Network Legos: Building Blocks of Cellular Wiring Diagrams

T. M. Murali and Corban G. Rivera

660 McBryde Hall, Department of Computer Science,
Virginia Polytechnic Institute and State University, Blacksburg VA 24061

Abstract. Publicly-available data sets provide detailed and large-scale information on multiple types of molecular interaction networks in a number of model organisms. These multi-modal universal networks capture a static view of cellular state. An important challenge in systems biology is obtaining a dynamic perspective on these networks by integrating them with gene expression measurements taken under multiple conditions.

We present a top-down computational approach to identify building blocks of molecular interaction networks by

(i) integrating gene expression measurements for a particular disease state (e.g., leukaemia) or experimental condition (e.g., treatment with growth serum) with molecular interactions to reveal an *active network*, which is the network of interactions active in the cell in that disease state or condition and

(ii) systematically combining active networks computed for different experimental conditions using set-theoretic formulae to reveal *network legos*, which are modules of coherently interacting genes and gene products in the wiring diagram.

We propose efficient methods to compute active networks, systematically mine candidate legos, assess the statistical significance of these candidates, arrange them in a directed acyclic graph (DAG), and exploit the structure of the DAG to identify true network legos. We describe methods to assess the stability of our computations to changes in the input and to recover active networks by composing network legos.

We analyse two human datasets using our method. A comparison of three leukaemias demonstrates how a biologist can use our system to identify specific differences between these diseases. A larger-scale analysis of 13 distinct stresses illustrates our ability to compute the building blocks of the interaction networks activated in response to these stresses.

1 Introduction

Rapid advances in high-throughput and large-scale biological experiments are inspiring the study of properties of sets of molecules that act in concert [13], how these sets interact with each other, and how these interactions change dynamically in response to perturbations. Such groups of molecules have been dubbed various names such as gene modules [5, 30, 34], module networks [31] and gene

sets [35]. One of the fundamental challenges of systems biology is to automatically compute such modules and the relationships between them by integrating multiple types of data and discovering patterns of coordinated activity contained in these data sets.

In this paper, we present a top-down computational approach that identifies building blocks of cellular networks by

(i) integrating gene expression measurements for a particular disease state (e.g., leukaemia) or experimental condition (e.g., treatment with growth serum) with molecular interactions to reveal an *active network*, which is the network of interactions active in the cell in that disease state or condition and

(ii) systematically combining active networks computed for different experimental conditions using set-theoretic formulae to reveal *network legos*, which are modules of coherently interacting genes and gene products in the wiring diagram. These network legos are potential building blocks of the wiring diagram, since we can express each active network as a composition of network legos.

We illustrate the essence of our method using an example. Armstrong et al. [2] demonstrated that lymphoblastic leukaemias involving translocations in the *MLL* gene constitute a disease different from conventional acute lymphoblastic (ALL) and acute myelogenous leukaemia (AML). The authors based their analysis on the comparison of gene expression profiles from individuals diagnosed with ALL, AML, and MLL. We reasoned that the networks of molecular interactions activated in these diseases may also show distinct differences. First, we computed networks of molecular interactions activated in each leukaemia, as described in Section 3.2. Next, we systematically combined these active networks in multiple ways into network legos using graph intersections and graph differences, using the method presented in Section 3.3. Our system generated all the possible 19 ($3^3 - 2^3$) combinations involving the ALL, AML, and MLL active networks and their complements and connected them in the directed acyclic graph (DAG) displayed in Figure 1. In this DAG, each node represents a single combination, e.g., the leftmost node on the top row represents the MLL active network while the leftmost node in the middle row represents the interactions activated in AML but not in MLL (the "formula" AML$\cap$!MLL). A solid blue edge directed from a child to a parent indicates that the formula for the child (e. g., MLL) appears as a part of the formula for the parent (e.g., MLL$\cap$!AML), while a green edge indicates that the child's formula (e.g., MLL) appears negated in the parent's formula (e.g., AML$\cap$!MLL). The DAG is a concise representation of all the formulae we compute and the subset relationships between the formulae. We did not consider complementation-only formulae such as !ALL$\cap$!AML since the resulting networks are unlikely to be biologically useful. In Section 4.1, we describe how function and pathway enrichment of these networks suggests differences and similarities between ALL, AML, and MLL. For instance, Figure 1 displays an example of the enrichment of the interactions in the KIT pathway in the computed

network legos. Interactions in this pathway are significantly enriched only in formulae that involve the AML active network; the most significant enrichment (3.5×10^{-7}) occurs in the formula AML $\cap$!ALL $\cap$!MLL.

