Integrative Biology

Cite this: *Integr. Biol.*, 2012, **4**, 1415–1427

www.rsc.org/ibiology

PAPER

SAMNet: a network-based approach to integrate multi-dimensional high throughput datasets†

Sara J. C. Gosline,[a] Sarah J. Spencer,[b] Oana Ursu[c] and Ernest Fraenkel*[a]

Received 28th March 2012, Accepted 26th August 2012
DOI: 10.1039/c2ib20072d

The rapid development of high throughput biotechnologies has led to an onslaught of data describing genetic perturbations and changes in mRNA and protein levels in the cell. Because each assay provides a one-dimensional snapshot of active signaling pathways, it has become desirable to perform multiple assays (*e.g.* mRNA expression and phospho-proteomics) to measure a single condition. However, as experiments expand to accommodate various cellular conditions, proper analysis and interpretation of these data have become more challenging. Here we introduce a novel approach called SAMNet, for Simultaneous Analysis of Multiple Networks, that is able to interpret diverse assays over multiple perturbations. The algorithm uses a constrained optimization approach to integrate mRNA expression data with upstream genes, selecting edges in the protein–protein interaction network that best explain the changes across all perturbations. The result is a putative set of protein interactions that succinctly summarizes the results from all experiments, highlighting the network elements unique to each perturbation. We evaluated SAMNet in both yeast and human datasets. The yeast dataset measured the cellular response to seven different transition metals, and the human dataset measured cellular changes in four different lung cancer models of Epithelial-Mesenchymal Transition (EMT), a crucial process in tumor metastasis. SAMNet was able to identify canonical yeast metal-processing genes unique to each commodity in the yeast dataset, as well as human genes such as β-catenin and TCF7L2/TCF4 that are required for EMT signaling but escaped detection in the mRNA and phospho-proteomic data. Moreover, SAMNet also highlighted drugs likely to modulate EMT, identifying a series of less canonical genes known to be affected by the BCR-ABL inhibitor imatinib (Gleevec), suggesting a possible influence of this drug on EMT.

Introduction

Cells respond to external stimuli at many levels, including changes in gene and subsequently protein expression levels, post-translational changes to proteins, changes in subcellular localization and changes in levels of small molecules. While some of these changes can be measured *via* mRNA expression assays,[1] alternative technologies are needed to capture the full response. For example, genetic screens can identify genetic

[a] *Dept. of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: fraenkel-admin@mit.edu*
[b] *Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[c] *Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
† Electronic supplementary information (ESI) available. See DOI: 10.1039/c2ib20072d

Insight, innovation, integration

The increasing use of high throughput technologies in biology has led to an overwhelming amount of data. As the cost of genome-wide assays has dropped, experiments across various cellular conditions at once are no longer uncommon. Here we introduce SAMNet, an optimization algorithm that uses the underlying protein–protein interaction network to integrate results from multiple types of assays across various conditions, highlighting genes and pathways that might have been missed by the original experiments but are relevant to the underlying cellular process. We illustrate how SAMNet can be used to integrate genetic mutant data and mRNA expression data across seven conditions in budding yeast as well as phosphorylation data and mRNA expression data in a model of Epithelial-Mesenchymal transition (EMT) in lung cancer.

mutations that change a cellular response to a particular perturbation,[2] phospho-proteomics assays can identify changes in protein activity,[3] transcription factor binding assays[4] can identify changes in binding activity and epigenetic screens can detect changes in chromatin structure.[5] However, each experiment only detects a fraction of the total cell state, making interpretation of the experiments challenging.

The cataloging of protein–protein interactions across species and conditions into databases such as STRING[6] has fueled the development of computational algorithms that search for relationships between various genes. These algorithms use the published interactome as a blueprint for putative signaling pathways then identify which signaling pathways best explain the changes measured with specific high-throughput assays. Given a set of genetic hits and differentially expressed mRNA, various approaches have been used to identify signaling pathways active in these experiments, such as dynamic programming-based methods,[7] probabilistic models of the underlying pathways,[8] and network-flow based optimization approaches.[9] Other network approaches, such as Steiner tree-based algorithms, have been shown to identify proteins that best explain the presence of genetic hits in the interactome (without expression data).[10] Steiner trees have also been used to explain expression changes downstream of phosphorylation activity.[11,12]

In this work we introduce SAMNet, for Simultaneous Analysis of Multiple Networks, an algorithm that uses a network flow model to integrate two distinct high-throughput experiments across multiple conditions. Our approach is motivated by the fact that cellular responses to many distinct biological perturbations show significant overlap, a fact that has been recognized since pioneering work by Gasch *et al.*[13] As a result, independent analysis of data from different perturbations will be biased toward revealing the common pathways at the expense of the specific responses. By adopting a multi-commodity flow-based approach, SAMNet identifies interactions from the protein–protein interaction network that are unique to each condition.

Network flow algorithms are a family of algorithms that select a combination of edges in a network that provide the best path from a designated source to a designated sink. The earliest mention of network flow in the context of the protein interaction network is the FunctionalFlow algorithm used to ascribe function to unknown proteins by quantifying the flow through the weighted interactome from proteins of known function.[14] ResponseNet, a single-commodity flow algorithm used phenotypic and mRNA expression data to study the effects of alpha-synuclein toxicity.[9] More recently, a multi-commodity variant was used to characterize the results of RNA interference experiments in yeast.[15,16] Information flow models make up a similar class of algorithms that model the interactome as an electrical circuit, where each edge acts a resistor and carries the current from an artificial source to each gene in the network to determine its importance. Information flow algorithms have also been used to integrate genetic and expression data within the protein interaction network[17–20] as well as random walk approaches.[21]

SAMNet uses a constrained optimization formulation based on the multiple commodity flow problem to model multiple experiments simultaneously as "commodities" that must transit from a common source to a common sink through a shared protein interaction network. Each edge in the interaction network has a particular capacity, and therefore must be 'shared' by all commodities. This constraint forces the algorithm to select interactions that are unique to each cellular perturbation, thus avoiding the selection of common stress pathways, a common pitfall of other optimization approaches. We test SAMNet on two distinct datasets. We model the effect of seven different transition metals on the budding yeast *Saccharomyces cerevisiae*[22] through integration of genetic mutant and mRNA expression data. Having shown that the algorithm can identify meaningful biological pathways across the 14 datasets (seven conditions, two assays each), we also used the algorithm in a model of Epithelial-Mesenchymal Transition (EMT) in human lung cancer cell lines.[23]

Our results indicate that SAMNet is a powerful tool for modeling diverse sources of high throughput data across multiple experiments. As the cost of performing these experiments decreases, the relative cost of analysis will only rise. By selecting relevant proteins and interactions that are unique to cellular perturbations, SAMNet provides a crucial step in the preliminary processing of these data and can be used to generate further hypotheses from the data.

Materials and methods

Network-based integration of 'omics' data

We modeled our approach on a previous algorithm, ResponseNet,[9] in which genetic hits were connected to an artificial node representing the "source" of "flow" and the differentially expressed genes were connected to an artificial node representing the "sink". The algorithm then selected the best edges and nodes through which the "flow" could run from the source to the sink based on a cost for each edge, ultimately representing the best combination of protein–protein and protein DNA interactions that explained the genetic and transcriptional data.

Similarly, we represent the proteins, mRNA and their interactions as a graph $G = (V,E)$ where the vertex set V represents proteins and mRNA while the edge set E represents putative physical interactions between them. Fig. 1 depicts G. The vertex set V is comprised of both proteins (squares and circles in Fig. 1) and mRNA (diamonds in Fig. 1). Edges among proteins (solid lines in Fig. 1) are derived from prior knowledge about protein–protein interactions and edges between proteins and mRNA are derived from inferred protein–DNA interaction networks (dashed lines in Fig. 1). A gene is included as an mRNA node if the gene is putatively transcribed by a protein present in the protein-interaction network and a gene is included as a protein node if the translated gene is known to interact with another protein. As such, it is possible to have a gene represented in both mRNA and protein form, as it can exist in both states in the cell.

