BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 12 2005, pages 2898–2905
doi:10.1093/bioinformatics/bti440

Systems biology

Differential network expression during drug and stress response

Lawrence Cabusora[1], Electra Sutton[1], Andy Fulmer[2] and Christian V. Forst[1,*]

[1]Los Alamos National Laboratory, PO Box 1663, Mailstop M888, Los Alamos, NM 87545, USA and [2]Miami Valley Labs, Procter & Gamble, PO Box 538707, Cincinnati, OH 45253-8707, USA

Received on September 8, 2004; revised on March 28, 2005; accepted on April 6, 2005
Advance Access publication April 19, 2005

ABSTRACT

Motivation: The application of microarray chip technology has led to an explosion of data concerning the expression levels of the genes in an organism under a plethora of conditions. One of the major challenges of systems biology today is to devise generally applicable methods of interpreting this data in a way that will shed light on the complex relationships between multiple genes and their products. The importance of such information is clear, not only as an aid to areas of research like drug design, but also as a contribution to our understanding of the mechanisms behind an organism's ability to react to its environment.

Results: We detail one computational approach for using gene expression data to identify response networks in an organism. The method is based on the construction of biological networks given different sets of interaction information and the reduction of the said networks to important response sub-networks via the integration of the gene expression data. As an application, the expression data of known stress responders and DNA repair genes in *Mycobacterium tuberculosis* is used to construct a generic stress response sub-network. This is compared to similar networks constructed from data obtained from subjecting *M.tuberculosis* to various drugs; we are thus able to distinguish between generic stress response and specific drug response. We anticipate that this approach will be able to accelerate target identification and drug development for tuberculosis in the future.

Contact: chris@lanl.gov

Supplementary information: Supplementary Figures 1 through 6 on drug response networks and differential network analyses on cerulenin, chlorpromazine, ethionamide, ofloxacin, thiolactomycin and triclosan. Supplementary Tables 1 to 3 on predicted protein interactions. http://www.santafe.edu/~chris/DifferentialNW

1 INTRODUCTION

One of the most important challenges for researchers in the *post-genomic era* is to move toward a new view of biology—a systems level approach. The surprisingly low estimate of 30 000–40 000 genes in the human genome strongly indicates that functional complexity may originate in locations and processes beyond the identification of particular genes. The highly successful approach of biology for the past thirty years has been to investigate individual genes or proteins, but *Systems Biology* moves beyond looking at such elements in isolation. Rather, it examines the behavior and relationships of all of the elements in a biological system in an attempt to model and, ultimately, to control its dynamical behavior.

One particular goal within systems biology is to develop the capability for analyzing biological interaction networks as they respond to different external conditions. Groundwork with respect to this goal has been laid in seminal contributions from Ideker *et al*. (2001, 2002) and Zien *et al*. (2000). This capability is of great importance for the understanding of system behavior and finally the prevention of infection from bacterial pathogens, i.e., *Mycobacterium tuberculosis* and others.

The paper will present a new method for analyzing biological response data, such as those from gene expression arrays, by combining it with computationally derived network information. It will further provide predictions of particular response sub-networks in the *M.tuberculosis* network after treatment with unspecific stress-inducers and comparison with specific antibacterial drugs. The impact of these studies will be the identification of response networks that could present potential drug-targets for novel antibiotics.

2 SYSTEMS AND METHODS

The method and analysis section is comprised of various steps, including, (i) *M.tuberculosis* network construction by computational methods, (ii) network filtering and response network identification by superimposing experimental gene expression data upon the computationally derived *M.tuberculosis* network and (iii) differential network expression analysis.

2.1 Graphs and networks

In the following we describe basic graph-theoretical properties and different network representations.

DEFINITION 1. *A 'typical' graph* $\Gamma = (\mathcal{V},\mathcal{E}) = (\mathcal{V}(\Gamma),\mathcal{E}(\Gamma))$ *consists of a vertex set* $\mathcal{V}$ *with vertices (or nodes)* $v \in \mathcal{V}$ *and an edge set* $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ *with edges* $\epsilon \in \mathcal{E}$.

Populating a graph Γ with biological information yields a biological network N as follows.

DEFINITION 2. *Let* $N = (\mathcal{V},\mathcal{E},\pi)$ *be a network with vertices* $v \in \mathcal{V}$ *and edges* $\epsilon \in \mathcal{E}$ *as well as a function* $\pi : X \to \mathcal{P}$ $(X = \mathcal{V} \cup \mathcal{E})$ *that maps vertices and edges onto respective properties* $p \in \mathcal{P}$.

In the case of biological networks, depending on the particular network representation, node properties can include gene, protein or chemical names, and edge properties may refer to specific interactions, such as *binding* or *catalysis*.

The mapping $\pi : X \to \mathcal{P}$ is at least surjective because for all $p \in \mathcal{P}$, there exists an $x \in X$ with $\pi(x) = p$.

*To whom correspondence should be addressed.

DEFINITION 3. *We define a generalized reaction graph (GRG), as a triple* $\mathcal{G} = (P, T, R)$*, where* P *is the set of places,* T *denotes the set of transitions and* R *indicates a set of reactions (directed edges) utilizing places that transform places into other places via a transition (thus,* $R \subseteq (P \times T) \cup (T \times P)$*). Vertices of* $\mathcal{G}$ *are all places or transitions.*

This definition is a simplification of the definition of *Petri-Nets* (Reisig, 1985), where transitions are actually represented by abstract nodes. In our definition, we consider transitions to be associated with components and thus simplify the strict Petri-Net definition.