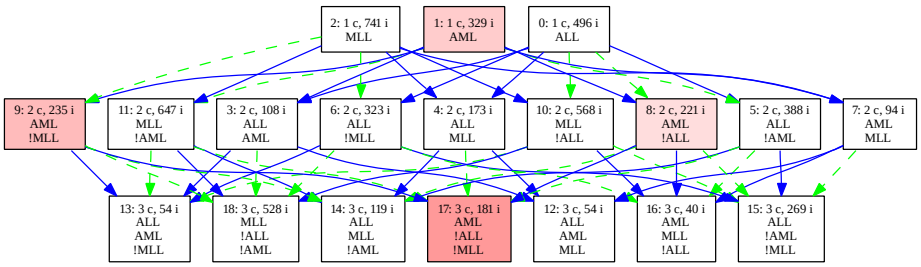

Fig. 1. The lattice connecting combinations of ALL, AML, and MLL active networks. Each node contains an index, the number of 'c'onditions, the number of 'i'nteractions and the active networks participating in the formula, with '!' indicating complementation. Colours indicate differential enrichment of the interactions in the KIT pathway in the computed combinations. Darker colours denote more significant enrichment values.

Given a wiring diagram and the transcriptional measurements for a particular condition, we use the gene expression data to induce edge weights in the wiring diagram. We find dense subgraphs [8] in this weighted graph to compute the active network for that condition. Given the active networks for a number of different conditions, we first represent the active networks in an appropriately-defined binary matrix and compute closed itemsets [1, 40] in the matrix. Each itemset simultaneously represents a set-theoretic combination of particular active networks and a subgraph of the wiring diagram; we call such a subgraph a "network block". We exploit the subset structure between blocks to arrange them in a DAG. When the number of active networks is large, we may compute a very large number of highly-similar blocks. Not all these blocks are likely to be network legos. We assess the statistical significance of each block by simulation and identify those that are maximally significant, i.e., more significant than any descendant or an ancestor in the DAG. We deem these blocks to be network legos.

We develop two measures to assess the quality of the network legos we compute. *Stability* measures to what degree we can recompute the same legos when we remove each active network in turn from the input. *Recoverability* measures to what extent we recoup the original active networks when we combine network legos. These two notions test two different aspects of network lego computation. Considering active networks to be the inputs and network legos to be the outputs, stability measures how much the outputs change when we perturb the inputs by removing one of the inputs at a time. In contrast, recoverability asks whether we can reclaim the inputs by combining the outputs; thus recoverability is a measure of how well the network legos serve as building blocks. To assess

the biological content of network legos, we measure the functional enrichment of the genes and interactions that belong to a network lego. For each function, we track its degree of enrichment in the DAG to visually highlight differences among the active networks, as displayed in Figure 1.

In addition to the ALL-AML-MLL analysis, we apply our approach to a collection of 178 arrays measuring the gene expression responses of HeLa cells and primary human lung fibroblasts to cell cycle arrest, heat shock, endoplasmic reticulum stress, oxidative stress, and crowding [21]. Overall, the dataset contains 13 distinct stresses over the two cell types. Our method computes 143 network legos. In this paper, we focus on a structural analysis of the network legos. We carefully examine the compositions of these network legos to demonstrate that they are true building blocks of the active networks for these 13 stresses. We demonstrate that our algorithm to construct network legos is stable: when we remove each active network and recompute network legos, we are able to recompute most network legos at least 95% fidelity. We also demonstrate that we can recover active networks with almost perfect accuracy by composing network legos. Further analysis of the network legos reveals that the active networks corresponding to cell cycle arrest contain interactions that are quite distinct from the network of interactions activated by the other stresses. When we remove the two cell cycle arrest data sets, we compute only 15 network legos. Of the 11 remaining active networks, we recover five with complete accuracy and one with 99.9% accuracy. We recover the other five active networks with accuracies ranging from 71% to 92%. Taken together, these statistics indicate that the network legos we detect are indeed building blocks of the networks activated in response to the stresses studied by Murray et al. [21].

There are two ways in which a biologist can use our system. In the first, our system allows the systematic comparison of responses to a small number of different conditions, diseases, or perturbations tested in the same lab. The ALL-AML-MLL comparison we presented earlier and discuss further in Section 4.1 is such an application. In the second, a biologist can analyse a specific condition of interest in the context of a large compendium of other conditions, compute the building blocks of the networks activated in these conditions, and ask how the building blocks compose the active network for the specific condition of interest to the biologist. In Section 4.2, we analyse 13 distinct stresses imparted to human cells to illustrate this application. In this respect, our work is similar to the approach developed by Tanay et al. [38]. They integrate a diverse collection of datasets into a bipartite graph representing connections between genes and gene properties. Their modules are statistically-significant biclusters [37] in this graph. They represent a target gene expression dataset as a bipartite graph and compute which already-computed modules respond in the target data set. Our approach differs from theirs in two respects. First, we represent differences and similarities between multiple conditions explicitly as a set theoretic formula involving the interaction network activated in each condition. Two, when we analyse a large compendium of gene expression data sets, we exploit the subset

structure between these formulae to detect network legos, statistically-significant building blocks of these active networks.