Network optimization formulation

In the original graph G, there are two subsets of nodes that represent the biological experiment in question, one representing the differentially expressed mRNA for each condition k,

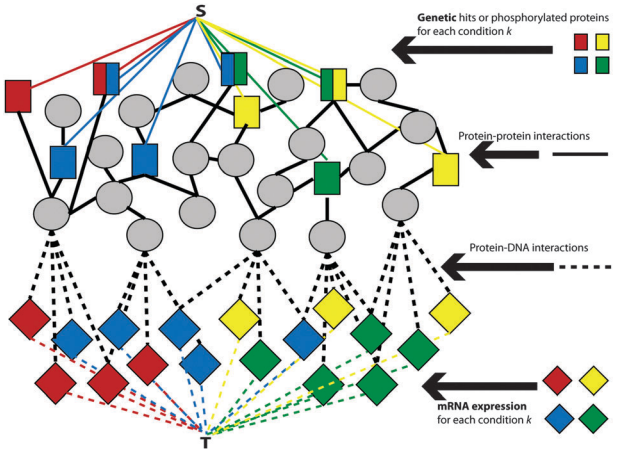

Fig. 1 The integration of four distinct data types into a single weighted graph with the auxiliary nodes S and T. Four different conditions are represented, with the genetic hits/phosphorylated proteins (squares) and differentially expressed mRNA (diamonds) derived from distinct experiments. Internal nodes (circles) are derived from the interactome. Black edges represent data from published interactions, colored edges represent chemical-specific data. Dashed edges represent protein–DNA interactions while solid edges represent protein–protein interactions.

labeled expr_k (diamonds in Fig. 1), and one representing the upstream modifiers, either genetic hits or phosphorylated proteins in each condition k, labeled hits_k (rectangles in Fig. 1). Proteins/mRNA identified in the original experiments that have no known interactions are omitted from the network.

While ResponseNet had a similar formulation, SAMNet differs from ResponseNet by representing each cellular condition as a *commodity*, which is an abstraction derived from the field of operations research to represent a collection of goods that must travel from one point (in this case, the source S) to another (the sink T). Each condition, for instance a cell state with a specific perturbation, is represented by its own commodity. This enhancement requires modifying the ResponseNet optimization from a basic network flow algorithm to a multi-commodity network flow formulation to allow for shared use of the same underlying network G without allowing flow to travel from the hits in one condition to the differentially expressed genes from another condition (hit_i to expr_j).

The non-zero edge weights $w_{ijk} > 0$ in G represent confidence in the interaction between the two proteins and are and equivalent across all commodities. We also add a capacity constraint cap_{ij} for each edge that is set to 1 in the original graph G.

The graph G is then augmented as follows to incorporate the specific perturbation data:

(1) $G = (V,E,C)$, where C represents the set of commodities, or conditions, to be evaluated. The sets of vertices and edges are the same for each commodity.

(2) $V' = V \cup \{S,T\}$, where S and T are auxiliary nodes representing the source and sink of the network.

(3) $E' = E \cup \{(S,i,k)\ \forall i \in \text{hits}_k, \forall k \in C\} \cup \{(j,T,k)\ \forall j \in \text{expr}_k,\ k \in C\}$. This update creates condition-specific edges between the source and genetic hits for a particular condition k, and also between mRNA differentially expressed in condition k and the sink.

4. Weights from the S to genetic hits w_{Sik} represent growth deficiency in the yeast data as defined by Jin *et al.*[22] and absolute log fold change in phosphorylation activity as described by Thomson *et al.*[23] We define the capacities from the source to genetic/phospho-proteomic hits such that they sum to 1 for each commodity: $\text{cap}_{Si} = \sum_{k \in C} \frac{w_{Sik}}{\sum_{j \in \text{hits}} w_{Sjk}}$.

5. Weights w_{iTk} from the mRNA nodes to T represent the absolute log-fold change of the mRNA under perturbation k in the original data.[22,23] We define the capacities from the expression values to the sink as the weights normalized to 1 for each commodity: $\text{cap}_{iT} = \sum_{k \in C} \frac{w_{iTk}}{\sum_{j \in \text{expr}} w_{iTk}}$.

We define the flow variable f_{ijk} to represent the flow from node i to node j for commodity k. We then use CPLEX version 12.4.0 (freely available for academic purposes from the IBM website) to solve the following linear program:

$$\min_f \sum_{k \in C} \left[\sum_{i \in \text{hits}_k} -\log(w_{Sik}) \times f_{Sik} + \sum_{i \in V, j \in V} -\log(w_{ijk}) \times f_{ijk} + \sum_{j \in \text{expr}_k} -\log(w_{jTk}) \times f_{jTk} \right] - \sum_{k \in C} \sum_{i \in \text{hits}_k} \gamma \times f_{Sik} \quad (1)$$

Subject to:

$$\sum_{j \in V'} f_{ijk} = \sum_{j \in V'} f_{jik} \quad \forall i \in V,\ k \in C \quad (2)$$

$$\sum_{i \in \text{hits}_k} f_{Sik} = \sum_{i \in \text{expr}_k} f_{iTk} \quad \forall k \in C \quad (3)$$

$$\sum_{k \in C} f_{ijk} \leq \text{cap}_{ij} \quad (4)$$

$$f_{ijk} \geq 0 \quad (5)$$

This linear program is comprised of an objective function (eqn (1)) and a series of constraints (eqn (2)–(5)) that together identify a putative set of edges that best explain the connection between upstream signaling changes and changes in mRNA expression. The objective function finds a balance between large networks that explain many connections but use low-confidence edges, and small networks that explain very little of the data but use high-confidence edges. This balance is achieved by maximizing the total flow in the network while minimizing the total cost of the weight of each edge multiplied by the flow passing through it (f_{ijk}). The parameter γ is a tuning parameter that effectively controls the size of the network by altering the balance between these two goals. Eqn (2) to (5) are constraints that are required for the following purposes: eqn (2) maintains the conservation of flow, forcing the flow entering a particular node to also leave that node, unless that node is the source S or the sink T. Eqn (3), called demand satisfaction, ensures that all flow is accounted for – everything that leaves the source S must reach the sink T. Eqn (4), the capacity constraint, forces all commodities to share the capacities of the edges. Eqn (5) ensures non-negative flow. The primary difference between this approach and the single-commodity flow in the ResponseNet algorithm[9] is eqn (4), which requires that the combined flow of all commodities

passing through an edge be limited to a single capacity value. This requirement prevents components of the response that are common to many conditions from dominating the networks.

Python scripts that run SAMNet, as well as the ensuing analysis including GO and KEGG enrichment determination are publicly available at http://www.github.com/sgosline/SAMNet.

Yeast transition metal dataset analysis

To evaluate the efficacy of the algorithm we used a published yeast dataset that measured both the growth phenotype and mRNA expression levels upon treatment with different metals.[22] In this study, the yeast deletion library was screened with seven different transition metals, each at their respective EC_{50} concentration (50% total effective concentration). Genetic hits were defined as mutations that cause cells treated with the transition metal to grow at least 50% slower than wild type treated with the same concentration of metal. mRNA expression data was also retrieved from the same set of experiments, and differentially expressed genes ($p < 0.01$ as defined by the original experiment) were included in our final set. The total number of genetic hits and differentially expressed genes are in Table 1.

The genetic hits and differentially expressed genes have very little overlap (at most eight genes for any of the seven commodities), as expected from previous analysis.[22] Furthermore, clustering either the genetic or the expression profiles suggests very different relationships among the transition metals, as shown in Fig. S1A and S1B (ESI†). Fig. S1A (ESI†) illustrates the clustering of growth inhibition values of genetic hits across all metals. Fig. S1B (ESI†) depicts the clustering of mRNA expression changes upon treatment with the same metals. The disagreement between dendrograms illustrates the differences in the two types of data.

To construct graph G described above with the yeast data, we represented edges between proteins with predicted protein–protein interactions derived from the STRING database[6] using interactions with supporting experimental evidence and a confidence score >0.6. Differentially expressed mRNA were connected to the network using predicted protein–DNA interactions derived from published ChIP data binding sites of the entire set of yeast transcription factors and then filtered for known transcription factor motifs as described by MacIsaac *et al.*[24] Only genetic hits that had predicted interactions (either with mRNA or with other proteins) were included in the network. mRNA nodes were distinct from protein nodes to avoid conflating the two types of molecules, as protein interactions cannot occur between un-translated mRNA.

Table 1 Sizes of Yeast metal datasets used for SAMNet. Genetic hits identified in the original screen that were not in the STRING interactome were removed from consideration

Metal treatment	Genetic hits	Differentially expressed mRNA	Overlap
Arsenic	38	566	1
Cadmium	49	898	6
Chromium	59	861	6
Copper	39	815	7
Mercury	3	877	0
Silver	2	814	0
Zinc	38	839	1

On the yeast interactome (6190 nodes and 114 973 edges in G'), the algorithm took $\sim$5 minutes to complete on a 64-bit server with four dual-core processors and 16 GB of RAM. We defined the predicted network as $F = \{f_{ijk} > 0\}$. We selected a γ parameter of 15 to maximize the robustness of the algorithm as described below. The resulting network had 1706 nodes and 2662 edges. The network can be found in Cytoscape format in the data/yeast_metal/metalOutput subdirectory of the online source code repository.

