BIOINFORMATICS

Vol. 29 ISMB/ECCB 2013, pages i210–i216
doi:10.1093/bioinformatics/btt240

A context-sensitive framework for the analysis of human signalling pathways in molecular interaction networks

Alexander Lan[1], Michal Ziv-Ukelson[1,*] and Esti Yeger-Lotem[2,3,*]

[1]Department of Computer Science, [2]Department of Clinical Biochemistry and Pharmacology and [3]National Center for Biotechnology in the Negev, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

ABSTRACT

Motivation: A major challenge in systems biology is to reveal the cellular pathways that give rise to specific phenotypes and behaviours. Current techniques often rely on a network representation of molecular interactions, where each node represents a protein or a gene and each interaction is assigned a single static score. However, the use of single interaction scores fails to capture the tendency of proteins to favour different partners under distinct cellular conditions.

Results: Here, we propose a novel context-sensitive network model, in which genes and protein nodes are assigned multiple contexts based on their gene ontology annotations, and their interactions are associated with multiple context-sensitive scores. Using this model, we developed a new approach and a corresponding tool, ContextNet, based on a dynamic programming algorithm for identifying signalling paths linking proteins to their downstream target genes. ContextNet finds high-ranking context-sensitive paths in the interactome, thereby revealing the intermediate proteins in the path and their path-specific contexts. We validated the model using 18 348 manually curated cellular paths derived from the SPIKE database. We next applied our framework to elucidate the responses of human primary lung cells to influenza infection. Top-ranking paths were much more likely to contain infection-related proteins, and this likelihood was highly correlated with path score. Moreover, the contexts assigned by the algorithm pointed to putative, as well as previously known responses to viral infection. Thus, context sensitivity is an important extension to current network biology models and can be efficiently used to elucidate cellular response mechanisms.

Availability: ContextNet is publicly available at http://netbio.bgu.ac.il/ContextNet.

Contact: estiyl@bgu.ac.il or michaluz@cs.bgu.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Complex diseases and viral infections are among the major problems in human health today. In an effort to broaden our understanding of the molecular basis of these diseases, they are increasingly interrogated using a variety of large-scale experimental techniques. Major techniques include sequencing efforts to reveal disease-related mutations, mRNA profiling to reveal genes that are differentially expressed during disease and siRNA screens to reveal disease-related proteins (e.g. Shapira *et al*., 2009), thereby revealing distinct subsets of the genes and proteins involved.

*To whom correspondence should be addressed.

Recent studies demonstrate the strength of integrative approaches in broadening our understanding of disease processes (reviewed in Ideker and Sharan, 2008; Schadt, 2009). Central to many integrative approaches is the molecular interaction network (interactome) paradigm, where interactome nodes represent proteins or genes, and interactome edges represent their physical and regulatory interactions. Interactomes provide a convenient framework for exploring the context within which disease genes operate, and they were successfully used to illuminate new disease genes (e.g. Guan *et al*., 2012; Magger *et al*., 2012), and their functions, as recently reviewed by Barabasi *et al*. (2011).

Because of the importance of signalling paths in health and disease, several computational efforts exploited the interactome framework for their elucidation. By connecting mutated proteins with their downstream differentially expressed targets, Yeang *et al*. (2004) identified intermediate proteins in the paths and assigned directionality to undirected protein–protein interactions (PPIs). Later studies identified interactome sub-networks relating the results of high-throughput genetic screening and mRNA profiling (Suthram *et al*., 2008; Tuncbag *et al*., 2012; Yeger-Lotem *et al*., 2009; Yosef *et al*., 2009). And yet another set of studies computed putative signalling paths by connecting membrane proteins to transcription factors (Bebek and Yang, 2007; Steffen *et al*., 2002; Tuncbag *et al*., 2012), while limiting the types and relative order of the proteins on the path (Scott *et al*., 2006; Steffen *et al*., 2002; Tuncbag *et al*., 2012). Although based on different computational techniques, each of these studies relied on a typical network representation, where each edge is assigned a single score based on its estimated reliability (e.g. Szklarczyk *et al*., 2011) or relevance to a specific cellular process (e.g. Cakmak and Ozsoyoglu, 2007; Myers *et al*., 2005; Yeger-Lotem *et al*., 2009).

Yet, single edge scores fail to capture the complexity of biological systems, where the activation of a specific protein may lead to multiple responses, depending on the current cellular condition. We illustrate this phenomenon using the human protein GRB2, an epidermal growth factor receptor-binding protein that is known to mediate several cellular signalling cascades (Fig. 1). Although GRB2 physically interacts with many different proteins, recent experimental analysis of its physical interactions has shown that its sub-network remodels itself dramatically in response to different stimuli (Bisson *et al*., 2011). For example, in the context of viral infection GRB2 tends to interact with interferon regulatory factor IRF5, whereas in the context of insulin signalling, it tends to interact with the insulin receptor substrate IRS1. Thus, examining GRB2 physical interactions regardless of cellular context will mask this important distinction.