The success of our approach stems from a number of factors. First, unlike other approaches that simultaneously integrate multiple gene expression data sets in the context of the network scaffold [5, 22], we compute individual active networks for each data set and associate the active network with the corresponding disease or perturbation. This approach allows us to explicitly compare and contrast different conditions. Second, we treat interactions (rather than genes or proteins) as the elementary objects of our analysis. Therefore, different network legos may share genes, allowing for the situation when a gene participates in multiple biological processes and is activated differently in these processes. Finally, we develop a simple but effective recursive method to assess the statistical significance of a network lego and to weed out sub-networks that masquerade as building blocks but contain true network legos. Taken together, formulae and network legos provide a dynamic and multi-dimensional view of cell circuitry obtained by integrating molecular interaction networks, gene expression data, and descriptions of experimental conditions.

2 Previous Work

A number of approaches, recently surveyed by Joyce and Palsson [17], have been developed to integrate diverse types of biological data and "mine" these datasets to find groups of molecules (usually genes and/or proteins) that act in concert to perform a specific biological task. Integrating information on available molecular interactions such as protein-protein, protein-DNA, protein-metabolite, and genetic interactions yields a multi-modal wiring diagram [32]. However, such a network typically provides a static view of the underlying cellular circuitry. A number of techniques attempt to obtain a dynamic view of cell state by overlaying measurements of molecular profiles (usually in the form of gene expression data) obtained under multiple conditions on the wiring diagram [10, 12, 15, 17, 20]. For instance, Han et al. [12] categorised hubs in *S. cerevisiae* protein interaction networks into "party" hubs, which interact with most of their partners simultaneously, and "date" hubs, which bind their different partners at different times or locations. Luscombe et al. [20] characterise topological changes in the structure of the *S. cerevisiae* transcriptional regulatory network under different conditions. The SAMBA algorithm [36] integrates a wide variety of data types in *S. cerevisiae* to identify gene modules with statistically significant correlated behaviour across diverse data sources. The bioPIXIE system [22] probabilistically integrates diverse genome-wide datasets and computes pathway-specific networks that include query genes input by a biologist.

Other methods have computed gene modules by focusing solely on gene expression data collected across multiple cellular conditions; they analyse large compendia of such data to reveal similarities and differences between multiple cellular conditions [30] or between organisms [7, 34], predict functional annotations [14, 18], reconstruct regulatory networks [41] and networks activated in

diseases [6], zero in on biomarkers for diseases [25, 26], and identify the gene products and associated pathways that a drug compound targets [10].

3 Algorithms

We describe the main computational ingredients of our approach in stages. We first define some useful terminology. Next, we present our method to integrate a cellular wiring diagram with the gene expression data for a single condition to compute the active network for that condition. Third, we describe how we combine active networks for different conditions to form blocks. Fourth, we discuss how we compute the statistical significance of blocks, arrange them in a DAG, and exploit the DAG to identify network legos, which are the most statistically-significant blocks in the DAG. Finally, we present our methods to measure the stability of network legos and assess how well we can recover active networks from the network legos.

3.1 Definitions

Our method takes as input (i) a cellular wiring diagram W representing known physical and/or genetic interactions between genes or gene products in an organism and (ii) a compendium of transcriptional measurements in the same organism obtained under various conditions such as diseases (e.g., breast cancer), stimuli (e.g., heat shock), or other perturbations (e.g., gene knock-out or over-expression). We assume that each gene expression dataset in the compendium contains measurements for multiple gene chips. For instance, a breast cancer dataset might include data from multiple patients while a heat shock dataset may measure gene expression at different time points.

Given a gene expression data set D_c for a condition c, we say that a gene *responds in* c if the expression values of the gene in D_c vary by more than an input threshold. Let g and h be two genes that respond in c and let $e = (g, h)$ be an interaction in W. We say that e is *active in* c if the expression profiles of g and h in D_c are correlated to a statistically-significant extent. The *active network* A_c *in* c is the sub-network of interactions in W that are active in c. We describe the details of how we detect responding genes, active interactions, and active networks in Section 3.2.

Let $\mathcal{A}$ denote the set of active networks for each of the conditions in the input compendium. We define a *block* to be a triple $(G, \mathcal{P}, \mathcal{N})$, where G is a subgraph of W; $\mathcal{P}$ and $\mathcal{N}$ are subsets of $\mathcal{A}$; $\mathcal{P} \neq \emptyset$; and $\mathcal{P} \cap \mathcal{N} = \emptyset$ such that

1. for each *positive* active network $P \in \mathcal{P}$, $G \subseteq P$,
2. for each *negative* active network $N \in \mathcal{N}$, $G \cap N = \emptyset$,
3. G is maximal, i.e., adding an edge to G violates at least one of the first two properties,
4. $\mathcal{P}$ is maximal, i.e., there is no $P \in \mathcal{A} - \mathcal{P}$ such that $G \subseteq P$, and
5. $\mathcal{N}$ is maximal, i.e., there is no $N \in \mathcal{A} - \mathcal{N}$ such that $G \cap N = \emptyset$.