Human EMT dataset analysis

To illustrate the ability of our algorithm to scale to a more complex organism and interpret other types of data, we evaluated previously published data that compared epithelial non-small cell lung cancer (NSCLC) cells to fixed mesenchymal cells as well as to cells with epithelial mesenchymal transition (EMT) artificially induced.[23] To better determine the role of distinct signaling pathways in EMT, this study stimulated the transition *via* three distinct mechanisms (which are known to work together in the cell) to identify the specific influence each pathway may have on the cell. We believed that SAMNet could better identify specific differences between the three modes of EMT induction by comparing them in a network context.

From this publication we collected mRNA expression levels and phospho-protein levels in H358 epithelial cells with EMT induced *via* three different mechanisms – over-expression of Zeb1, over-expression of Snail, or stimulation with TGFβ. mRNA fold changes values were collected from the original manuscript and only those mRNA that exhibited at least an absolute fold change difference of two and $p < 0.05$ were included in the set of differentially expressed mRNA. The authors also collected mRNA expression changes between two epigenetically fixed mesenchymal cells – Calu6 and H1703 – and compared them with the two epithelial cells (H358 and H292). To average the effects of two cell lines together, mRNA were considered to be differentially expressed if the absolute change between the average mRNA in both fixed cell lines and the average mRNA in the epithelial was greater than 1.5. Phospho-peptides were identified by tandem mass-spectrometry with proteins selected as differentially phosphorylated if peptides containing a phosphoserine, phosphothreonine or phosphothyronine were identified at $\geq 95\%$ confidence, fold changes between those peptides were in the upper or lower distribution quartiles ($>75\%$ or $<25\%$) and the changes in expression represented $p < 0.05$ according to a t-test. The number of differentially phosphorylated proteins and differentially expressed mRNA are described in Table 2.

We connected phosphorylated proteins to putative transcription factors using the PSIQUIC interactome,[25] selecting only those edges with a confidence score greater than 0.5. We then connected the interactome to differentially expressed mRNA using putative protein–DNA interactions derived as follows. We downloaded DNase I hypersensitivity data from the ENCODE consortium performed on the A549 cells,

Table 2 Sizes of Human EMT datasets used for SAMNet

EMT State	Phospho-proteins	Differentially expressed mRNA	Overlap
Fixed	132	131	13
Snail	14	1019	5
TGFβ	58	1020	28
Zeb1	14	1019	6

another lung cancer cell line.[5] We then added an edge between a transcription factor t and an mRNA m if the TRANSFAC MATCH algorithm[26] identified a binding site within a DNase I hypersensitive site for t within 5 kb of the transcription start site of m and m was the closest gene to that site. We ran MATCH using the minFP.prf file that provides thresholds for each motif that are high enough to minimize false positive identifications of transcription factor binding sites.

With the human interactome (limited to interactions with a confidence score >0.5) and four conditions, the results took <1 minute to complete. We selected a γ parameter of 14 to maximize the robustness (described below), at which point the final network had 357 nodes and 411 edges. The final network in cytoscape format can be found in the data/human_emt/emtOutput subdirectory in the online source code repository.

Identifying parameters yielding robust networks

To implement SAMNet we needed to identify the optimal parameters for network flow (γ) and construction of the transcription factor–gene network. Ideally, these parameters would show the best performance in recovering true signaling networks. However, as no signaling network is completely known, there are no gold standard datasets that can be used for this purpose. Instead, we assumed that the optimal parameters would identify networks robust to noise in the input data.

To determine the optimal value of γ, we ran SAMNet after omitting fractions of the input data and then calculated the specificity and sensitivity of the networks obtained from the random subsamples. More specifically, we generated 300 different sets of input for each dataset as follows. Fifty of the sets were missing a randomly chosen 10% of the genetic hits (phospho-proteins for the Human dataset) and 50 sets were missing a randomly chosen 10% of the differentially expressed genes. Similarly, 50 randomly chosen inputs were missing either 30% or 50% of either the genetic hits or differentially expressed genes. We varied the network flow parameter γ in both the yeast and human datasets, rerunning the optimization on each of the 300 subsets of the data to identify the value at which the resulting networks were most similar to the original network. Specifically, for each resulting network p, we calculated the fraction of nodes in the original network found in p (specificity) and how many nodes in p were in the original network (sensitivity). We then averaged the specificity and sensitivity measurements across all 100 resulting networks for each fraction of data left out (10%, 30%, 50%) to arrive at the values in Tables S1 and S3, ESI.†

The results are in Fig. S2 and Table S1 (ESI†) for Yeast. Human results are in Table S3 and Fig. S4 and S5 (ESI†). Careful analysis of the values in Tables S1 and S3 (ESI†) revealed that a γ value of 15 for the yeast dataset and 14 for the human dataset result in the highest specificity and sensitivity over all the commodities.

Because using DNAse I hypersensitive sites followed by motif search has only recently become a common way of determining tissue-specific binding sites,[27] we varied the distance between motif match and transcription start site to determine if this could have an impact on the robustness of the network as defined above. While increasing the distance between transcription factor binding site and transcription start site could lead to erroneous edges in the network, we evaluated the specificity and sensitivity of SAMNet using transcription factor binding sites up to one, three, five and ten kilobases upstream of the transcription start site. We found that allowing for transcriptional binding up to five kilobases upstream of the transcription start site provided the network that was most robust to random variation of input data across the distances tested (Table S3 and Fig. S5, ESI†).

Network visualization and functional interpretation

We used Cytoscape[28] to visualize the networks. This tool enabled us to select for high flow nodes or edges as depicted in Fig. 3 and also to focus on different subsets of nodes that we found to be interesting (Fig. 5).

To identify terms that were over-represented in specific commodities within the network, we used the GOstats and Category packages from Bioconductor[11] to compute the hyper-geometric probability of a given GO term or KEGG pathway (respectively) being over-represented within a specific set of terms compared to the entire network. We used GOStats to compute the conditional p-value for the GO enrichment to account for the graphical hierarchy of the ontology because standard false discovery rate (FDR) p-values are not reliable given the relationship between each of the terms in the ontology. For each commodity k, we identified the vertex set n that have at least one edge carrying commodity k and searched for categories with a higher expected number of proteins in n than expected by chance ($p < 0.01$ for Yeast, $p < 0.05$ for Human) given the size of the entire flow network (1706 nodes in Yeast dataset, 357 nodes in Human). While using such a small background reduces the significance of the enrichment p-values, we believe it compensates for biases in the interactome and the input data to only focus on those processes that are distinct for each commodity. The most significant yeast terms are shown in Table 3 (to save space only those terms with $p < 0.001$ are shown, full results are shown in Table S2, ESI†). For the human dataset we found KEGG terms to be more informative and thus included those in Table 4 ($p < 0.05$) but still listed all GO terms ($p < 0.05$) in Table S4 (ESI†). We used a higher p-value threshold for the human data because the lower number of nodes led to a decrease in statistical significance.

We also used functional enrichment to compare nodes identified by SAMNet for a specific commodity and ResponseNet on the same data. For each value of γ we calculated GO terms

Table 3 GO terms enriched ($p < 0.001$) for proteins ascribed to be related to a single metal treatment by SAMNet