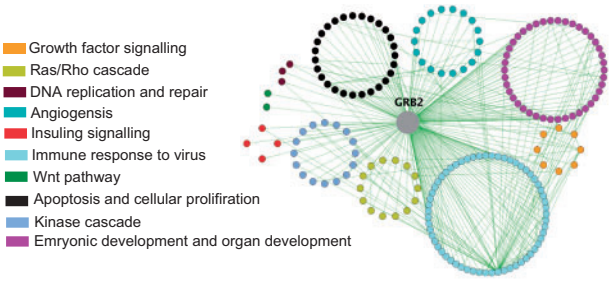


Fig. 1. The human protein GRB2 interacts through high-confidence PPIs with proteins from 10 distinct cellular processes

Here, we describe a novel computational framework that aims to capture the context dependence of molecular interactions. There are several methods that search for paths or transitions between functional contexts in interactomes (e.g. Banks *et al.*, 2008; Pandey *et al.*, 2007). In contrast, the novelty of our approach is not in the consideration of context but rather in the proposed computational model that allows for the consideration of multiple candidate context pairs for each molecular interaction. We provide a context-sensitive algorithm that scores context-specific paths leading from a source protein to a differentially expressed gene. The algorithm selects a single context pair per interaction, based on the context selected for the preceding interaction in the path. Paths are then scored according to the selected context pairs, and high-scoring paths are reported.

Context dependence has often been associated with tissues or cell types (Schaefer, 2012), and with different biological states, such as cell cycle stages (de Lichtenberg *et al.*, 2005) or response to stimuli (Barrios-Rodiles *et al.*, 2005). In this study, we demonstrate our context-sensitive framework by using gene ontology (GO) terms as the biological context. We first show the validity of the proposed model by using a set of manually curated human pathways (Paz *et al.*, 2011). We then demonstrate the ability of the framework to identify relevant interaction paths by analysing the response of human primary lung cells to influenza infection (Shapira *et al.*, 2009). Notably, a significant fraction of the proteins that the algorithm predicted were indeed found to affect the infection process, and the success rate of the prediction was significantly correlated with the path score. We implemented our proposed framework as a tool, ContextNet, which is publicly available at http://netbio.bgu.ac.il/ContextNet.

2 RESULTS

Our results include the proposed context-sensitive framework, its implementation as a tool, a statistical assertion that known cellular pathways in human are indeed context-sensitive and finally the application of our framework to identify and interpret the proteins and pathways underlying the response of human cells to viral infection.

2.1 A context-sensitive framework for identifying signalling paths

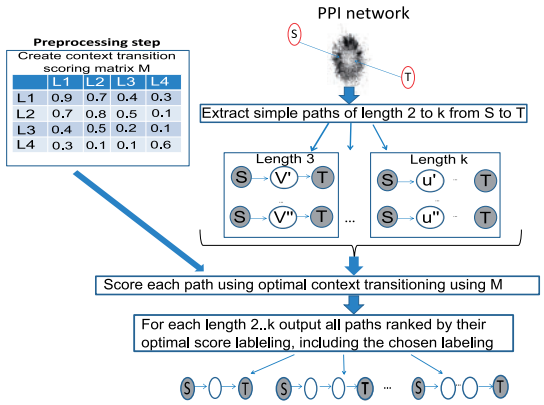
Our proposed framework consists of an interactome model, a context-based scoring scheme and a path interpretation and scoring algorithm.

2.1.1 Interactome model Our model of the human interactome consists of distinct nodes that represent either human proteins or genes, and edges that represent their experimentally detected PPIs and protein–DNA interactions that were downloaded from several databases (see Section 5), resulting in 176 849 interactions among 14 362 proteins and genes. We then added context to the network by using GO annotations (Ashburner *et al.*, 2000) as follows. Each GO annotation was considered as a distinct label and was associated with the corresponding gene and protein nodes. Each interaction was associated with the Cartesian product of the labels of the interacting nodes.

2.1.2 Context-based scoring matrix We constructed a context-transition scoring matrix, M, which assigns to each pair of labels a score that reflects the likelihood of the pair in known pathways. To compute the context-transition scores, we used the SPIKE database of manually curated human pathways (Paz *et al.*, 2011). Specifically, we extracted from SPIKE all signalling paths that connect a protein to a gene with the last edge in the path being a transcription regulation edge. We then directed these 18 438 simple paths from the protein to the gene and combined them into a set of 6762 unique directed interactions. From these interactions, we calculated the conditional probability $P(l_j \mid l_i)$ that reflects the likelihood of observing a directed interaction pointing from a node with label l_i to a node with label l_j. We further fine-tuned this context-transition scoring scheme as described in Section 5.