Intuitively, we can form G by taking the intersection of all the active networks in $\mathcal{P}$ and removing any edge that appears in any of the active networks in $\mathcal{N}$. In other words,

$$G = \Big(\bigcap_{P \in \mathcal{P}} P \Big) \bigcap \Big(\bigcap_{N \in \mathcal{N}} !N \Big) = \Big(\bigcap_{P \in \mathcal{P}} P \Big) - \Big(\bigcup_{N \in \mathcal{N}} N \Big),$$

where "$\cap$" (respectively, "$\cup$") denotes the intersection (respectively, union) of the edge sets of two graphs and "!" denotes the complementation (with respect to W) of the edge set of a graph. We require that $\mathcal{P}$ contain at least one active network so that G is not formed solely by the intersection of the networks in $\mathcal{N}$; such a block is unlikely to be biologically interesting. We also require that $\mathcal{P}$ and $\mathcal{N}$ be disjoint so that G is not the empty graph. Requiring $\mathcal{P}$ and $\mathcal{N}$ to be maximal ensures that we include all the relevant active networks in the construction of G. These criteria imply that it is enough to specify $\mathcal{P}$ and $\mathcal{N}$ to compute G uniquely; we include G in the notation for a block for convenience and drop $\mathcal{P}$ and $\mathcal{N}$ when they are understood from the context. We refer to $\big(\bigcap_{P \in \mathcal{P}} P\big) - \big(\bigcup_{N \in \mathcal{N}} N\big)$ as the *formula* for the block.

Let $\mathcal{B}$ be a set of blocks. Given two blocks $(G_1, \mathcal{P}_1, \mathcal{N}_1)$ and $(G_2, \mathcal{P}_2, \mathcal{N}_2)$ in $\mathcal{B}$, we say that $G_1 \prec G_2$ if

(i) $\mathcal{P}_1 \subseteq \mathcal{P}_2$ and $\mathcal{N}_1 \subseteq \mathcal{N}_2$ or
(ii) $\mathcal{P}_1 \subseteq \mathcal{N}_2$ and $\mathcal{N}_1 \subseteq \mathcal{P}_2$.

We say that $G_1 < G_2$ if there is no block $G_3 \in \mathcal{B}$ such that $G_1 \prec G_3 \prec G_2$. The operators $<$ and $\prec$ represent partial orders between blocks, with $\prec$ being the transitive closure of $<$. Given a set $\mathcal{B}$ of blocks, let $\mathcal{D}_{\mathcal{B}}$ denote the directed acyclic graph representing the partial order $<$: each node in $\mathcal{D}_{\mathcal{B}}$ is a block in $\mathcal{B}$ and an edge connects two blocks related by $<$. For a block G, let $\sigma_G \in [0, 1]$ denote the statistical significance of G. We describe a method to compute this value in Section 3.4. We define a *network lego* to be a block $(G, \mathcal{P}, \mathcal{N}) \in \mathcal{B}$ such that $\sigma_G < \sigma_H$, for every $H \in \mathcal{B}$ where $G \prec H$ or $H \prec G$. In other words, $(G, \mathcal{P}, \mathcal{N})$ is a network lego if it is more statistically significant that blocks formed by combining any subset of $\mathcal{P}$ and $\mathcal{N}$ or by combining any superset of $\mathcal{P}$ and $\mathcal{N}$. In this sense, we claim that G is a building block of the active networks in $\mathcal{A}$.

3.2 Computing active networks

Given a gene expression dataset for a disease state or an experimental condition c, we use a variational filter to remove all genes whose expression profile has a small dynamic range from the wiring diagram W. More specifically, we log-transform and zero centre each gene's expression values. We discard a gene and all its interactions in the wiring diagram W if all the transformed expression values of the gene lie between -1 and 1 [30]. We deem the remaining genes to have responded in the condition. To each interaction $e = (g, h)$ remaining

in W, we assign a weight equal to the absolute value of the Pearson's correlation coefficient of the expression profiles of the genes g and h, reasoning that this weight indicates how "active" the interaction is in the experimental condition. We discard edges whose weights are not statistically significant (based on a permutation test) at the 0.01 level. Let W_c be the resulting weighted interaction network. To mitigate the effect of isolated interactions in W_c, we search for pockets of concerted activity in W_c as follows. We define the *density* of a graph to be the total weight of the edges in the graph divided by the number of nodes in the graph. It is possible to find the subgraph of largest density using linear programming or parametric network flows [8]. We use a simpler greedy algorithm that finds a subgraph whose density is at least half the maximum density [8]. We repeatedly apply this approximation algorithm, remove the edges of the subgraph it computes, and re-invoke the algorithm on the remaining graph until the density of the remaining graph is less than the density of W_c. We deem the union of the computed dense subgraphs to be the active network A_c for the condition.