Commodity	Term	GOBPID	*p*-Value
Arsenic	Negative regulation of transcription from RNA polymerase II promoter	GO:0000122	1.35×10^{-4}
Arsenic	Tubulin complex assembly	GO:0007021	2.52×10^{-4}
Arsenic	Osmosensory signaling pathway	GO:0007231	3.17×10^{-4}
Arsenic	Cellular response to abiotic stimulus	GO:0071214	3.17×10^{-4}
Arsenic	Regulation of catalytic activity	GO:0050790	4.16×10^{-4}
Arsenic	Filamentous growth of a population of unicellular organisms	GO:0044182	4.33×10^{-4}
Arsenic	Pseudohyphal growth	GO:0007124	5.07×10^{-4}
Arsenic	Regulation of cell communication	GO:0010646	5.85×10^{-4}
Arsenic	Regulation of signaling process	GO:0023051	5.85×10^{-4}
Arsenic	Negative regulation of signal transduction	GO:0009968	9.40×10^{-4}
Cadmium	Covalent chromatin modification	GO:0016569	1.76×10^{-4}
Cadmium	Signaling	GO:0023052	4.45×10^{-4}
Cadmium	Response to DNA damage stimulus	GO:0006974	6.33×10^{-4}
Cadmium	Response to stress	GO:0006950	8.19×10^{-4}
Cadmium	TOR signaling pathway	GO:0031929	9.83×10^{-4}
Chromium	Negative regulation of transcription	GO:0016481	8.13×10^{-6}
Chromium	Negative regulation of RNA metabolic process	GO:0051253	1.08×10^{-5}
Chromium	Negative regulation of biosynthetic process	GO:0009890	1.77×10^{-5}
Chromium	Negative regulation of nitrogen compound metabolic process	GO:0051172	2.75×10^{-5}
Chromium	Negative regulation of biological process	GO:0048519	1.89×10^{-4}
Chromium	Negative regulation of cellular metabolic process	GO:0031324	1.97×10^{-4}
Chromium	Regulation of cell division	GO:0051302	4.35×10^{-4}
Chromium	Chromatin silencing	GO:0006342	5.87×10^{-4}
Chromium	Regulation of gene expression, epigenetic	GO:0040029	5.87×10^{-4}
Copper	Endocytosis	GO:0006897	1.57×10^{-5}
Copper	rRNA processing	GO:0006364	3.29×10^{-5}
Copper	Actin polymerization or depolymerization	GO:0008154	4.01×10^{-4}
Copper	Proteasome assembly	GO:0043248	6.62×10^{-4}
Copper	ncRNA metabolic process	GO:0034660	7.07×10^{-4}
Mercury	Cell wall organization or biogenesis	GO:0071554	2.12×10^{-5}
Mercury	Cellular macromolecule biosynthetic process	GO:0034645	2.28×10^{-5}
Mercury	Signal transmission	GO:0023060	6.99×10^{-5}
Mercury	UFP-specific transcription factor mRNA processing during unfolded protein response	GO:0030969	2.47×10^{-4}
Mercury	Reproductive developmental process	GO:0003006	2.63×10^{-4}
Mercury	Barrier septum formation	GO:0000917	5.99×10^{-4}
Mercury	Regulation of signal transduction	GO:0009966	7.53×10^{-4}
Mercury	Regulation of cellular component size	GO:0032535	8.90×10^{-4}
Mercury	Cell communication	GO:0007154	9.36×10^{-4}
Silver	Meiotic DNA double-strand break formation	GO:0042138	8.45×10^{-6}
Silver	Mitochondrial signaling pathway	GO:0031930	4.05×10^{-5}
Silver	SCF-dependent proteasomal ubiquitin-dependent protein catabolic process	GO:0031146	4.68×10^{-5}
Silver	Double-strand break repair *via* homologous recombination	GO:0000724	8.96×10^{-5}
Silver	G1/S transition of mitotic cell cycle	GO:0000082	6.81×10^{-4}
Silver	DNA catabolic process	GO:0006308	8.59×10^{-4}
Zinc	Golgi vesicle transport	GO:0048193	9.74×10^{-5}
Zinc	Golgi to vacuole transport	GO:0006896	3.97×10^{-4}
Zinc	Cell cycle phase	GO:0022403	8.49×10^{-4}

identified as enriched ($p < 0.05$ according to Fisher's exact test) for both SAMNet and ResponseNet. We then calculated the fraction of terms unique to a particular algorithm compared to all terms identified by both algorithms in Fig. 2C. To illustrate that the terms identified were unique to specific commodities, we performed the same comparison across terms that were not shared between two or more commodities in Fig. 2D. For most commodities, SAMNet was able to identify more unique GO terms for each commodity than the corresponding ResponseNet network.

Scanning network for putative drug targets

To determine if the network was over-represented among various drug-targets, we downloaded a list of drug-protein interactions from PharmGKB (http://www.pharmGKB.org).[29] We then computed, for each drug in the database, the number of targets that were found in the EMT network and computed the probability of finding this many drug targets by chance *via* Fisher's exact test. Full results are in Table S5 (ESI†).

Results

SAMNet identifies condition-specific genes to enable multi-dimensional data analysis

The primary enhancement of SAMNet over previous optimization algorithms is the ability to model multiple conditions simultaneously to reveal condition-specific response pathways. The capacity constraint (eqn (4)) in the optimization criteria requires that the flow through an edge for each commodity must only be enough such that the sum of flow over all commodities is less than the edge capacity. Therefore, the algorithm must consider all commodities when determining how much flow for each commodity can be sent along each edge in the network. The goal is to leverage the availability of

Table 4 KEGG pathways enriched ($p < 0.05$) for proteins ascribed to a single EMT state

Commodity	Term	KEGGID	*p*-Value
Fixed	Chronic myeloid leukemia	05220	0.000478681
Fixed	Epithelial cell signaling in *Helicobacter pylori* infection	05120	0.000680791
Fixed	Adipocytokine signaling pathway	04920	0.004669895
Fixed	Pancreatic cancer	05212	0.00722651
Fixed	Acute myeloid leukemia	05221	0.010175165
Fixed	Jak-STAT signaling pathway	04630	0.013123511
Fixed	Huntington's disease	05016	0.013942901
Fixed	Peroxisome	04146	0.031660636
Fixed	T cell receptor signaling pathway	04660	0.032080957
Fixed	Small cell lung cancer	05222	0.032080957
Fixed	Cell adhesion molecules (CAMs)	04514	0.036944687
Fixed	Melanoma	05218	0.048088525
tgfb	GnRH signaling pathway	04912	0.000957486
tgfb	ECM-receptor interaction	04512	0.006782428
tgfb	MAPK signaling pathway	04010	0.015789683
tgfb	Pathogenic *Escherichia coli* infection	05130	0.018867568
tgfb	Axon guidance	04360	0.024357116
tgfb	Gap junction	04540	0.032635449
tgfb	NOD-like receptor signaling pathway	04621	0.032635449
tgfb	Neurotrophin signaling pathway	04722	0.035733824
Zeb1	Spliceosome	03040	2.21×10^{-5}
Zeb1	Histidine metabolism	00340	0.033517425
Snail	Endometrial cancer	05213	0.002423778
Snail	Adherens junction	04520	0.018937741
Snail	Non-small cell lung cancer	05223	0.041822122

multiple conditions, creating a unified network that highlights pathways that are distinct to each condition without ruling out the possibility that two cellular conditions can indeed share interactions if the experimental results dictate such behavior.

To determine if this enhancement had a significant impact on the result, we compared the SAMNet network with the result of running each perturbation separately with the same value of γ using ResponseNet.[22] Our results in Fig. 2 show the graph statistics for the SAMNet graph compared to the individual ResponseNet graphs using the same parameters. When run with the same parameters, SAMNet identifies for each commodity a subset of the ResponseNet graph run on the same data that is highly enriched in condition-specific nodes and edges when compared to other conditions. Fig. 2A illustrates how 40–60% of the nodes identified by SAMNet are unique to each commodity while ~80% of the nodes in each individual ResponseNet network are shared across all experiments. We get similar results when we compare fraction of unique edges in the network in Fig. 2B, with SAMNet identifying more commodity-specific interactions than ResponseNet run individually on each data set. We also compared SAMNet to the Prize Collecting Steiner Forest (PCSF) algorithm which takes an alternate optimization approach to identify highly likely edges in an interaction network.[30] Our results, depicted in Fig. S7 (ESI†), illustrate that while the PCSF identifies more distinct nodes and edges than ResponseNet, more than 50% of each solution is shared with other commodities. These results indicate that the while PCSF outperforms ResponseNet in identifying relevant networks, it is not as well suited as SAMNet for finding the pathways specific to each member of a set of perturbations.

To further compare SAMNet with ResponseNet with respect to the ability to generate functionally relevant networks we calculated the GO terms enriched for each commodity as described in Materials and Methods. For every value of γ we computed the GO terms enriched for each set of nodes involved in a particular chemical treatment identified by either SAMNet or ResponseNet. The results, shown in Fig. 2C, illustrate that for all but the lowest value of γ SAMNet identifies more GO terms for each commodity than ResponseNet. Because we are interested in GO terms that are unique for each commodity, we eliminated each GO term that was enriched in more than one commodity to determine if SAMNet was still able to identify more unique GO terms than ResponseNet, shown in Fig. 2D.