2.1.3 Algorithm for context-sensitive path interpretation and scoring Our framework identifies top-scoring context-sensitive paths connecting a protein to a differentially expressed gene. An overall illustration of the framework is shown in Figure 2. Given a source protein and a target gene, the algorithm first identifies in the network all simple paths of lengths from 2 to k that connect the source to the target, where k is a predefined user parameter. Next, it uses the context-transition scoring matrix M (described earlier in the text) to rank each path, favouring the strongest contextual interpretation. Finally, the algorithm returns the top scoring paths, along with the chosen label for each node in each path.

We now turn to describe the labelling of a specific candidate path, P (Fig. 3). As each node in P is associated with several labels, the optimization problem at hand is that of selecting an ordered set of labels, one label per each node in P, such that the sum of context-transition scores for consecutive labels in this ordered set is maximized. For this purpose, our method constructs for P a directed acyclic graph, denoted the 'context-label network', as follows. Each potential functional label of a node at index j of P, appears as a contextual-label vertex in the j-th column of the corresponding context-label network. A directed edge is added between labels x and y in consecutive columns of the context-label network, and its weight is set to the context-transition score of the two labels, $M(x, y)$. Additional 'skip' edges with constant gap scores are added to the context-label network, to increase interpretation flexibility by supporting poor or lacking context annotations of some nodes. Each putative contextual interpretation of P then corresponds to a directed path through its context-label network, where the path begins in one of the vertices of the first column, ends in one of the vertices



	L1	L2	L3	L4
L1	0.9	0.7	0.4	0.3
L2	0.7	0.8	0.5	0.1
L3	0.4	0.5	0.2	0.1
L4	0.3	0.1	0.1	0.6

Fig. 2. A high-level overview of our framework for computing context-sensitive molecular interaction paths

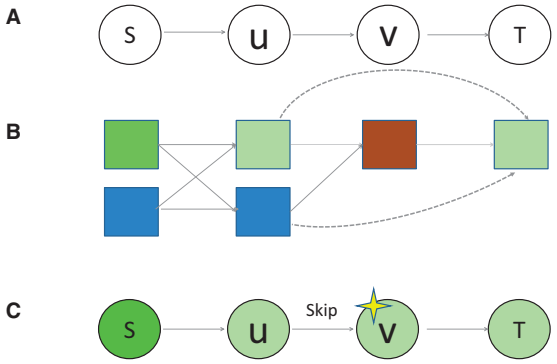


Fig. 3. (**A**) A sample path P connecting a source node S to a target node T. (**B**) Below each node is the set of its labels, where shades of colours are used to denote label similarity according to the context-transition scoring matrix. Edges connect labels corresponding to consecutive nodes in P, and a dashed edge demonstrates a legitimate context gap (Skip step). (**C**) The best-scoring label assignment for P

of the last column and traverses through at most one vertex from each column of the context-label network. Our method computes a heaviest context-labelling path for *P* via a dynamic programming algorithm, as described in Section 5.

2.1.4 ContextNet publicly available tool We implemented our framework as an interactive internet tool and made it publicly available in http://netbio.bgu.ac.il/ContextNet. Given an input consisting of source proteins and target genes, our tool enumerates simple paths in the human interactome connecting the two sets, computes their interpretations and ranks them by their context-labelling scores. The output reports the top-scoring paths and their contextual interpretation.

3 APPLICATIONS OF THE FRAMEWORK TO THE INTERPRETATION OF HUMAN SIGNALLING AND VIRAL INFECTION PATHWAYS

3.1 Known cellular signalling pathways are context sensitive

Our first step was to validate whether known cellular paths are context sensitive. For this purpose, we exploited the SPIKE database of manually curated pathway maps, where each map describes a specific cellular pathway composed of tens of proteins (Paz *et al.*, 2011). We applied our framework to 21 of these maps in a leave-two-out statistical significance test. Specifically, we computed a context-transition scoring matrix *M* based on the paths included in 19 maps, and then used the matrix *M* and the label-selection algorithm to calculate the best score for each signalling path in the two left-out maps. To estimate the statistical significance of the context-labelling score of each path, we repeatedly randomized the set of labels associated with each node and recalculated the path score (see Section 5). If the original path was not context sensitive, one would expect that its original score would be similar to its scores based on randomized labels. However, if the original path was indeed context sensitive, then its original score would be significantly higher than scores based on randomized labels ($P \leq 0.05$). We found that for paths of length two edges, 43% of the 195 paths scored significantly better than random. Furthermore, >70% of the 650 paths of length 3 and >85% of the thousands of paths of length 4–5 scored significantly better than random (Fig. 4). Based on this, we conclude that most manually curated signalling paths of length 3–5 are indeed context sensitive.