Analogous to the network defined in Definition 2, we expand the GRG and define a *generalized reaction network* (GRN) $\mathcal{N}$ by introducing a property mapping $\pi : (P \cup T) \rightarrow \mathcal{P}$, yielding $\mathcal{N} = (P, T, R, \pi)$.

A GRG can easily be simplified into a *node graph* where places are translated into nodes, and transitions are contracted to edges, as well as into an *edge graph*, where transitions are converted to nodes, and places are contracted to edges.

2.2 Biological network construction

In order to construct the *M.tuberculosis* network, we identified three type of interactions relevant for such a network (and disregarded other potential sources of interaction information, such as protein–DNA binding or multi-state protein phosphorylation by kinases during signaling, due to insufficient information): (i) protein interaction, (ii) metabolic reactions and (iii) co-expression in regulons. For protein interaction data, we used identified component genes involved in fusion events according to Enright *et al*. (1999); Enright and Ouzounis (2001). This method for recognizing possible interacting proteins is called *the Rosetta Stone* approach (Enright and Ouzounis, 2001; Marcotte, 2000); it is based on the observation that individual genes in one organism that are fused into a single chain in another organism are likely to interact physically with each other. Employing the Rosetta Stone method using about 80 completely sequenced and published microbial genomes, we observed 113 interacting proteins. Between these 113 proteins we identified 257 interactions. In the graph, the proteins are represented as vertices, and edges connect interacting proteins. Each edge is assigned a Z-score z which is a statistical measure of similarity for the pair of components at its endpoints (Enright *et al*., 1999; Enright and Ouzounis, 2001). We used a conservative cut-off value of $|z| \leq 2$ yielding 106 proteins and 233 interactions that we used for the construction of the *M.tuberculosis* protein interaction network (Supplementary Tables 1 and 2). The remaining protein pairs with Z-scores above the threshold involving 24 interactions were disregarded (Supplementary Table 3).

We combined the protein interaction network with the metabolic reaction data of *M.tuberculosis* from the BioCyc (http://biocyc.org) and KEGG (Ogata *et al*., 1999) databases. The enzymes and substrates in the metabolic reactions are denoted as places and transitions. Our goal was to use metabolic reactions that would assemble into textbook-style metabolic pathways, without additional links introduced through ubiquitous chemicals. Since co-factors such as ATP and substrates such as water are ubiquitous in the metabolic reactions, we deleted the substrates and edges associated with those nodes; specifically, substrates with the number of interactions higher than 100 were removed from the network. A list with a number of ubiquitous substrates is shown in Table 1. The reaction graph was then constructed by contracting the substrate edges, i.e., by connecting enzymes that shared at least one common substrate by an edge and then deleting all the substrate vertices from the graph. The regulon (or coordinately regulated operon) data of *M.tuberculosis* were obtained from *Escherichia coli* by homology (McGuire and Church, 2000). We added the proteins as vertices to the graph and connected any co-regulated genes with new edges.

2.3 Experimental data

Through a separate study by Boshoff *et al*. (2004), gene expression information from *M.tuberculosis* (H37Rv) was obtained after its growth in Middlebrook 7H9 supplemented with albumin/dextrose/NaCl/glycerol, Dubos medium or defined minimal medium, as previously described (Schnappinger *et al*., 2003). Cultures were grown before adding either drug or solvent, and RNA was isolated at selected intervals thereafter. For each drug-treated culture, a parallel culture was treated with an equivalent amount of solvent [DMSO (dimethylsulphoxide), ethanol or water] for the same amount of time. RNA from the latter culture was used as the reference sample to which the drug-treated sample was compared. Each treatment condition and each drug concentration was repeated a minimum of two independent times.

The gene expression data has been made available by Helena Boshoff and is accessible through the Gene Expression Omnibus at NCBI (GEO; http://www.ncbi.nlm.nih.gov/geo) with GEO platform accession number GPL1396.

Table 1. Ubiquitous chemical substrates

Chemical	Number of interactions
Water	22 781
ATP	10 539
H^+	4 872
Phosphate	2 446
CO_2	2 295
NADPH	2 040
NADH	1 800
Coenzyme A (CoA)	1 400
Donor-H_2	893
2-Oxo-glutarate	884
Glutamate	735
Adenosyl-homo-Cystein	455
Pyruvate	287
NAD	217
Acetate	154
O_2	121
Fumarate	119
NH_3	119
Succinyl-CoA	117
FAD	110

3 ALGORITHM AND IMPLEMENTATION

3.1 Overview

The algorithm is implemented as follows. After the large biological network (the *M.tuberculosis* network, Section 2.2) is constructed, it is stored as a Petri-Net, converted into a node graph (Section 2.1), stripped of ubiquitous chemicals according to a maximal permissible node degree (Section 2.2) and loaded with gene expression values (Sections 2.3 and 3.3). In addition, input parameters such as the k-value for k-shortest path calculation and the maximal path-length l are set. In the next step, the network is filtered according to the parameters and a set of seed nodes (Section 3.2). The core of the algorithm is the component that scores particular sub-networks, yielding a sub-network with optimal score. Further components of the algorithm include a statistical analysis routine for sub-network score validation (Section 3.4) and a network algebra component for differential network analysis (Section 3.5). A flow diagram of the corresponding processes is shown in Figure 1.

REMARK 1. *In the course of this paper we use the node graph to represent and to describe the biological network* N*. However, all algorithms are sufficiently general to be applied to GRNs. In the case*

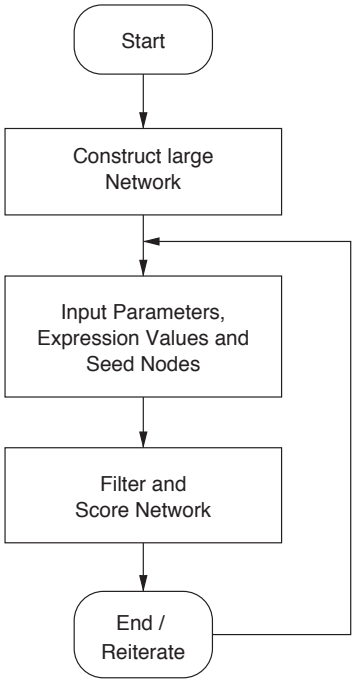

Fig. 1. Global architecture of the algorithm.

of different implementations for different network representations, references are being made.