3.3 Computing blocks

We reduce the problem of computing blocks to the problem of computing closed itemsets in a binary matrix [1, 40]. We construct a binary matrix M where each column corresponds to an interaction in the wiring diagram W. The matrix M contains two rows for each active network $A \in \mathcal{A}$: the *positive* row corresponds to the interactions in A and the *negative* row to the interactions in $W - A$. In the positive row corresponding to A, we set a cell to be one if and only if the corresponding interaction belongs to A; this cell is zero in the negative row for A. Thus, M is a qualitative representation of which interactions are present in each active network and which are present in its complement.

In a binary matrix such as $\mathcal{B}$, an *itemset* (R, C) is a subset R of rows and a subset C of columns such that the sub-matrix spanned by these rows and columns only contains ones [1]. A *closed* itemset [40] is an itemset with the property that each row (respectively, column) not in the itemset contains a zero in at least one column (respectively, row) in the itemset. Therefore, it is not possible to add a row or a column to the itemset without introducing a zero into the corresponding sub-matrix. We can partition R into two subsets R_P and R_N where R_P (respectively, R_N) consists of all the positive (respectively, negative) rows in R. There is a natural mapping from a closed itemset (R, C) to a block $(G, \mathcal{P}, \mathcal{N})$:

1. G is the subgraph of W induced by the interactions corresponding to the columns in C;
2. $\mathcal{P}$ is the set of active networks corresponding to the rows in R_P; and
3. $\mathcal{N}$ is the set of active networks corresponding to the rows in R_N.

We compute closed itemsets in $\mathcal{B}$ to satisfy the maximality requirements in the definition of a block. We do not compute any itemsets where all rows correspond

to complements of active networks, since such itemsets are unlikely to be biologically relevant (they correspond to blocks where $\mathcal{P} = \emptyset$). To construct closed itemsets, we use our implementation of the *Apriori* algorithm [1]. We have modified the original *Apriori* algorithm to construct closed itemsets. We convert each itemset to the corresponding block and formula. Finally, we connect the resulting set of blocks $\mathcal{B}$ in the DAG $\mathcal{D}_{\mathcal{B}}$ as per the partial order $<$.

3.4 Statistical significance of a block

To measure the statistical significance of a block, we construct an empirical distribution of block sizes. We repeatedly select a subset of rows uniformly at random from the binary matrix M, compute the columns common to these rows, and convert the resulting itemset into a block. We ensure that the random subset of rows does not contain an active network and its complement, since such a subset will trivially result in an itemset with zero columns. Given a block $(G, \mathcal{P}, \mathcal{N})$ computed in the real dataset, let m be the number of interactions in G. To estimate the statistical significance σ_G of $(G, \mathcal{P}, \mathcal{N})$, we only consider the distribution formed by random blocks $(H, \mathcal{P}', \mathcal{N}')$ where $|\mathcal{P}| = |\mathcal{P}'|$ and $|\mathcal{N}| = |\mathcal{N}'|$. We set σ_G to be the fraction of such blocks that have more than m interactions. Since the number of interactions in a block will decrease with an increase in $|\mathcal{P}|$ or in $|\mathcal{N}|$, these constraints ensure that we compare G with appropriate random blocks in order to estimate σ_G. We only retain blocks that are significant at the 0.01 level. We compute the DAG defined by these blocks. We perform two topological traversals of this DAG, one from the roots to the leaves and the other from the leaves to the roots, to identify the maximally-significant blocks. The resulting set of blocks are the network legos we desire to compute. Let $\mathcal{L}$ denote the set of network legos.

3.5 Stability and recoverability analysis

It is clear that the set $\mathcal{L}$ of network legos we compute depend on the active networks in $\mathcal{A}$. To assess this dependence, we modify a method for suggested by Segal et al. [30]. We remove each network $N \in \mathcal{A}$ in turn and recompute network legos from the set $\mathcal{A} - \{N\}$. Let $\mathcal{L}_N$ denote the resulting set of network legos. For each network lego L in $\mathcal{L}$, we compute the most similar network lego L' in $\mathcal{L}_N$ using the set-similarity measure $(|L \cap L'|/|L \cup L'|)$ and store this measure as $s_{L,N}$. Given a similarity threshold t, for each network lego L in $\mathcal{L}$, we compute the fraction of networks in $\mathcal{A}$ such that $s_{L,N} \geq t$. The higher this fraction is, the more resilient L is to perturbations in the input.