3.2 The context-sensitive framework successfully identifies proteins associated with influenza infection of human cells

During infection, influenza proteins were shown to interact with human proteins to recruit the cellular mechanism for viral proliferation. However, the pathways and the intermediate proteins involved in the infection process are just beginning to emerge. In an effort to identify these critical pathways, several large-scale analyses were recently performed. In particular, (Shapira *et al.*, 2009) reported a large-scale analysis of influenza infection of human primary lung epithelial cells, in which they identified PPIs between 10 influenza proteins and 87 human proteins and performed extensive mRNA profiling of the infected human cells. Based on these data, Shapira *et al.* (2009) predicted the involvement of 1756 human genes in the infection process, which they tested by RNA silencing. They found that 616 of the 1756 tested genes were indeed siRNA positive, namely, had a significant effect on viral propagation and interferon production when silenced. Here, we took advantage of this wealth of information to assess our context-sensitive framework and to identify potential signalling paths through which viral proteins may modify the cellular transcriptional program. To this end, we calculated the set of all simple paths linking the human interactors of the viral proteins to each of the human genes exhibiting differential expression after infection (see Section 5). We then scored each path using our context-sensitive algorithm, focusing on paths of length three to five.

Figure 5A demonstrates the biological relevance of the top 5% scoring paths compared with a background set consisting of all paths of same length. The biological relevance was measured by the percentage of paths that contained at least one connecting intermediate protein (not source or target) that was found to be siRNA positive in the experiment described earlier in the text. As

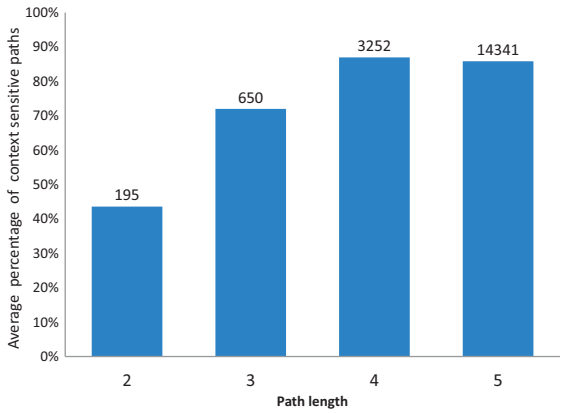


Fig. 4. Results of the SPIKE randomization test for paths of lengths two to five edges are shown. The bars indicate the percentage of paths, per length, that scored significantly higher than random

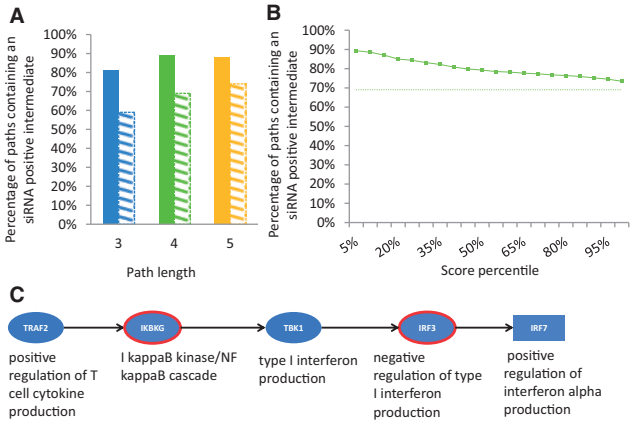


Fig. 5. (**A**) The biological relevance (y-axis) of the top 5% scoring paths (filled bars) compared with a background set consisting of all paths of same length (striped bars), for lengths 3–5 edges (x-axis). Top-scoring paths were more likely to include a biologically relevant protein. (**B**) A graph showing the correlation between context labelling score of a path (x-axis, in score-ranking percentiles) and its likelihood to contain a biologically relevant intermediate protein apart from the source and target (y-axis), for paths of length four. The x-axis shows the percentile of top-scoring paths of the 49 002 scored paths, which constitute 1.24% of the simple paths of length four between the sources and targets. A dotted line marks the background distribution over all simple paths of length four. Top-scoring paths were more likely to contain a biologically relevant intermediate protein (Pearson $r = 0.98$, P-value $< 10^{-12}$). Similar statistically significant correlations were obtained for paths of lengths three and five (Supplementary Fig. S1). (**C**) A top-ranking path connecting the viral interacting protein TRAF2 to the differentially expressed gene IRF7. This path ranked best of the 998 connecting paths for this source and this target. Two of the three intermediate proteins in this path were found to be siRNA positive (red border). Proteins in the path have closely related labels relevant for infection, as shown below each protein

shown, in all paths of lengths three to five, the top-scoring paths were more likely to include a biologically relevant protein. The advantage of top-scoring context-sensitive paths was also observed when compared against shortest paths of similar lengths (Supplementary Fig. S1). We then extended this analysis to test whether the context-labelling score was also correlated with the biological relevance. Indeed, we found that paths of higher scores were also more likely to contain a connecting biologically relevant protein. Figure 5B shows the correlation for paths of length four (Pearson $r = 0.98$, $P < 10–12$), and similar statistically significant correlations were obtained for paths of length three and five (Supplementary Fig. S2). Figure 5C exemplifies a top-ranking path connecting the viral interacting protein TRAF2 to the differentially expressed gene IRF7. This path ranked best of the 998 connecting paths for this source and this target. Notably, two of the three intermediate proteins in this path were found to be siRNA positive, and the contexts that our framework selected for them were indeed related to viral infection as shown in Figure 5C. These results demonstrate again that our context-sensitive framework helps identify biologically relevant proteins and contexts.