3.2 Sub-network filtering

A first-pass analysis in the iterative process of *Network Scoring* is performed during *sub-network filtering*. Here, particular nodes are tagged as *seed nodes*, and shortest pathways between them are identified by Dijkstra's algorithm (Dijkstra, 1959). Successively longer pathways are identified by Yen's algorithm for the k-shortest simple paths problem, using an implementation by Hershberger *et al.* (2003). Yen's algorithm, despite being over thirty years old, is still the best known for the k-shortest simple paths problem with respect to its worst-case running time, i.e., $O(kn(m + n \log n))$ time for a graph with m nodes and n edges. For the purposes of sub-network filtering, an additional constraint is built into the pathfinding algorithm: if the length of a (weighted) path exceeds a particular, pre-specified length l, then the path is disregarded and not included in the sub-network.

Seed nodes are either provided by the user or are extracted by machine learning techniques such as genetic algorithms or singular value decomposition to cluster expression data. With a given set of seed nodes, which may change during the iterative process between sub-network filtering and network scoring (Section 3.3), sub-networks are identified that are spanned by the seed nodes. Thus, if $v_s \in N$ is a seed node in network N, then $v_s \in \partial N$, with ∂N denoting the boundary of N.

The algorithm performs the following functions: (i) compilation of a list of seed nodes, (ii) computation of all possible pairs of seed nodes from this list, (iii) calculation of shortest and k-shortest paths (with maximal path-length l) between each pair of seed nodes using Dijkstra's and Yen's algorithms, (iv) recording of all nodes and edges on identified paths, (v) filtering (deletion, hiding) of all other nodes and edges that are not on the selected paths, (vi) sub-network scoring, (vii) reiteration.

By mathematical means, the above procedure extracts a sub-network from the large biological network that is spanned by the seed nodes. The pathways between each pair of seed nodes hold the desired properties of being k-shortest paths with maximal path length l. Thus, by scoring a particular network with given seed nodes, the last step of reiteration is not necessary. Only for the identification of the optimally scored network and for the refinement of seed nodes is reiteration required (see Sections 3.3 and 3.4).

3.3 Network scoring

Network scoring uses expression values as metrics for weighted edges in the network. In a node graph, as we are using, genes, proteins and other cellular components are coded as nodes which are connected by edges in the biological network. Because the sub-network filtering step assumes weights on edges for scoring, such edge weights must be calculated from node scores, i.e., gene expression levels.

REMARK 2. *On the other hand, if the biological network is represented as an edge graph, then edges represent genes, proteins and other cellular components, and nodes refer to interactions. Then edges can be weighted by gene expression values directly.*

We use the *Bioconductor* (http://www.bioconductor.org) modules in combination with the *R* package to analyze the gene expression data and to extract p-values p_m for each expressed gene m. Following Ideker *et al.* (2002), we convert the p_m into a z-score $z_m = \Phi^{-1}(1 - p_m)$, where Φ^{-1} denotes the inverse normal distribution function.

REMARK 3. *For the edge scores, we use either the product probability $p_i \cdot p_j$ or the correlation coefficient ρ_{ij} for an edge m with origin node i and terminus j, yielding for example $z_m = \Phi^{-1}(1 - p_i \cdot p_j)$.*

To calculate a total score of the sub-network N, we then sum the z_m over all m, given the constraint of the k-shortest paths with maximal path-length l between each two seed nodes in N:

$$z_N = \frac{1}{\sqrt{m}} \sum_{\text{shortest } N(k,l)} z_m . \qquad (1)$$

Given a particular set of seed nodes, the shortest path approach already guarantees a best scored sub-network, but to obtain an optimized set of seed nodes for a better scoring sub-network, we have to search through the network. We reduce this problem by finding the best scoring pathways between pairs of randomly selected seed nodes out of a list ordered by, for example, fold change after gene expression analysis. We choose pairs of nodes by a rank-weighted random distribution. The shortest pathway between these nodes is identified, s_N [Equation (2)] calculated and the node pair recorded. After convergence to highest scoring pathways, a sub-network score is computed using nodes with the highest pathway score.

Although the alternating calculation of node pairs yielding best scoring pathways and sub-network filtering is a heuristic approach, it is advantageous to the network optimization method due to its speed. Particular combinations of node pairs may assemble into a suboptimal sub-network [e.g. node a_1 in node pair (a_1, a_2) is distant from node b_1 in pair (b_1, b_2)]. However, such a suboptimal seed node combination will be detected in the next sub-network filtering by the path-length restrictions. Long and poorly scored pathways will disrupt the network in unconnected components. Only seed nodes

within the giant component will be included in the next round of pathway optimization.

3.4 Statistical analysis

We perform a statistical analysis to validate the significance of the identified sub-graph compared to other sub-networks. Through a Monte Carlo approach, we compare z_N for a particular sub-network to a set of randomly sampled reference sub-networks of size m (using the same expression profile but different seed nodes). The sub-network scores are corrected in the standard fashion:

$$s_N = \frac{z_N - \langle z_m \rangle}{\sigma_m} \tag{2}$$

This correction guarantees a mean of 0 and a standard deviation of 1 for the scores of randomized sub-networks.