The full SAMNet network for yeast contains 1706 nodes and 2662 edges (Fig. S6, ESI†). We summarize the final network in Fig. 3, which depicts those edges that consume the highest amount of flow. While it omits most nodes, even the summarized network in Fig. 3 provides a mechanistic explanation of how divergent genetic hits can converge on common yeast stress response pathways as well as shared pathways across various metals. For example, the vacuolar (H-)-ATPase (V-ATPase) complex[9] is targeted by silver and zinc. While many elements of this complex are genetic hits (VPH2, VMA7, VMA8, VMA6, VMA4, VMA2, VMA21 and VMA22 in zinc and VMA9 in silver), SAMNet identifies other members of V-ATPase complex as relevant, such as VPH1, which was not identified as a genetic hit, or VMA21 and VMA2, which were genetics hits in the zinc treatment but not in the silver treatment. Furthermore, the high degree of similarity between the silver and zinc treatments, while indicated in the original clustering of the mRNA expression data (Fig. S1, ESI†) was not evident in the genetic hit data and illustrates how SAMNet can infer pathways even with missing data. Because flow is forced through both genetic hits and differentially expressed genes equally, the algorithm can compensate for missing data in one type of assay.

Across more than half of the conditions, SAMNet implicates RAV1 and SKP1, members of the RAVE complex which is also a regulator of the V-ATPase complex.[31] The large amount of flow passing through these proteins corresponds

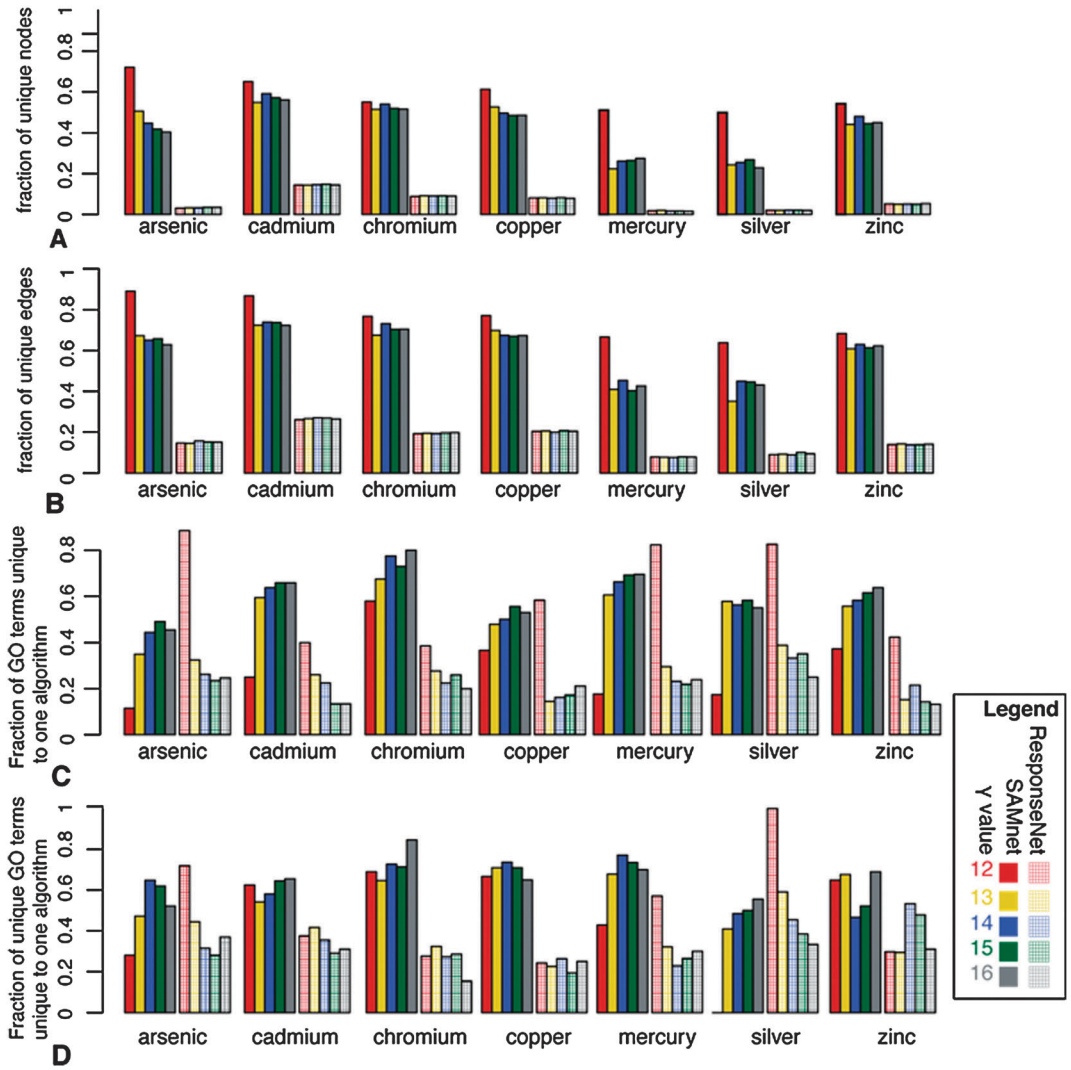

Fig. 2 Comparison of SAMNet and ResponseNet. Fraction of unique (A) nodes and (B) edges in each commodity of SAMNet for the various values of γ compared to the original ResponseNet networks on each independent condition. Fraction of (C) all and (D) unique (to a specific condition) GO terms identified by SAMNet (per commodity) and ResponseNet (on corresponding condition).

to their centrality in cellular processes,[9] confirmed by the essentiality of SKP1. Proteins with flow shared across most of the commodities encompass general stress-related functions such as HSF1, TOR1, GCN4 and GLN3 which are involved in cellular stress as well as MSN2, an environment stress regulator.[32] Lastly we also see metal-specific proteins involved in the shared response, such as YAP family members CIN5 and YAP1.[13] It is important to note that many proteins that are important in the processing of heavy metals, such as VPH1 and YAP1, were not detected in the original genetic or mRNA experiments for any commodities despite their well-known role in metal processing. Overall, however, SAMNet identifies putative nodes involved in each commodity beyond those originally detected and thus facilitates discovery of underlying biological processes involved.

SAMNet can identify biological processes affected by different perturbations

One of the primary challenges of identifying signaling pathways specific to various cellular perturbations is the fact that many responses share similar pathways. By forcing all perturbations to "share" flow through capacitated edges, SAMNet is forced to distribute flow across multiple relevant pathways. As mentioned above, 40–60% of the nodes in each commodity are unique to that commodity, allowing sufficient sample size to search for enriched biological processes in the Gene Ontology (GO) graph. Table 1 shows the GO Biological Process terms uniquely enriched in sets of commodity-specific nodes at $p < 0.001$.

Our approach recovers many of the effects of each metal that were not identifiable with the combined single commodity approach. For example, the proteins ascribed to the mercury commodity are over-represented among cell wall biogenesis-related genes, which has been documented in Hg^{+} resistant strains of Yeast.[33] Cadmium has been identified as playing a role in chromatin modification in human cell lines.[34] Zinc-specific proteins are enriched in vesicle transport which has been observed at the phenotypic level.[35] Lastly, the overwhelming number of RNA and ribosomal-associated terms in the copper commodity also has experimental support,