3.3 Highlighting and interpreting the multi-faceted functionality of the viral protein PB2

The PB2 protein is a subunit of the influenza virus RNA polymerase, and it is known to be a major virulence determinant of influenza viruses. It was recently demonstrated that PB2 regulates interferon expression during infection (Graef *et al.*, 2010; Shapira *et al.*, 2009). However, the molecular mechanisms by which it acts are just beginning to emerge (Graef *et al.*, 2010). To illuminate these mechanisms, we applied our framework to identify and interpret PB2 downstream interactions. Therefore, we ranked, by context-sensitive scores, the millions of paths connecting the 28 human proteins that were found to interact with PB2 to the 527 target genes that were found to be differentially expressed during influenza infection (see Section 5). We combined the 1% top-ranking paths into the network shown in Figure 6. Importantly, the network we obtained clearly highlights the role of PB2 in interferon regulation. The four human proteins that interact with PB2 through top-ranking paths were all found as likely upstream regulators of interferon expression (Fig. 6, red sub-network). Moreover, for five of the eight differentially expressed genes in the PB2-induced sub-network, our framework selected interferon-related context labels. Notably, the assignment of interferon labels to these genes is statistically significant, as only 39 of the 527 target genes are associated with interferon (Fisher exact test $P = 8.2*10^{-5}$). Thus, our framework correctly uncovered the key downstream effect of PB2 and suggested the cellular pathways by which it acts.

4 DISCUSSION

A well-known limitation of current computational models of PPI networks is the weak handling of interaction context. PPI network models are typically constructed by combining interactions from various measurements, regardless of the biological context in which they were measured, such as specific stimuli, tissues, cellular components and disease states. Previous context-sensitive approaches to network interpretation limited molecular interaction pathways to a single context, such as a single tissue or cell type (Schaefer, 2012), or followed a predefined context-transition template, e.g. defining a flow from membrane to nucleus

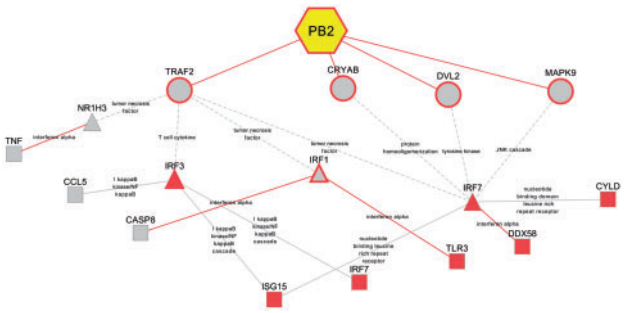


Fig. 6. The PB2 network was computed by merging the top 1% scoring paths connecting PB2 interactors and genes that were differentially expressed on infection. Full lines represent direct physical interactions; dashed lines represent indirect interactions, computed by omitting the intermediate nodes along the path. Triangle nodes represent transcription factors, square nodes represent differentially expressed genes and circular nodes represent direct viral interactors. The labels chosen by our framework are depicted next to nodes. The interferon-labelled nodes are shown in red, and nodes and edges leading from PB2 to these nodes have a red border

(Scott *et al*., 2006). Here, we presented a novel framework that assigns biological context to molecular interaction pathways while allowing for context variations and switches that may not be obvious to the outside observer before our analysis. Our framework also computes a context-sensitivity score for the context-annotated pathway, which can be used to prioritize pathways.

To capture the dynamic context of molecular interaction networks, we added another dimension to the PPI network model, which takes into account, for each node, a set of labels corresponding to its potential functions. We also added edges to connect these labels, and set the weights of these edges to reflect the probability of transitions between the connected labels. The weights of these transitions were inferred from known human pathways. Our model is also flexible enough to provide some support for poorly annotated proteins by leaning on the labels assigned to their upstream and downstream neighbours. This is achieved via special skip edges, thus enabling the inclusion and interpretation of the 16% of the interacting proteins that are not yet annotated.

The proposed framework can be generalized to handle various contexts. In this study, we used GO terms as context labels and known human pathways as guides for constructing the label-transition–scoring matrix. Our analysis of SPIKE maps revealed that proteins in known pathways were mostly associated with GO biological process terms (76%) and not with molecular function (14%) or cellular component GO terms (10%). Therefore, we focused our analysis on GO biological process terms, which we further filtered to remove high-entropy, low–information-content terms. The remaining terms were associated with a relatively small number of genes (median of 19 genes per term, including genes associated with descendent terms), indicating the high specificity of these terms.

We validated our framework using thousands of known human pathways. We found that >70% of the known paths of lengths three to five edges were indeed context sensitive (Fig. 4).