3.5 Differential network expression analysis

In principle, graph comparison is an NP-hard problem which typically can only be addressed by exhaustive enumeration techniques. On the other hand, methods for comparative network analysis for biological systems have been developed in the past. Such methods have been proven powerful in a number of applications including for metabolic (Dandekar *et al*., 1999; Forst and Schulten, 1999, 2001; Ogata *et al*., 2000) and protein interaction networks (Kelley *et al*., 2003) as well as for correlation of protein interaction networks with gene expression (Nakaya *et al*., 2001).

Here we present two approaches: a simple approach utilizing node and edge labels of the given biological network, i.e., using only matching components for the graph comparison; and a correlation-based approach taking into account the commonality of how sub-networks have been generated by our method. With respect to the simple approach we can define the following network algebra terms.

DEFINITION 4. *We consider two networks N_1 and N_2. The intersection $I(N_1, N_2)$ is defined as follows:*

$$I(N_1, N_2) = \{v, \epsilon : \pi(v) = \pi(v') \wedge [\pi(\epsilon) = \pi(\epsilon') \text{ if } \pi(o(\epsilon)) = \pi(o(\epsilon')) \wedge \pi(\tau(\epsilon)) = \pi(\tau(\epsilon'))]\}$$

with $o(\epsilon)$ and $\tau(\epsilon)$ being the origin and terminal nodes of edge ϵ.

In plain words, we take the intersection between all $v \in \mathcal{V}$ and $v' \in \mathcal{V}'$ as well as the intersection between corresponding edges $\epsilon \in \mathcal{E}$ and $\epsilon' \in \mathcal{E}'$ under the condition that $\forall v, v' : \pi(v) = \pi(v')$ and $\forall \epsilon, \epsilon' : \pi(\epsilon) = \pi(\epsilon')$. An edge $\epsilon \in N$ is only chosen if both the originating and terminating nodes have π-corresponding nodes in N'. Other network algebraic terms, such as symmetric difference or union, are defined similarly.

In addition to the general network algebraic operations, we have also implemented a correlation-based graph comparison technique that takes advantage of the common method by which sub-networks are generated. Specifically, we are given a set of networks $N_1, N_2, \ldots, N_n$, each of which was generated from a large biological network by sub-network filtering over different gene expression profiles. We create a 'correlated intersection' I_c [with vertex set $\mathcal{V}(I_c)$ and edge set $\mathcal{E}(I_c)$] as follows. We initially set $\mathcal{V}(I_c) = \cup_{i \in [1,n]} \mathcal{V}(N_i)$, where π-corresponding nodes from different networks are identified into a single node in $\mathcal{V}(I_c)$. Then we let $\mathcal{E}(I_c)$ be such that if $\epsilon \in \mathcal{E}(N_i)$ has endpoints v_1 and v_2, then there is an $\epsilon' \in \mathcal{E}(I_c)$ with endpoints in $\mathcal{V}(I_c)$ corresponding to v_1 and v_2. [There is no need for duplicate edges between any pair of nodes in $\mathcal{E}(I_c)$.] We then iterate over the $\epsilon' \in \mathcal{E}(I_c)$: ϵ' has endpoints v'_1 and v'_2 which correspond to genes. Each network N_i was generated under a certain set of conditions that gave specific expression values to these genes. Thus, for each N_i, we have expression values v_{1i} and v_{2i} for the genes in question. We take the Pearson correlation coefficient ρ between the two sets of variables v_{1i} and v_{2i}. If ρ is less than a given threshold value, then we remove edge ϵ' from I_c; otherwise, it remains in the final correlated graph. (This can also be done with other statistical correlation coefficients, e.g., the Kendall τ or the Spearman ρ.) The procedure can be used with networks generated from a similar set of conditions, e.g., multiple trials using the same experimental conditions, in which case the procedure yields a network which has been 'smoothed out' for experimental variation. The procedure can also be given networks generated from markedly different conditions, in which case it yields a network roughly analogous to a refined intersection which takes more of the data into account than the simple on–off presence of a node/edge in the networks.

4 RESULTS

First line drugs used for therapy against tuberculosis infection include ethambutol, pyrazinamide and isoniazid (INH). Isoniazid is known to inhibit the biosynthesis of mycolic acid, disrupting the FAS-II fatty acid synthesis pathway that *M.tuberculosis* utilizes to construct lipids used in its cell wall (Wilson *et al*., 1999). We examined the drug response network of *M.tuberculosis* induced by the chemical INH and compared it to a generic stress response network induced by exposure to hydrogen-peroxide (H_2O_2). We also examined other drugs targeting fatty acid biosynthesis in the bacterial pathogen, as well as certain drugs with completely different mechanisms.

Isoniazid specifically interferes with the FAS-II subsystem of the cell that extends fatty acids from 26 to ∼56 carbon atoms for the biosynthesis of mycolic acid. Isoniazid itself is a pro-drug: it is modified to its biologically active form inside the pathogen by the mycobacterial catalase-peroxidase enzyme *katG*. In the presence of *katG* or Mn^{2+} ions, INH forms a complex with NAD, an INH–NAD adduct. This adduct has been shown to act as a slow, tight-binding competitive inhibitor of *InhA*, an enoyl reductase that catalyzes the NADH-dependent reduction of long chain *trans*-2-enoyl-acyl carrier proteins (ACPs). Experimental results by Rawat *et al*. (2003) indicate a mechanism of at least two steps in which an initial enzyme-inhibitor complex is rapidly formed and then slowly changed to a final inhibited complex. However, a second target of INH has been proposed by Barry and co-workers (1998). *KasA*, one of three ketoacyl synthases in the FAS-II pathway, has been identified to be inhibited by the activated INH complex.