If the network legos in $\mathcal{L}$ are true building blocks of the active networks in $\mathcal{A}$ that they spring from, it should be possible to recover each active network in $\mathcal{A}$ from the network legos in $\mathcal{L}$. For each active network A, we define

$$\mathcal{L}_A = \{(G, \mathcal{P}, \mathcal{N}) \in \mathcal{L} | A \in \mathcal{P}\},$$

the set of network legos in $\mathcal{L}$ where A does not appear negated in the network lego. We compute the union of the network legos in $\mathcal{L}_A$ and the fraction of A's

edge set that appears in the union. The larger this fraction is, the more "recoverable" A is from the computed network legos.

4 Results

We applied the algorithm described in the previous section to human data sets. We obtained a network of 31108 molecular interactions between 9243 human gene products by integrating the interactions in the IDSERVE database [24], the results of large scale yeast two-hybrid experiments [27, 33], and 20 immune and cancer signalling pathways in the Netpath database (http://www.netpath.org). The IDSERVE database includes human curated interactions from BIND [4], HPRD [23], and Reactome [16], interactions predicted based on co-citations in article abstracts, and interactions that transferred from lower eukaryotes based on sequence similarity [19]. We derived functional annotations for the genes in our network from the Gene Ontology (GO) [3] and from MSigDB [35]. In addition, we annotated each Netpath interaction in our network with the name of the pathway it belonged to. We used these annotations to compute the functional enrichment of the nodes and edges in the network legos using the hypergeometric distribution with FDR correction.

4.1 ALL, AML, and MLL

We continue the analysis of ALL, AML, and MLL that we started in Section 1. Since the three leukaemias induce only 19 blocks, we did not compute the statistical significance of the blocks. Instead, we treated every block as a network lego. To assess the biological content of the results and to illustrate one type of analysis our approach facilitates, we computed functions enriched in the genes and interactions in the networks corresponding to the 19 formulae. Figure 1 demonstrates that the interactions in the KIT pathway are differentially enriched in the 19 networks. The darker the colour of a node, the more statistically significant is the enrichment of this pathway in the corresponding network. We first note that the only formulae enriched in this pathway are the ones that involve AML (and not the complement of AML). The statistical significance is the lowest (FDR-corrected p-value 3.5×10^{-7}) for the formula $\mathrm{AML} \cap !\mathrm{ALL} \cap !\mathrm{MLL}$, indicating that this pathway may be activated in AML and not in ALL or in MLL. Evidence in the literature supports this conclusion. The c-KIT receptor is activated in almost all subtypes of AML [29]. Similarly, Schnittger et al. [28] report that "mutations in codon D816 of the KIT gene represent a recurrent genetic alteration in AML". We note that gain-of-function mutations in c-Kit have been observed in many human cancers [9]. Our analysis only suggests that in the context of ALL, AML, and MLL, the KIT pathway may be activated only in AML.

4.2 Human Stresses

We computed network legos by applying our methods to the human interaction network and the gene expression responses of HeLa cells and primary human lung fibroblasts to heat shock, endoplasmic reticulum stress, oxidative stress, and crowding [21]. The dataset we analysed includes transcriptional measurements obtained by Whitfield et al. [39] for studying cell cycle arrest by using a double thymidine block or with a thymidine-nocodazole block. Overall, the dataset contains 13 distinct stresses over the two cell types. The authors note that each type of stress resulted in a distinct response and that there was no general stress response unlike in the case of *S. cerevisiae* [11]. Therefore, this dataset poses a challenge to our system. Can we find network legos that combine active networks for multiple stresses? In this paper, we focus on the topological and quantitative aspects of our results.

The number of genes in the 13 active networks we computed ranged from 165 (for crowding of WI38 cells) to 1148 (for the thymidine-nocodazole block) with an average of 684 genes per active network. The number of interactions ranged from 257 to 3667 with an average of 1874 interactions per active network. Theoretically, we can compute 1586131 ($3^{13} - 2^{13}$) blocks involving 13 distinct active networks. Our method computed 444201 blocks, indicating that the remaining combinations of active networks are not closed or yield blocks without any interactions. We computed a null distribution of block sizes using a million random samples. Of the 444201 blocks, 12386 blocks were statistically significant at the 0.01 level. We identified 143 network legos in the DAG induced by the relation $<$ on these blocks. We observed that all but one of the 143 network legos involved at least six distinct active networks, indicating that these network legos are not the result of combining a small number of active networks. The following table displays the distribution of the number of legos involving k conditions, where $5 \leq k \leq 12$. Interestingly, no network lego involved all 13 active networks.