We proved the use of our framework by applying it to reveal the molecular interaction paths involved in viral infection. Using our method, we successfully identified biologically relevant context-sensitive paths connecting viral proteins with the downstream human transcriptional response. We implemented our framework as an interactive internet tool and made it publicly available in http://netbio.bgu.ac.il/ContextNet.

We have shown the value of introducing GO context to the analysis of signalling pathways. Future work could include optimization variants of this problem that will be applicable to longer paths and to a wider scope of context. For example, the current algorithm selects the context of an interaction based on the context selected for the preceding interaction in the path. Important extensions would be to reflect longer contextual histories that go beyond the first-order neighbourhoods and to extend the sought context units from linear paths to networks. Another extension would be to integrate additional context schemes, such as pathway context enrichment (Pandey *et al*., 2007), network schemas (Banks *et al*., 2008), semantic similarity (Pesquita *et al*., 2009), tissue associations (Barshir *et al*., 2013) or protein localization (Scott *et al*., 2006), and to enhance them with interaction-confidence scores. The new framework we presented and its extensions may be applied to a variety of network-related problems where context-sensitive relationships are meaningful.

5 METHODS

5.1 Human interactome model

Experimentally detected human undirected PPIs were assembled from four major PPI databases, including BIOGRID (Stark *et al*., 2011), DIP (Salwinski *et al*., 2004), IntAct (Aranda *et al*., 2010) and MINT (Ceol *et al*., 2010). Directed transcription regulation interactions between transcription factors and their target genes were downloaded from the TRANSFAC database (Matys *et al*., 2006). We also included manually curated directed and undirected interactions from the SPIKE database (Paz *et al*., 2011). GO terms and annotated human proteins and genes were downloaded from the GO database (Ashburner *et al*., 2000), February 2012 release. To increase the specificity and reliability of the GO data used in this study, we filtered out GO terms assigned to >600 genes and GO annotations with evidence codes IPI, NAS and ND. Each GO term was denoted by a distinct label ℓ.

5.2 Learning the label-transition scoring matrix M

To study the frequency of label transitions in signalling paths, we exploited the manually curated pathway maps from the SPIKE database (Paz *et al*., 2011). From each map we extracted the set of all simple paths linking a protein to a gene with the last edge in the path being a transcription regulation edge. We then directed each path from the upstream protein to its downstream target gene, and combined all paths per map into one set of unique directed edges. For each node v, we denote as L_v the set of all labels of v (GO annotations in this study). For each edge (u, v) in some path, and for each $x \in L_u$ and $y \in L_v$, we computed the label pairwise frequency by counting how many times y appears after x in some edge in the set of all edges. This yielded the set F,

consisting of pairs of consecutive labels, annotated by their frequencies.

We observed that some labels in P are non-specific, for example, 'Transcription Factor activity', which always appears as the target node in every edge pointing to a transcription factor. Therefore, we applied an entropy measurement criterion to filter out non-specific labels from F (i.e. labels with high entropy). For this, we calculated for every ordered pair, $(x, y) \in F$, the forward conditional probability $P(y \mid x) =$ (Occurrences of y after x)/(Total occurrences of x), as well as the backward conditional probability $P(x \mid y) =$(Occurrences of x prior to y)/(Total occurrences of y). It is important to note that $P(x \mid y)$ and $P(y \mid x)$ are not necessarily the same. Therefore, for each label x, we computed its forward individual entropy $h(Y \mid x)$ across all occurrences of consecutive label pairs $(x, y) \in F$, as $h(Y \mid x) = \sum_{y \in Y} -P(y \mid x) log(P(y \mid x))$. Note that, in the aforementioned formulation, x is a specific label (thus denoted by a lowercase letter) and Y is a random variable (thus denoted by an uppercase letter). All pairs $(x, y) \in F$, such that the forward individual entropy of label x was found to exceed a threshold (0.7), were filtered out from F. Similarly, we calculated the backward individual entropy $h(X \mid y) = \sum_{x \in X} -P(x \mid t) log(P(x \mid t))$ for each label y and filtered out all pairs $(x, y) \in F$ such that the backward individual entropy of label y was found to exceed a threshold (0.7). In-between filtration steps, both individual and pairwise label frequencies were re-calculated based on the remaining labels in F. Finally, we set $M(x, y) = log_{10}(1 - P(y \mid x))$, for all $(x, y) \in F$.

We used 1428 terms, covering 7001 genes, which constitute 48.7% of the interactome. These terms were then filtered based on their entropy, leaving 1203 terms covering 5919 genes, which are 41.2% of the interactome. To examine the level (within the GO hierarchy) of terms that is useful for context definition, we computed the size of each GO term appearing in M. The size of a GO term was defined as the number of genes associated with it, including the genes associated with its descendant GO terms. We found that the GO terms in the entropy-filtered M had a median size of 19 genes. This indicates that GO terms that have a small size, and, therefore, are low level, are more informative than high-level, non-specific terms that are associated with many genes. The GO terms participating in M, their entropy and the number of genes they cover, are provided in the ContextNet website at: http://netbio.bgu.ac.il/ContextNet/SuppTable1.xlsx.