Although INH has been the most effective and widely used drug for the treatment of tuberculosis since the 1950s, it requires activation by *katG*. Thus, a substantial fraction of all *M.tuberculosis* isolates that are resistant to INH show *katG* mutations. Consequently, compounds that inhibit the ultimate molecular target(s) of INH but do not require activation by *katG* have tremendous promise as novel drugs for combating multi-drug resistant *M.tuberculosis*.

To understand the interlocked mechanism of INH activation and *M.tuberculosis*' drug response, we studied the *M.tuberculosis* response against INH treatment and a generic stress response against H_2O_2. Gene expression profiles were obtained by our collaborator Helena Boshoff. The INH expression profile was obtained

by exposing *M.tuberculosis* to 0.2 μg/ml INH for 6 h against a reference of *M.tuberculosis* exposure to ethanol. The H_2O_2 measurements were performed in a concentration of 4 mM for 2 h. In addition to these treatments, we utilized expression profiles from experiments involving other *M.tuberculosis* drugs, such as cerulenin, chlorpromazine, ethionamide, thiolactomycin and triclosan. The following experimental conditions for exposing *M.tuberculosis* to drugs and solvents were used: 0.5 μg/ml cerulenin for 6 h against DMSO, 10 μg/ml chlorpromazine for 6 h against ethanol, 12 μg/ml ethionamide for 6 h against ethanol, 5 μg/ml ofloxacin for 6 h against ethanol, 130 μg/ml thiolactomycin for 6 h against DMSO and 50 μg/ml triclosan for 6 h against ethanol. All experiments were performed in duplicate with no or minimal differences between corresponding response networks.[1]

The networks generated from this data are shown in Figures 2 and 3; the seed nodes used were the union of the most significant response nodes to INH and H_2O_2 exposure individually, as determined by gene expression analysis. We note some interesting features of the graphs. Genes known to encode for proteins involved in the classical DNA damage response (*recA*, *uvrA*) appeared in the calculated H_2O_2 response network. A glance at the graph shows that each was up-regulated (Fig. 2), as well as *furA–katG*, a gene pair central to the oxidative stress response of *M.tuberculosis*. In the same network motif as *recA* and *uvrA* there resides a conserved hypothetical gene *Rv*2840*c* that is strongly up-regulated during H_2O_2 treatment. The network connectivity of *Rv*2840*c* with *recA* and *uvrA* suggests an involvement with the DNA-repair network of *M.tuberculosis*. This also applies to *infB*, the translation initiation factor IF-2 that may also play a role in DNA-repair. Genes that are strongly up-regulated but that are not members of an up-regulated network motif are *Rv*1464, a member of the *NifS*-family and the hypothetical protein *Rv*1831.

On the other hand, the entirety of the FAS-II fatty acid synthase pathway (except *acpM*, which was not included in the interaction data used to construct the original, whole network) appears in the INH response network (Fig. 3); these include enoyl-[acyl-carrier-protein] reductase *inhA*, 3-oxoacyl-[acyl-carrier-protein] synthase *kasA*/*B*, and mycolytransferase *fbpC*2. All except *inhA* show strong up-regulation. Some of these nodes were designated as seed nodes, and such nodes were also observed in the H_2O_2 response network; nevertheless, it is clear that none of them was under a particular up-regulation there. Furthermore, the specific removal of these genes (*kasA*, *kasB*, *fabD*) from the seed node list did not affect their presence in the INH response sub-network: the newly calculated network continued to contain each of them (data not shown). Furthermore, a similarly recalculated H_2O_2 response network lacked all of these genes. Also involved in the up-regulated *kasA*/*B* network motif are the strongly up-regulated putative 1-acylglycerol-3-phosphate *O*-acyltransferase *Rv*2182*c* and phenolpthiocerol synthase *ppsA*. 1-Acylglycerol-3-phosphate *O*-acyltransferase does play a role in lipid bio-synthesis. On the other hand, *ppsA* is known to be involved in the synthesis of phenolglycolipids, a class of components in the *M.tuberculosis* membrane that are known to play a role in pathogenicity (Constant *et al*., 2002).

Interestingly, the up-regulated FAS-II pathway is closely connected to a similarly up-regulated fatty acid degradation pathway assembled by the enoyl-hydratases *echA*3, *echA*6 and *echA*8; by acyl-CoA ligases *fadD*5, *fadD*7 and *fadD*8; as well as by the putative acyl-CoA dehydrogenase *fadE*24. Other up-regulated pathways include a proposed electron transport/nitrate utilization pathway involving cytochrome oxidase *cydB*, NADH dehydrogenase *ndh* and nitrate reductases *narG* and *narJ*, as well as the putative oxidoreductase *Rv*0197, the last putative. Alkylhydroperoxidases *ahpC* and *ahpD* are up-regulated in both the H_2O_2 and the INH response networks.

The picture that the INH response sub-network draws confirms the response of *M.tuberculosis* after INH treatment. The interference of INH within *M.tuberculosis* triggers a response both in the FAS-II pathway as well as in the fatty acid degradation pathway. The latter pathway is activated for the degradation of cell-membrane lipids made premature due to the interruption of their biosynthesis by the inactivation of *inhA*. Up-regulation of *ahpC* and *ahpD* for detoxification induces down-regulation of *katG*. The up-regulation of *cydD*, *ndh* and *Rv*0197 suggests degradation of INH through nitrate and nitrite to ammonia.