#conditions	5	6	7	8	9	10	11	12
#legos	1	6	10	36	34	20	28	8

In light of the statement by Murray et al. [21] that each type of stress resulted in a distinct response, it is important to ask whether most of our network legos primarily involve complemented active networks. Over all network legos $(G, \mathcal{P}, \mathcal{N})$, we counted the total size of the positive active networks (those in the sets $\mathcal{P}$) and the total size of the negative active networks (those in the sets $\mathcal{N}$). Interestingly, more than 40% of the active networks appeared in the positive sets, indicating that the network legos we found were not primarily focussed on what made the stresses unique. Rather, a large fraction of the network legos represented features common to multiple stresses. The active networks that appeared most often in the positive form were the two treatments that resulted in cell cycle arrest. Each participated in as many as 119 network legos. In most of these network legos, almost all the other active networks appeared in complemented form. The complements of the cell cycle arrest active networks did not participate in any network legos. This observation indicates that the interactions

activated by cell cycle arrest are quite distinct from the network of interactions activated by the other stresses.

We obtained very good stability and recovery results. Upon the removal of each active network, we were able to recompute each network lego with at least 95% fidelity. We were also able to recover 11 active networks with 100% accuracy by composing network legos. The two active networks we could not recover completely were the double thymidine network (97% recovery) and the thymidine-nocodazole network (86% recovery). When we tested the recoverability of active networks using the blocks at the roots of the DAG connecting statistically-significant blocks, the recovery for these two active networks dropped to 85% and 75% respectively. This result underscores the fact that identifying network legos as those that are maximally statistically-significant in the DAG of blocks is a useful concept.

Since the cell-cycle treatments resulted in active networks that were quite distinct from those for the other stresses, we repeated the analysis after removing the double thymidine and thymidine-nocodazole active networks. The 11 remaining active networks yielded only 77117 blocks (out of the 175099 possible). Of these, 1629 blocks were statistically significant. These blocks yielded 15 network legos. This much smaller set of network legos suggests that a number of the 143 network legos in the complete analysis were needed to capture unique aspects of the cell cycle active networks. Each network lego involved at least seven active networks. No network lego involved all 11 stresses. The ratio of total size of the positive active networks and the negative active networks in the 15 network legos was 1:2. As many as eight network legos had only one active network in $\mathcal{P}$—the fibroblast active network upon treatment with menadione—indicating that this stress results in an active network that is quite unique compared to the other 10 active networks. Of the 11 active networks, we recovered five with complete accuracy and one with 99.9% accuracy. We recovered the remaining with accuracies ranging from 71% to 92%. Taken together, these statistics indicate that the network legos we detect are indeed building blocks of the networks activated in response to the stresses studied by Murray et al. [21].

5 Discussion

We have presented a novel approach for combining gene expression data sets with multi-modal interaction networks. This combination provides a dynamic view of the interactions that are activated in the wiring diagram under different conditions. We represent similarities and differences between the network of interactions activated in response to different cell states both as a set theoretic formula involving cell states and as a network lego, a functional module of co-expressed molecular interactions. A novel contribution of our work is the DAG that relates all cell states (and the active networks corresponding to the cell states). This DAG provides a high-level abstract view of the similarities and differences between cell states.

Acknowledgments We thank Naren Ramakrishnan for useful suggestions on the algorithm for computing blocks. We thank Aravind Subramanian for providing us an updated version of the gene sets in MSigDB.

References

1. R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the Twentieth International Conference on Very Large Databases*, pages 487–499, Santiago, Chile, 1994.
2. S. Armstrong, J. Staunton, L. Silverman, R. Pieters, M. den Boer, M. Minden, S. Sallan, E. Lander, T. Golub, and S. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, 30(1):41–7, 2002.
3. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. the Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, 2000.
4. G. D. Bader, D. Betel, and C. W. V. Hogue. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31(1):248–50, 2003.
5. Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, 21(11):1337–42, 2003.
6. K. Basso, A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human B cells. *Nat Genet*, 37(4):382–90, 2005.
7. S. Bergmann, J. Ihmels, and N. Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*, 2(1):E9, 2003.
8. M. Charikar. Greedy approximation algorithms for finding dense components in graphs. In *Proceedings of APPROX*, 2000.
9. D. Cozma and A. Thomas-Tikhonenko. Kit-Activating Mutations in AML: Lessons from PU.1-Induced Murine Erythroleukemia. *Cancer Biol Ther*, 5(6):579–81, 2006.
10. D. di Bernardo, M. J. Thompson, T. S. Gardner, S. E. Chobot, E. L. Eastwood, A. P. Wojtovich, S. J. Elliott, S. E. Schaus, and J. J. Collins. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotechnol*, 23(3):377–83, 2005.
11. A. P. Gasch, P. T. Spellman, C. M. Kao, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–57, 2000.
12. J. Han, N. Bertin, T. Hao, D. Goldberg, G. Berriz, L. Zhang, D. Dupuy, A. Walhout, M. Cusick, F. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, 2004.
13. L. Hartwell, J. Hopfield, S. Leibler, and A. Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–52, 1999.
14. H. Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, 21 Suppl 1:i213–i221, 2005.