5.3 The dynamic programming algorithm for path interpretation

To compute the strongest contextual interpretation of paths from the PPI graph, our algorithm uses the pre-computed context-transition scoring matrix M computed as described in the previous section. Given, as input, a path P and the context-transition scoring matrix M, the algorithm computes an output consisting of the score for the strongest contextual interpretation of P, as well as the corresponding annotation of the nodes of P, in form of a sequence of pairs $<(S, \ell_s), (v_2, \ell_2)..(v_k, \ell_k), (T, \ell_t)>$, where S denotes the source node, T denotes the target node and ℓ_i denotes the label assigned to node v_i in P, selected from among all possible context labels suggested for node v_i.

To this end, a 'context-label network' is constructed for P, as described in the Section 2, in the form of a directed acyclic grid graph G' (see Fig. 3 for an illustration), where the j-th column of vertices in G' (presented in the figure under the corresponding j-th node of P) represents all the potential context labels for that node.

We define two edge types within the possible context transitions. A *Switch* edge represents a transition from label x to label y in an adjacent column in G'; it is added to the graph if the context-transition score of the ordered pair of labels it connects is above a given threshold, and its weight is set to $M(x, y)$. A *Skip* edge (shown in dashed lines in the figure), skips over an adjacent column in the grid to a label in the next one. A *Skip* edge connects two similar labels: x in column i of G' and y in column $i+2$ of G', if column $i+1$ does not contain any label z such that $M(x, z) > Threshold$. The weight of the *Skip* edge is set to $M(x, y) +$ SkipPenalty. When reconstructing an optimal solution interpreting P (i.e. the optimal context-label assignments to the nodes of P), the skipped node is assigned the same label as the one chosen for the source node of the skip edge (for example, in Figure 3C, node v is green even though it does not have a green label in its column). The Skip edge allows us to deal with poorly annotated genes and to suggest an overall context-acceptable path interpretation.

We can now reduce the problem of label assignment to that of finding the heaviest path in an edge-weighted–directed acyclic graph, with a predefined constraint d_{max} on the maximum number of *Skip* edges allowed. Dynamic programming is then applied to solve the reduced problem, implementing the recursion later in the text (Fig. 7), where x and y denote label nodes in G', *Predecessors*(y) denotes the set of vertices that have edges leading to y in G', and *Labels*(v) denotes the set of labels of v. The final context-labelling score is reported as $max_{y \in Labels(T)} S(y, d_{max})$.

The dynamic programming algorithm traverses all the nodes in the context-label network (G') in increasing column order and applies the recursion given in Figure 7 to compute the score for each label node. Therefore, the time and space requirements of the dynamic programming algorithm implementing this recursion is $O(E + V)$, where E and V denote the number of edges and vertices in G', respectively.

5.4 Evaluating the context sensitivity of known cellular paths in SPIKE

We conducted 10 trials as follows. In each trial, we computed a context-transition scoring matrix as described earlier in the text by using 19 of the 21 SPIKE maps. We then used the matrix to

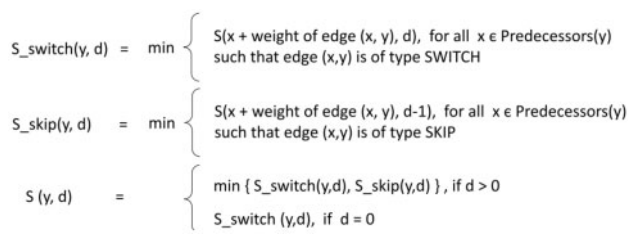


Fig. 7. The recursion for computing optimal label assignment

calculate the score of paths extracted from the two left-out maps. To test the significance of the score of each path, we shuffled the label assignments between nodes in the human interactome and re-scored the path. We repeated the randomization 40 times per path. Paths whose original score was better than the score obtained in at least 38 of the 40 shuffles ($P \leq 0.05$) were considered statistically significant and were marked as successful paths.

5.5 Inferring influenza infection pathways

We extracted from (Shapira *et al*., 2009) the following data sets: (i) the 36 human proteins that interact with influenza proteins and that are annotated with a GO term that appeared in M, which we denoted as sources; (ii) the 527 infection-related differentially expressed genes (as defined in Shapira *et al*.) that are connected to at least one of the source proteins, which we denote as targets; and (iii) 1756 genes that were silenced and their effect on infection was measured, which we denote as siRNA positive. Using our framework, we identified and scored all paths of lengths three to five edges in the interactome that connect these sources to these targets. We did not consider paths of length two because only 40% of them were context-sensitive according to the SPIKE analysis (Fig. 4). We used a context-sensitive scoring matrix M that we computed based on all SPIKE maps. Annotations based on IEA evidence codes, which are less reliable, were considered to increase the number of annotated genes in the interactome; however, they were weighted 50% lower than annotations based on other evidence codes. We then counted the fraction of paths per length that contained a predicted intermediate protein (other than the source and target) that was found to be siRNA positive.