The differential response network (Fig. 4) shows the common responses between the generic DNA-damage/stress response introduced by H_2O_2 and the specific drug response against INH. Up-regulated in both networks are alkylhydroperoxidases *ahpC* and *ahpD*. These genes are involved in detoxification and are suspected to play a role in INH resistance by interfering with the *katG* activity. The latter gene is a member in both networks, together with *furA*, as part of the DNA-damage pathway. *FurA* is up-regulated in the H_2O_2 network and slightly up-regulated in the INH network. The down-regulation of *katG* in the INH network seems to conform with the interference with *ahpC*/*D*. *KatG* is up-regulated during H_2O_2 response.

Other commonly up-regulated genes include the beta chain of the tryptophan synthase, *trpB*, the phosphoenolpyruvate carboxykinase *pckA* and *regX*3, a two-component response regulator involved in *M.tuberculosis* virulence and infection (Parish *et al*., 2003). The latter, together with the alternative sigma factor *sigH*, is a member of a network motif that regulates major components of the oxidative and heat stress response, as well as a hypothetical protein, *Rv*1222. A second commonly up-regulated network motif involves a member of the DNA-repair network, excinuclease ABC subunit A, *uvrA* and a conserved hypothetical protein, *Rv*2840*c*.

Computing the network intersection between the H_2O_2 and INH response networks yields a sub-network that assembles about 60% of the nodes and a third of the edges of the individual INH and H_2O_2 networks. Given the large component of commonly up-regulated genes including virulence factors and generic stress response network motifs, we conclude that INH does trigger essential key components that are common to the two networks.

Other drugs known to work against the FAS-II pathway in *M.tuberculosis* are ethionamide and thiolactomycin. Comparing response networks between INH and these drugs by means of our intersection method yielded networks that clearly reflected the common mode of action (Supplementary Figures 1 and 2). We note the presence of the FAS-II genes *kasA*, *kasB*, *fabD* and *fbpC*2. Examining the response network generated by cerulenin—a drug that inhibits both the FAS-I and FAS-II pathways—also shows a large number of FAS-II gene expressions being affected in the response of *M.tuberculosis* (Supplementary Figure 3).

[1]Data available upon request.

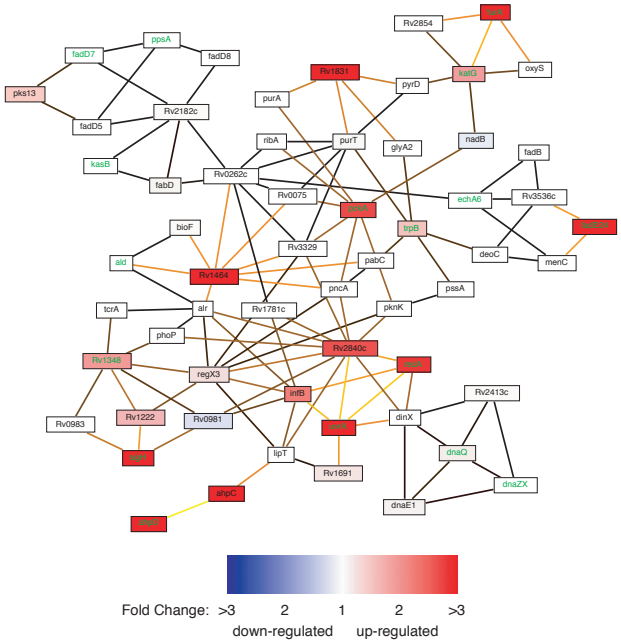

Fig. 2. (Top) Hydrogen-peroxide response sub-network. Red nodes denote up-regulated genes; blue nodes indicate down-regulated genes. Green node labels are seed nodes. Parameters are $k = 3$ and $l = 14$. The network consists of 60 nodes, 17 of which are seed nodes, and 117 edges. (Bottom) Specific fold change for a particular node is color-coded according to the color bar. For color coding of edges, see Figure 3.

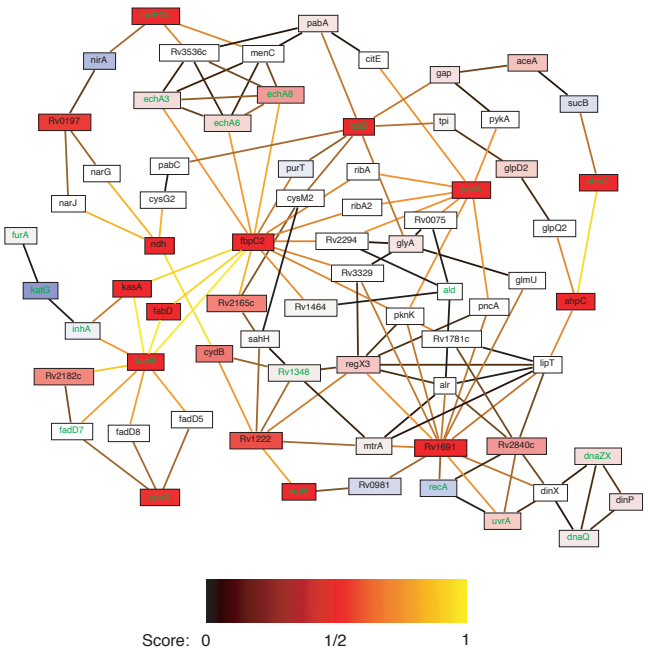

Fig. 3. (Top) Isoniazid response sub-network. Color-coding is similar to Figure 2. Parameters are $k = 2$ and $l = 13$. The network has 71 nodes and 131 edges. Eighteen nodes are seed nodes. (Bottom) Individual edge-scores used to calculate pathway scores are color-coded according to the color bar.