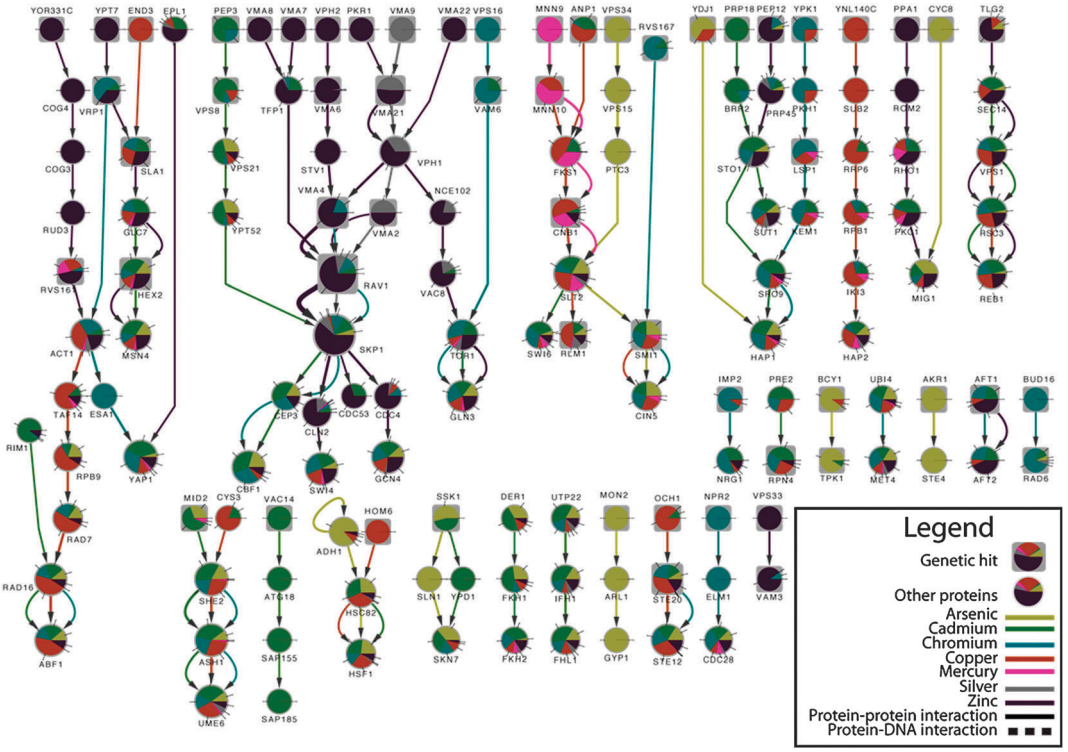

Fig. 3 Subset of Yeast transition metal interactome with edge flow values greater than 0.005. Larger node size indicates more flow, while color indicates conditions in which that node/edge is determined to be active. The direction of edges represents flow from genetic hits (rectangles) to mRNA nodes and the color of the edges represents the commodity that was selected to use that edge. Pie charts correspond to fraction of flow for each commodity passing through the node. Graph generated and filtered by Cytoscape.[28]

as copper has been implicated in hepatic RNA-processing defects in mouse disease models.[36]

SAMNet identifies key mediators of epithelial-mesenchymal transition (EMT)

Having shown that SAMNet can identify condition-specific undetected proteins, we moved to a less characterized system in a more complex organism to determine if the algorithm can identify relevant pathways. Using the same underlying network formulation, we identified protein–protein and protein–DNA interactions that best explain changes in phospho-protein levels upstream of mRNA expression levels across four models of EMT. These four models included H358 cells induced with over-expression of Zeb1, H358 cells induced with over-expression of Snail, H358 cells induced by stimulation with TGFβ and Calu6/H1703 cells to represent an epigenetically fixed mesenchymal model (see Methods). For each model, phospho-proteomic and mRNA expression fold changes were collected and run together as a separate commodity in the SAMNet algorithm. The final network, comprised of 357 nodes and 411 edges, is depicted in Fig. 4a.

The nodes in the network representing a large amount of flow (indicated by the size of the node) are generally well-known mediators of cancer. For example, high-flow nodes GRB2, SRC and EGFR have been identified as regulators of cancer progression.[37] Key transcriptional regulators ESR1 and TP53 were identified by the algorithm as regulating differentially expressed genes across multiple conditions.[38] Fig. 4B depicts the network surrounding β-catenin (CTNNB1) and TCF7L2, also known as TCF4.

This interaction is a hallmark of EMT in which E-Cadherin (CDH1 in Fig. 4B) becomes phosphorylated and releases β-catenin from the membrane, causing it translocate to the nucleus where it activates TCF7L2/TCF4 and LEF1 transcription factors.[39] While these proteins were found to be slightly active in the fixed mesenchymal cell line (grey), they exhibit strong mRNA regulatory effects in the TGFβ and Snail-induced cell lines suggesting that this activity may be related to the transition from the epithelium to the mesenchyme, since it is not present in the fixed cell line. TCF7L2 and CTNN1B were absent from the original experiments due to lack of detectable fold change (as their interaction is activated by translocation). Nevertheless, SAMNet was able to identify these proteins as key mediators of EMT. Based on these results, we would suggest that perturbing the E-Cadherin pathway might disrupt TGFβ and Snail-induction of EMT to a greater degree than Zeb1-induction.

SAMNet can specifically identify signaling changes in various EMT models

The large degree of dissimilarity between transition metal treatments in the yeast dataset made it fairly straightforward to identify condition-specific proteins involved in each perturbation. Therefore it was surprising that, given the high degree of similarity between the various experiments in the EMT data, we were able to identify unique KEGG pathways

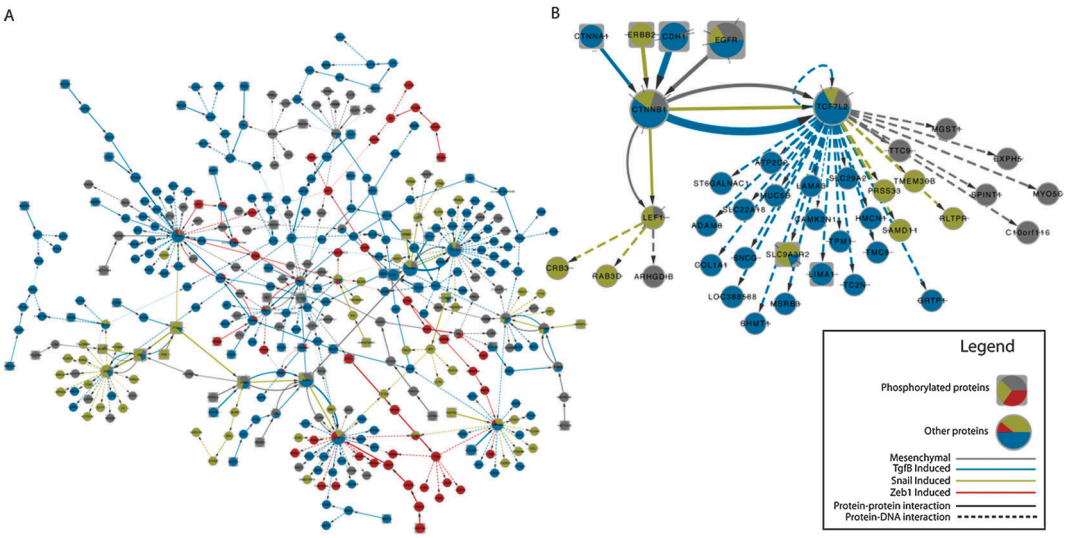

Fig. 4 (A) Full EMT network annotated in a similar fashion to Fig. 3. Phosphorylated proteins are represented by rectangles while non-phosphorylated proteins have grey borders. Pie charts represent flow distribution through nodes, while edge color represents the condition in which that edge was selected. (B) CTNNB1 and TCF7L2 and their interacting nodes.

($p < 0.05$) among nodes ascribed to each condition, described in Table 4.

Many of the over-represented KEGG terms are in line with what is expected of the various cellular conditions. For example, the fixed condition is highly enriched in genes involved in epithelial cell signaling and the TGFβ model includes EMC–receptor interaction related genes. Also, various cancer pathways (including non-small cell lung cancer) are identified across all conditions. However, there are less-expected patterns as well. The JAK-STAT pathway was unique to the fixed mesenchymal model, suggesting that this pathway and its anti-apoptotic effects are not present in cancer cells until after the transition to the mesenchyme. The Snail condition was enriched in adherens junction-related proteins while TGFβ was enriched in Gap junction related proteins, suggesting a possible division of tasks across various EMT signaling proteins. Interestingly, the spliceosome pathway appeared highly enriched among Zeb1-related proteins. A recent study of EMT in a human mammary epithelial cell line (HMLE) identified alternative splicing as a key mechanism of EMT, leading to the many alternatively spliced isoforms that can make cells more invasive.[40] When we searched for enriched KEGG terms in the ResponseNet networks in a similar fashion, we only found enriched terms for the fixed commodity (Table S6, ESI†) suggesting that SAMNet is a necessary improvement to study these pathways.

To further investigate condition-specific pathways, we manually selected sub-networks of interest from the larger EMT network to illustrate how SAMNet can be used to generate further hypotheses from multiple high throughput experiments. Fig. 5A highlights the transcriptional role of ESR1 predominantly in the Snail induced model. Snail has been found to repress ESR1 during EMT.[41] This same work identified significant cross-talk between the TGFβ pathway and the Snail-ESR1 pathway as well, suggesting that the identification of this transcription factor is biologically relevant. Based on the SAMNet results, we suggest that estrogen receptor agonists and antagonists are more likely to alter Snail-induced EMT compared to Zeb1 and TGFβ induction.