5.6 PB2 analysis

We computed all paths connecting the 28 annotated human proteins that interact with PB2 (source proteins) and to the 527 differentially expressed genes that are on a path of length <6 edges from a source protein (target genes). The annotation of 39 of the 527 differentially expressed genes was associated with interferon.

ACKNOWLEDGEMENT

The authors thank the ISMB anonymous referees for their helpful comments.

Funding: European Union Seventh Programme under the FP7-PEOPLE-MCA-IRG Funding scheme (256360 to E.Y.-L.); United States-Israel Binational Science Foundation (BSF) (2009323 and 2011296 to E.Y.-L.); Israel Science Foundation (ISF) (478/10 to M.Z.-U.); Frankel Center for Computer Science at Ben Gurion University of the Negev (to A.L. and M.Z.-U.).

Conflict of Interest: none declared.

REFERENCES

Aranda,B. *et al.* (2010) The intact molecular interaction database in 2010. *Nucleic Acids Res*., **38**, D525–D531.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25.

Banks,E. *et al.* (2008) Organization of physical interactomes as uncovered by network schemas. *PLoS Comput. Biol*., **4**, e1000203.

Barabasi,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet*., **12**, 56–68.

Barrios-Rodiles,M. *et al.* (2005) High-throughput mapping of a dynamic signaling network in mammalian cells. *Science*, **307**, 1621–1625.

Barshir,R. *et al.* (2013) The tissuenet database of human tissue protein-protein interactions. *Nucleic Acids Res*., **41**, D841–D844.

Bebek,G. and Yang,J. (2007) Pathfinder: mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics*, **8**, 335.

Bisson,N. *et al.* (2011) Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the grb2 adaptor. *Nat. Biotechnol*., **29**, 653–658.

Cakmak,A. and Ozsoyoglu,G. (2007) Mining biological networks for unknown pathways. *Bioinformatics*, **23**, 2775–2783.

Ceol,A. *et al.* (2010) Mint, the molecular interaction database: 2009 update. *Nucleic Acids Res*., **38**, D532–D539.

de Lichtenberg,U. *et al.* (2005) Dynamic complex formation during the yeast cell cycle. *Science*, **307**, 724–727.

Graef,K.M. *et al.* (2010) The pb2 subunit of the influenza virus RNA polymerase affects virulence by interacting with the mitochondrial antiviral signaling protein and inhibiting expression of beta interferon. *J. Virol*., **84**, 8433–8445.

Guan,Y. *et al.* (2012) Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS Comput. Biol*., **8**, e1002694.

Ideker,T. and Sharan,R. (2008) Protein networks in disease. *Genome Res*., **18**, 644–652.

Magger,O. *et al.* (2012) Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol*., **8**, e1002690.

Matys,V. *et al.* (2006) Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*., **34**, D108–D110.

Myers,C.L. *et al.* (2005) Discovery of biological networks from diverse functional genomic data. *Genome Biol*., **6**, R114.

Pandey,J. *et al.* (2007) Functional annotation of regulatory pathways. *Bioinformatics*, **23**, i377–i386.

Paz,A. *et al.* (2011) Spike: a database of highly curated human signaling pathways. *Nucleic Acids Res*., **39**, D793–D799.

Pesquita,C. *et al.* (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol*., **5**, e1000443.

Salwinski,L. *et al.* (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res*., **32**, D449–D451.

Schadt,E.E. (2009) Molecular networks as sensors and drivers of common human diseases. *Nature*, **461**, 218–223.

Schaefer,M.H. (2012) Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Comput. Biol*., **9**, e1002860.

Scott,J. *et al.* (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. *J. Comput. Biol*., **13**, 133–144.

Shapira,S.D. *et al.* (2009) A physical and regulatory map of host-influenza interactions reveals pathways in h1n1 infection. *Cell*, **139**, 1255–1267.

Stark,C. *et al.* (2011) The biogrid interaction database: 2011 update. *Nucleic Acids Res*., **39**, D698–D704.

Steffen,M. *et al.* (2002) Automated modelling of signal transduction networks. *BMC Bioinformatics*, **3**, 34.

Suthram,S. *et al.* (2008) eqed: an efficient method for interpreting eqtl associations using protein networks. *Mol. Syst. Biol*., **4**, 162.

Szklarczyk,D. *et al.* (2011) The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*., **39**, D561–D568.

Tuncbag,N. *et al.* (2012) Steinernet: a web server for integrating 'omic' data to discover hidden components of response pathways. *Nucleic Acids Res*., **40**, W505–W509.

Yeang,C.H. *et al.* (2004) Physical network models. *J. Comput. Biol*., **11**, 243–262.

Yeger-Lotem,E. *et al.* (2009) Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet*., **41**, 316–323.

Yosef,N. *et al.* (2009) Toward accurate reconstruction of functional protein networks. *Mol. Syst. Biol*., **5**, 248.