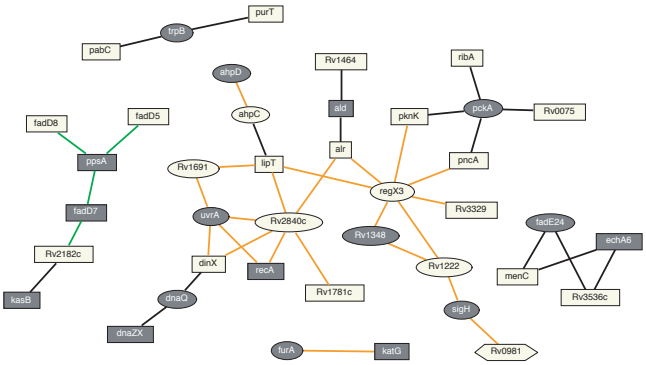

Fig. 4. Intersection between the INH and H_2O_2 response networks (Figs 2 and 3). Dark gray nodes are seed nodes, black edges denote bio-chemical reactions and orange edges indicate relationships through operons. Oval shaped nodes indicate common up-regulation and the hexagonal node denotes common down-regulation. The network consists of 40 nodes, 17 of which are seed nodes, and 42 edges.

Triclosan has been demonstrated to attack the FAS-II system *in vitro* too. However, Boshoff *et al*. suggest that triclosan's primary mode of action *in vivo* is actually upon respiration (Boshoff *et al*., 2004). Our results support this conclusion (Supplementary Figure 4): comparison of a triclosan response network to any of the networks generated by the aforementioned drugs that attack the FAS-II system gave sub-networks of rather small sizes (26 nodes instead of a typical 37 nodes on average) lacking the complete lipid biosynthesis pathway that is present in the isoniazid, ethionamide and thiolactomycin response networks. All drug response networks for drugs affecting the FAS-I or FAS-II pathways have in common two network hubs around mycolytransferase *fbpC*2 and the conserved hypothetical protein *Rv*1691. Together with *fbpA* and *fbpB*, *fbpC*2 forms the antigen 85 complex that plays a key role in the pathogenesis of *M.tuberculosis* (Kremer *et al*., 2002). In addition to its mycolytransferase activity, *fbpC*2 is also involved in the cell wall metabolism. The function of the second hub *Rv*1691 is unknown; a BLAST search did not show any hit to other genes with known function.

On the other hand, comparing the triclosan response with that generated by chlorpromazine yielded much more of a correlated intersection (Supplementary Figure 5). Chlorpromazine, among other actions, has been shown to inhibit respiration in *M.tuberculosis*, and we note in the correlated intersection the presence of genes dealing with respiration and oxidative stress: *cydB*, *appC* and *nirB*. This supports the claim that triclosan does indeed affect respiration in *M.tuberculosis*.

For an example of a drug that did trigger a DNA-repair response, we examined ofloxacin, one of the class of drugs known as fluoroquinolones. The precise mechanism by which any of the fluoroquinolones interfere with DNA synthesis has not yet been elucidated, but our approach enables us to discern immediately the nature of the action of ofloxacin. We compare the ofloxacin response network with the response of *M.tuberculosis* to ultraviolet light exposure (Supplementary Figure 6). The resulting correlated intersection clearly exhibits genes associated with the SOS gene repair response (*uvrA*, *uvrB*, *recA*, *infB*).

5 DISCUSSION

We have developed a novel method using *k*-shortest path algorithms to examine and to interpret expression data in the context of network connectivity. The method is flexible enough to accommodate new information about protein or gene interactions or even to incorporate completely novel connections. The algorithms used scale easily with the size of the network. Although the present *M.tuberculosis* network has been constructed by computational methods and information acquired from interaction and network databases, interaction information obtained from experimentally derived protein interactions or regulon information can easily be included in the network. In the simplest case, new interaction data from literature information can be added to the network by establishing novel links. To ameliorate existing interactions by experimental data, confidence values similar to *Z*-scores, as in the case of protein interaction, can be added. All nodes and edges are labeled in the network and have additional properties, such as *Z*-scores, attached.

Comparison of the drug response network induced by INH and the generic stress response network induced by H_2O_2 revealed features unique to each graph. In particular, we not only detected the presence of such genes as the *recA* and *uvr* family or the FAS-II group, but we were able to see how they interacted with other genes in their respective networks. We specifically identified a close connection between the FAS-II network and the fatty acid degradation pathway in the INH network as well as a putative electron-transfer, INH degradation network assembled by cytochromes, nitrate and nitrite reductases.

This approach to constructing response networks enables easy comparison of the generated graphs. This in turn allowed identification of the similarities between the responses of *M.tuberculosis* to the presence of H_2O_2 and INH. In particular, in the differential network we identified genes and network components that are relevant for stress response as well as responsible for INH activation, such as the *furA*/*katG* pair. Another class of common network responses involved detoxification and virulence networks including *ahpC*/*D*, *ppsA* and *regX*3.

The analysis of other *M.tuberculosis* drugs, including cerulenin, chlorpromazine, ethionamide, ofloxacin, thiolactomycin and triclosan, revealed insights into common response among the drugs targeting the FAS-pathways as well as differences from triclosan, chlorpromazine and ofloxacin, which have different targets. As expected, we observed a great deal of correspondence between the response networks of INH, ethionamide and thiolactomycin. Common to all these drugs that inhibit the FAS-pathway are the strongly up-regulated network motifs involving the highly connected hubs *fbpC*2 and *Rv*1691. We noted evidence of triclosan's *in vivo* action with observed activity against the FAS-II pathway but primarily targeting the citric acid (TCA) cycle and respiration. Network comparison with another *M.tuberculosis* drug, chlorpromazine, identified conserved and highly up-regulated network motifs involving genes of the TCA cycle and cytochromes. With respect to a representative of a third class of *M.tuberculosis* drugs, ofloxacin, we identified a common network response between this drug and UV radiation, indicating an interference with DNA synthesis by triggering DNA repair response.