Fig. 5B shows interactions with NFKB2, a subunit of the NFκβ complex and a node that is uniquely selected by the induced models. Interestingly, while much is known about the NFκβ complex in its entirety, very little is known about its individual components[42] and this network provides a putative mechanism by which early EMT can regulate cancer progression. We hypothesize that specific inhibition of NFKB2 could inhibit the transition of these cells.

Lastly, we focused on the elements of the spliceosome that were selected by the network, labeled in grey in Fig. 5C along with their immediate neighbors. Interestingly, five out of nine of the spliceosome-related proteins were phosphorylated in either the Zeb1 or Snail induced models. This suggests that phosphorylation of members of the spliceosome can alter the splicing behavior during induction of EMT. In each of the cases cited above, the nodes found by SAMNet could be detected by ResponseNet network under some parameter settings. However, many of the condition-specific events were muted as the nodes were shared by other conditions, making identification of KEGG pathways impossible (Table S6, ESI†). By identifying compact, condition-specific networks, SAMNet makes it easier to generate high-priority hypotheses for experiments.

SAMNet network can be used to identify novel drugs to treat lung cancer

To explore other applications of SAMNet, we scanned the proteins identified in EMT across all genes in PharmGKB, an online repository of drug–gene interactions to see if the network was enriched in targets of known cancer drugs. We performed Fisher's exact test to search for drugs that had a significantly large number of interactions with genes in the full EMT network. We identified 47 compounds with a significant ($p < 0.001$) number of interacting genes in the network.

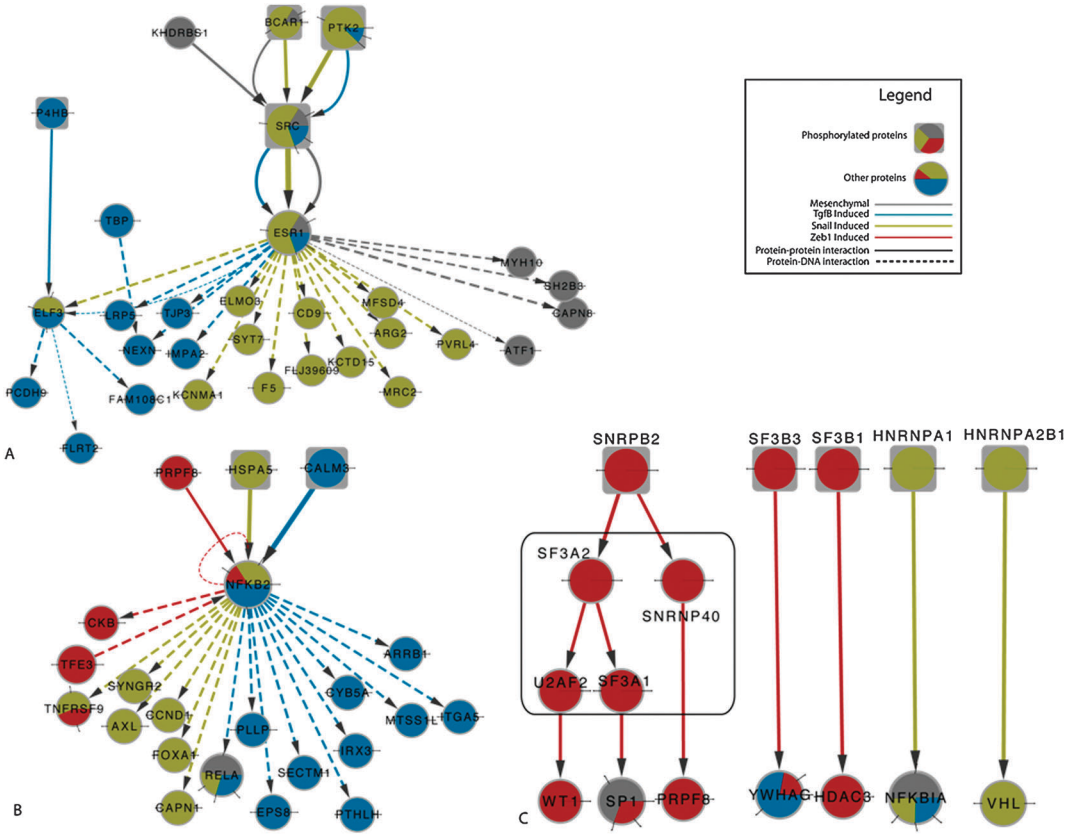

Fig. 5 (A) Role of ESR1 in Snail and TGFβ pathways (B) NFKB2 activity in induced models only (C) Spliceosome-related proteins (according to KEGG) and their interacting partners identified by SAMNet in two induced models. Differentially phosphorylated proteins are represented by rectangles while those inferred by SAMNet are encircled in the black square.

Table S5 (ESI†) describes the drugs from PharmGKB, the overlap of their predicted targets with genes in the network, and the relative contribution of each commodity to the set of genes. Interestingly, the most significant cancer-related compound was imatinib, also known as Gleevec, a BCR-ABL inhibitor that was designed to treat a specific mutation in chronic myelogenous leukemia (CML) that has predicted effects on genes across all four EMT models. While this drug has not been approved to treat non-small cell lung cancers, previous work has found that it potentiates cisplatin to enhance cell death of NSCLC cell line A549.[43] Another study identified that the same compound can inhibit TGFβ-induced cellular proliferation suggesting that Gleevec's synergy with cisplatin is directly related to EMT induction.[44] These two studies, coupled with the over-representation of Gleevec-affected genes in our network, suggest that Gleevec could have an effect on targeting the growth of NSCLC cells through EMT-initiated pathways. The next relevant drugs identified were gemcitabine and gefitinib, both approved drugs for many carcinomas including non small-cell lung cancer.

Discussion

Before the development of high throughput technologies, biological hypotheses were tested one at a time between a control and a test condition (*e.g.* a healthy and diseased tissue). To analyze these results, scientists only needed to plot the values in two dimensions to determine if there was a difference between the samples. However, as the number of conditions has increased along with the number of assays performed, analysis of these high throughput datasets has failed to keep pace. Examples of large, multidimensional datasets include the cancer genome atlas (TCGA),[45] which has a large repository of cancer tumor data across 20 cancers including genetic, mRNA expression, miRNA expression and other forms of data. Within breast cancer alone, there is also a large amount of cell line data[46] measuring the response of 24 different drugs in over 400 cancer cell lines. These datasets provide the ability to identify specific differences between various classes of patients or cell lines with greater statistical power than a basic two-condition test. However, as these large experiments become more common, the need for tools that capitalize on the increased availability of data has only increased.

Here we introduce SAMNet, an algorithm that is able to identify unique pathways active across multiple experiments while still taking into account results of multiple assays. By forcing each experimental condition to share edges in a capacitated and weighted network, our approach can distinguish protein interactions that are distinct to specific conditions from those that are shared. Given the structure of protein–protein interaction networks weighted by evidence, most constrained optimization approaches[9,11] will always select the highest confidence edges that explain the data,

even if these same edges can also explain other, unrelated, data. The selected edges can also be biased toward the experimental platform at hand, and not the differences between the cellular conditions. While these algorithms use permutation tests to identify proteins/pathways that are specifically over-represented in the final network, SAMNet eliminates the need for this step by considering all conditions at once and selecting the best edges for each. While multi-commodity flow has been used previously in the context of the protein interaction network,[15,16] the model was much more constrained with the end goal of identifying relevant RNA interference hits that explain changes in expression of a single gene. SAMNet is data-agnostic and is easily applied to various experimental setups and data types.

Our ability to demonstrate SAMNet on both a yeast system with highly dissimilar treatments as well as a human system with highly related experiments shows the algorithm is a useful and broadly applicable tool to help scientists interpret high throughput data. We illustrate how the algorithm can identify biological processes uniquely affected in one condition *versus* another. By generating specific hypotheses SAMNet can aid experimentalists in designing specific follow-up experiments, such as targeting the sub-networks in Fig. 5, to affect cells in one state (*e.g.* Snail-induced epithelial cells) while not affecting others. As more large scale and collaborate efforts generate data across various conditions and patients we believe that SAMNet will provide a useful tool to integrate these experiments, enhance functional enrichment and provide specific subnetworks that can best explain the observed results.

Funding

This work is supported by NIH grants U54CA112967 and R01GM089903 and used computing resources funded by the National Science Foundation under Award No. DB1-0821391. O.U acknowledges support from the Alexander J. Denner and the Alexander Laats funds within the MIT Undergraduate Research Opportunities Program.

