), R55 (2006)
3. Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P., Vidal, M.: Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437(7062), 1173–1178 (2005)
4. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzflaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., Wanker, E.E.: A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122(6), 957–968 (2005)
5. Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboue, P.A., Weng, W., Wilbur, W.J., Hatzivassiloglou, V., Friedman, C.: Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. Biomed Inform.* 37(1), 43–53 (2004)
6. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, D., Califano, A.: Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl 1), S1–7 (2006)
7. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.: A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302(5644), 449–453 (2003)

8. Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., Chinnaiyan, A.M.: Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.* 23(8), 951–959 (2005)
9. Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjana, V., Muthusamy, B., Gandhi, T.K., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H.N., Rashmi, B.P., Ramya, M.A., Zhao, Z., Chandrika, K.N., Padma, N., Harsha, H.C., Yatish, A.J., Kavitha, M.P., Menezes, M., Choudhury, D.R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S.K., Madavan, V., Joseph, A., Wong, G.W., Schiemann, W.P., Constantinescu, S.N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobel, G.C., Dang, C.V., Garcia, J.G., Pevsner, J., Jensen, O.N., Roepstorff, P., Deshpande, K.S., Chinnaiyan, A.M., Hamosh, A., Chakravarti, A., Pandey, A.: Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* 13(10), 2363–2371 (2003)
10. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., Apweiler, R.: Intact: an open source molecular interaction database. *Nucleic Acids Res.* 32(Database issue), D452–D455 (2004)
11. Bader, G.D., Betel, D., Hogue, C.W.: Bind: the biomolecular interaction network database. *Nucleic Acids Res.* 31(1), 248–250 (2003)
12. Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D.: Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30(1), 303–305 (2002)
13. Kumar, A., Agarwal, S., Heyman, J.A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Cheung, K.H., Miller, P., Gerstein, M., Roeder, G.S., Snyder, M.: Subcellular localization of the yeast proteome. *Genes Dev.* 16(6), 707–719 (2002)
14. Zeller, K.I., Jegga, A.G., Aronow, B.J., O'Donnell, K.A., Dang, C.V.: An integrated database of genes responsive to the myc oncogenic transcription factor: identification of direct genomic targets. *Genome Biol.* 4(10), R69 (2003)
15. Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A.: Reverse engineering of regulatory networks in human b cells. *Nat. Genet.* 37(4), 382–390 (2005)
16. Wang, K., Banerjee, N., Margolin, A., Nemenman, I., Califano, A.: Genome-wide discovery of modulators of transcriptional interactions in human b lymphocytes. In: Apostolico, A., Guerra, C., Istrail, S., Pevzner, P., Waterman, M. (eds.) RECOMB 2006. LNCS (LNBI), vol. 3909, pp. 348–362. Springer, Heidelberg (2006)
17. Alterovitz, G., Xiang, M., Mohan, M., Ramoni, M.F.: Go pad: the gene ontology partition database. *Nucleic Acids Res.* 35, D322–D327 (2006)
18. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25(1), 25–29 (2000)
19. Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., Wingender, E.: Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 31(1), 374–378 (2003)

20. Mewes, H.W., Frishman, D., Mayer, K.F., Munsterkotter, M., Noubibou, O., Pagel, P., Rattei, T., Oesterheld, M., Ruepp, A., Stumpflen, V.: Mips: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* 34(Database issue), D169–D172 (2006)
21. O'Brien, K.P., Remm, M., Sonnhammer, E.L.: Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* 33(Database issue), D476–D480 (2005)
22. Vazquez, A., Flammini, A., Maritan, A., Vespignani, A.: Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.* 21(6), 697–700 (2003)
23. Ge, H., Liu, Z., Church, G.M., Vidal, M.: Correlation between transcriptome and interactome mapping data from *saccharomyces cerevisiae*. *Nat Genet.* 29(4), 482–486 (2001)
24. Hua, X., Miller, Z.A., Wu, G., Shi, Y., Lodish, H.F.: Specificity in transforming growth factor beta-induced transcription of the plasminogen activator inhibitor-1 gene: interactions of promoter dna, transcription factor *myc3*, and *smad* proteins. *Proc Natl Acad Sci. USA* 96(23), 13130–13135 (1999)