The strength of differential network expression analysis lies in the identification of the modes of action of drugs by comparative studies. Similar response networks indicate similar modes of action, as has been shown here in the case of drugs targeting the FAS-pathways. By differential network expression, novel drugs with similar response networks to drugs with known modes of action can be easily identified early in the drug development to prevent repeated and costly clinical trials. On the other hand, response networks by themselves serve as a valuable representation and computational model of physiological responses. Identification of single highly connected genes, such as *Rv*1464, *Rv*1691, or *Rv*2840*c* or small and strongly expressed network motifs provides information about potential drug targets. Future experiments could specifically target these identified network motifs to either verify the importance of these particular genes and network motifs for *M.tuberculosis* or refute the prediction.

In summary, differential network expression provides an excellent systems biology tool to identify and analyze systems level responses by a comparative approach. In particular, the case study on the H_2O_2 and INH response of *M.tuberculosis* highlighted the interconnectivity of cellular processes. The presented method is sufficiently flexible to accommodate a variety of types of biological network information and experimental data. Initial differential network analysis studies have already been performed in the case of mammalian signaling networks and therapeutic drug responses. Differential network analysis not only offers insights into the mode of action of antimicrobial drugs but also provides information on potential key targets for future drug-development efforts.

ACKNOWLEDGEMENTS

This work was supported by grants from the NSF student training program to LC, from Los Alamos National Laboratory (LDRD-20040184ER), and from the Procter & Gamble Corp. by a Cooperative Research and Development Agreement (CRADA LA01C10461). We gratefully thank Helena Boshoff, NIAID, for providing *M.tuberculosis* gene expression data and Hyun Park, Carnegie Mellon, for coding the initial program version. We are also indebted to Ivo Hofacker, University of Vienna, Austria and Peter Stadler, University of Leipzig, Germany for essential contribution to the network algebra as well as to Jeff Blanchard, University of Massachusetts, and Robert Küffner, University of Munich, Germany, for numerous and fruitful discussions on pathway scoring.

REFERENCES

Boshoff,H.I.M. *et al*. (2004) The transcriptional responses of *M.tuberculosis* to inhibitors of metabolism: novel insights into drug mechanisms of action. *J. Biol. Chem*., **279**, 40174–40184.

Constant,P. *et al*. (2002) Role of the *pks15/1* gene in the biosynthesis of phenolglycolipids in the *Mycobacterium tuberculosis* complex. *J. Biol. Chem*., **277**, 38148–38158.

Dandekar,T. *et al*. (1999) Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J*., **343**, 115–124.

Dijkstra,E.W. (1959) A note on two problems in connection with graphs. *Num. Math*., **1**, 269–271.

Enright,A.J. and Ouzounis,C.A. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol*., **2**, R34.

Enright,A.J. *et al*. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–89.

Forst,C.V. and Schulten,K. (1999) Evolution of metabolism: a new method for the comparison of metabolic pathways using genomic information. *J. Comp. Biol*., **6**, 343–360.

Forst,C.V. and Schulten,K. (2001) Phylogenetic analysis of metabolic pathways. *J. Mol. Evol*., **52**, 471–489.

Hershberger,J. Maxel,M. and Suri,S. (2003) Finding the *k*-shortest simple paths: a new algorithm and its implementation. In *5th Workshop on Algorithm Engineering and Experiments*. SIAM Conferences.

Ideker,T. *et al*. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Ideker,T. *et al*. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18** (Suppl. 1), S233–S240.

Kelley,B.P. *et al*. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl Acad. Sci. USA*, **100**, 11394–11399.

Kremer,L. *et al*. (2002) The *M.tuberculosis* antigen 85 complex and mycolyltransferase activity. *Lett. Appl. Microbiol*., **34**, 233–237.

Marcotte,E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol*., **10**, 259–365.

McGuire,A.M. and Church,G.M. (2000) Predicting regulons and their *cis*-regulatory motifs by comparative genomics. *Nucleic Acids Res*., **28**, 4523–4530.

Mdluli, K. *et al*. (1998) Inhibition of a *Mycobacterium tuberculosis* ketoacyl ACP synthase by isoniazid. *Science*, **280**, 1607–1610.

Nakaya,A. *et al*. (2001) Extraction of correlated gene clusters by multiple graph comparison. *Genome Inf*., **12**, 44–53.

Ogata,H. *et al*. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*., **27**, 29–34.

Ogata,H. *et al*. (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res*., **28**, 4021–4028.

Parish,T. *et al*. (2003) The *senX3-regX3* two-component regulatory system of *Mycobacterium tuberculosis* is required for virulence. *Microbiology*, **149**, 1423–1435.

Rawat,R. *et al*. (2003) The isoniacid–NAD adduct is a slow, tight-binding inhibitor of *inha*, the *Mycobacterium tuberculosis* enoyl reductase: adduct affinity and drug resistance. *Proc. Natl Acad. Sci. USA*, **100**, 13881–13886.

Reisig,W. (1985) *Petri Nets. An Introduction*. EATCS Monographs on Theoretical Computer Science. Springer Verlag.

Schnappinger,D. *et al*. (2003) Transcriptional adaptation of mycobacterium tuberculosis within macrophages: Insights into the phagosomal environment. *J. Exp. Med*., **198**, 693–704.

Wilson,M. *et al*. (1999) Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization. *Proc. Natl Acad. Sci. USA*, **96**, 12833–12838.

Zien,A., Küffner,R., Zimmer,R. and Lengauer,T. (2000) Analysis of gene expression data with pathway scores. In *Proceedings of ISMB'00*. ISCB, American Association for Artificial Intelligence, pp. 407–417.