Acknowledgements

The authors would like to thank the following people whose work facilitated this project. Dr Esti Yeger-Lotem, Dr Laura Riva and Dr Kenzie MacIsaac provided scripts to format the protein–DNA interactions. Dr S. Carol Huang provided assistance in formatting the STRING and PSIQUIC interactions. Dr Nurcan Tuncbag assisted in parsing the PharmGKB drug–protein interactions. Meena Subramanian aided in the comparison of the Prize Collecting Steiner Forest algorithm. Dr David Karger and Bernard Haupler provided helpful discussions when this project was in an early stage.

References

1 T. R. Hughes, M. J. Marton, a. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. a. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, a. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard and S. H. Friend, *Cell*, 2000, **102**, 109–126.
2 C. H. Ho, J. Piotrowski, S. J. Dixon, A. Baryshnikova, M. Costanzo and C. Boone, *Curr. Opin. Chem. Biol.*, 2011, **15**, 66–78.
3 J. Cox and M. Mann, *Annu. Rev. Biochem.*, 2011, **80**, 273–299.
4 G. M. Euskirchen, J. S. Rozowsky, C.-L. Wei, W. H. Lee, Z. D. Zhang, S. Hartman, O. Emanuelsson, V. Stolc, S. Weissman, M. B. Gerstein, Y. Ruan and M. Snyder, *Genome Res.*, 2007, **17**, 898–909.
5 ENCODE Project Consortium, *Nature*, 2007, **447**, 799–816.
6 D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen and C. von Mering, *Nucleic Acids Res.*, 2011, **39**, D561–D568.
7 A. C. Haugen, R. Kelley, J. B. Collins, C. J. Tucker, C. Deng, C. A. Afshari, J. M. Brown, T. Ideker and B. Van Houten, *Genome Biol.*, 2004, **5**, R95.
8 C.-H. Yeang, T. Ideker and T. Jaakkola, *J. Comput. Biol.*, 2004, **11**, 243–262.
9 E. Yeger-Lotem, L. Riva, L. J. Su, A. D. Gitler, A. G. Cashikar, O. D. King, P. K. Auluck, M. L. Geddie, J. S. Valastyan, D. R. Karger, S. Lindquist and E. Fraenkel, *Nat. Genet.*, 2009, **41**, 316–323.
10 M. S. Scott, T. Perkins, S. Bunnell, F. Pepin, D. Y. Thomas and M. Hallett, *Mol. Cell. Proteomics* 2005, **4**, 683–692.
11 S.-S. C. Huang and E. Fraenkel, *Sci. Signaling*, 2009, **2**, ra40.
12 M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J.-M. François and R. Zecchina, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 882–887.
13 A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein and P. O. Brown, *Mol. Biol. cell*, 2000, **11**, 4241–4257.
14 E. Nabieva, K. Jim, A. Agarwal, B. Chazelle and M. Singh, *Bioinformatics*, 2005, **21**(Suppl 1), i302–i310.
15 R. Singh and B. Berger, in *International Symposium of Intelligent Sys in Mol Bio (ISMB), PLoS Track*, 2007, Extended abstract.
16 R. Singh, *PhD Thesis*, Massachusetts Institute of Technology, 2011.
17 P. V. Missiuro, K. Liu, L. Zou, B. C. Ross, G. Zhao, J. S. Liu and H. Ge, *PLoS Comput. Biol.*, 2009, **5**, e1000350.
18 X. Ren, X. Zhou, L.-Y. Wu and X.-S. Zhang, *BMC Syst. Biol.*, 2010, **4**, 72.
19 Y. Chen, T. Jiang and R. Jiang, *Bioinformatics*, 2011, **27**, i167–i176.
20 S. Suthram, A. Beyer, R. M. Karp, Y. Eldar and T. Ideker, *Mol. Syst. Biol.*, 2008, **4**, 162.
21 B. Zhang, Z. Shi, D. T. Duncan, N. Prodduturi, L. J. Marnett and D. C. Liebler, *Mol. BioSyst.*, 2011, **7**, 2118–2127.
22 Y. H. Jin, P. E. Dunlap, S. J. McBride, H. Al-Refai, P. R. Bushel and J. H. Freedman, *PLoS Genet.*, 2008, **4**, e1000053.
23 S. Thomson, F. Petti, I. Sujka-Kwok, P. Mercado, J. Bean, M. Monaghan, S. L. Seymour, G. M. Argast, D. M. Epstein and J. D. Haley, *Clin. Exp. Metastasis*, 2011, **28**, 137–155.
24 K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo and E. Fraenkel, *BMC Bioinf.*, 2006, **7**, 113.
25 B. Aranda, H. Blankenburg, S. Kerrien, F. S. L. Brinkman, A. Ceol, E. Chautard, J. M. Dana, J. De Las Rivas, M. Dumousseau, E. Galeota, A. Gaulton, J. Goll, R. E. W. Hancock, R. Isserlin, R. C. Jimenez, J. Kerssemakers, J. Khadake, D. J. Lynn, M. Michaut, G. O'Kelly, K. Ono, S. Orchard, C. Prieto, S. Razick, O. Rigina, L. Salwinski, M. Simonovic, S. Velankar, A. Winter, G. Wu, G. D. Bader, G. Cesareni, I. M. Donaldson, D. Eisenberg, G. J. Kleywegt, J. Overington, S. Ricard-Blum, M. Tyers, M. Albrecht and H. Hermjakob, *Nat. Methods*, 2011, **8**, 528–529.
26 A. E. Kel, E. Gössling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis and E. Wingender, *Nucleic Acids Res.*, 2003, **31**, 3576–3579.
27 H. H. He, C. A. Meyer, M. W. Chen, V. C. Jordan, M. Brown and X. S. Liu, *Genome Res.*, 2012, **22**, 1015–1025.
28 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, *Genome Res.*, 2003, **13**, 2498–2504.
29 E. M. McDonagh, M. Whirl-Carrillo, Y. Garten, R. B. Altman and T. E. Klein, *Biomarkers Med.*, 2011, **5**, 795–806.
30 N. Tuncbag, A. Braunstein, A. Pagnani, S.-S. C. Huang, J. Chayes, C. Borgs, R. Zechina and E. Fraenkel, in *RECOMB*, 2012, pp. 127–1477.
31 J. H. Seol, A. Shevchenko and R. J. Deshaies, *Nat. Cell Biol.*, 2001, **3**, 384–391.

32 P. Natarajan, J. Wang, Z. Hua and T. R. Graham, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 10614–10619.
33 B. Ono, H. Ohue and F. Ishihara, *J. Bacteriol.*, 1988, **170**, 5877–5882.
34 G. Banfalvi, M. Gacsi, G. Nagy, Z. B. Kiss and A. G. Basnakian, *Apoptosis*, 2005, **10**, 631–642.
35 B. Ezaki and E. Nakakihara, *Yeast*, 2012, **29**, 17–24.
36 C. Geoffroy-Siraudin, M.-H. Perrard, F. Chaspoul, A. Lanteaume, P. Gallice, P. Durand and M.-R. Guichaoua, *Toxicol. Sci.*, 2010, **116**, 286–296.
37 A. Gardarin, S. Chédin, G. Lagniel, J.-C. Aude, E. Godat, P. Catty and J. Labarre, *Mol. Microbiol.*, 2010, **76**, 1034–1048.
38 J. S. Biscardi, R. C. Ishizawar, C. M. Silva and S. J. Parsons, *Breast Cancer Res.*, 2000, **2**, 203–210.
39 N. G. Iyer, H. Ozdag and C. Caldas, *Oncogene*, 2004, **23**, 4225–4231.
40 X. Tian, Z. Liu, B. Niu, J. Zhang, T. K. Tan, S. R. Lee, Y. Zhao, D. C. H. Harris and G. Zheng, *J. Biomed. Biotechnol.*, 2011, **2011**, 567305.
41 A. Dhasarathy, M. Kajita and P. A. Wade, *Mol. Endocrinol.*, 2007, **21**, 2907–2918.
42 A. Fusco and M. Fedele, *Nat. Rev. Cancer*, 2007, **7**, 899–910.
43 N. D. Perkins, *Nat. Rev. Cancer*, 2012, **12**, 121–132.
44 S. Matsuyama, M. Iwadate, M. Kondo, M. Saitoh, A. Hanyu, K. Shimizu, H. Aburatani, H. K. Mishima, T. Imamura, K. Miyazono and K. Miyazawa, *Cancer Res.*, 2003, **63**, 7791–7798.
45 Cancer Genome Atlas Research Network, *Nature*, 2008, **455**, 1061–1068.
46 J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel and L. A. Garraway, *Nature*, 2012, **483**, 603–607.