15. T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–40, 2002.
16. G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, 33(Database issue):D428–32, 2005.
17. A. R. Joyce and B. O. Palsson. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*, 7(3):198–210, 2006.
18. H. Lee, A. Hsu, J. Sajdak, J. Qin, and P. Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome Res*, 14(6):1085–94, 2004.
19. B. Lehner and A. G. Fraser. A first-draft human protein-interaction map. *Genome Biol*, 5(9):R63, 2004.
20. N. Luscombe, M. Babu, H. Yu, M. Snyder, S. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–12, 2004.
21. J. I. Murray, M. L. Whitfield, N. D. Trinklein, R. M. Myers, P. O. Brown, and D. Botstein. Diverse and specific gene expression responses to stresses in cultured human cells. *Mol Biol Cell*, 15(5):2361–74, 2004.
22. C. L. Myers, D. Robson, A. Wible, M. A. Hibbs, C. Chiriac, C. L. Theesfeld, K. Dolinski, and O. G. Troyanskaya. Discovery of biological networks from diverse functional genomic data. *Genome Biol*, 6(13):R114, 2005.
23. S. Peri, J. Navarro, R. Amanchy, T. Kristiansen, C. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. Shivashankar, B. Rashmi, M. Ramya, Z. Zhao, K. Chandrika, N. Padma, H. Harsha, A. Yatish, M. Kavitha, M. Menezes, D. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S. Anand, V. Madavan, A. Joseph, G. Wong, W. Schiemann, S. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G. Blobe, C. Dang, J. Garcia, J. Pevsner, O. Jensen, P. Roepstorff, K. Deshpande, A. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363–71, 2003.
24. A. K. Ramani, R. C. Bunescu, R. J. Mooney, and E. M. Marcotte. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol*, 6(5):R40, 2005.
25. D. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. Chinnaiyan. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A*, 101(25):9309–14, 2004.
26. D. R. Rhodes, S. Kalyana-Sundaram, V. Mahavisno, T. R. Barrette, D. Ghosh, and A. M. Chinnaiyan. Mining for regulatory programs in the cancer transcriptome. *Nature Genetics*, 37(6):579–583, May 2005.
27. J. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. Berriz, F. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. Goldberg, L. Zhang, S. Wong, G. Franklin, S. Li, J. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. Sikorski, J. Vandenhaute, H. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. Cusick, D. Hill, F. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–8, 2005.

28. S. Schnittger, T. M. Kohl, T. Haferlach, W. Kern, W. Hiddemann, K. Spiekermann, and C. Schoch. KIT-D816 mutations in AML1-ETO-positive AML are associated with impaired event-free and overall survival. *Blood*, 107(5):1791–9, 2006.
29. S. Schwartz, A. Heinecke, M. Zimmermann, U. Creutzig, C. Schoch, J. Harbott, C. Fonatsch, H. Loffler, T. Buchner, W. D. Ludwig, and E. Thiel. Expression of the C-kit receptor (CD117) is a feature of almost all subtypes of de novo acute myeloblastic leukemia (AML), including cytogenetically good-risk AML, and lacks prognostic significance. *Leuk Lymphoma*, 34(1-2):85–94, 1999.
30. E. Segal, N. Friedman, D. Koller, and A. Regev. A module map showing conditional activity of expression modules in cancer. *Nat Genet*, 36(10):1090–8, 2004.
31. E. Segal, M. Shapira, A. Regev, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2):166–76, 2003.
32. R. Sharan and T. Ideker. Modeling cellular machinery through biological network comparison. *Nat Biotechnol*, 24(4):427–33, 2006.
33. U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–68, 2005.
34. J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–55, 2003.
35. A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, and J. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 2005.
36. A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A*, 101(9):2981–6, 2004.
37. A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. In *Proceedings of ISMB 2002*, pages S136–S144, 2002.
38. A. Tanay, I. Steinfeld, M. Kupiec, and R. Shamir. Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. *Molecular Systems Biology*, 1(1):msb4100005–E1–msb4100005–E10, March 2005.
39. M. L. Whitfield, G. Sherlock, A. J. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown, and D. Botstein. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell*, 13(6):1977–2000, 2002.
40. M. J. Zaki and C.-J. Hsiao. CHARM: An efficient algorithm for closed itemset mining. In *SIAM International Conference on Data Mining*, pages 457–473, 2002.
41. X. J. Zhou, M. C. Kao, H. Huang, A. Wong, J. Nunez-Iglesias, M. Primig, O. M. Aparicio, C. E. Finch, T. E. Morgan, and W. H. Wong. Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol*, 23(2):238–43, 2005.