BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 7 2012, pages 947–954
doi:10.1093/bioinformatics/bts039

Systems biology

Advance Access publication February 04, 2012

Discriminating response groups in metabolic and regulatory pathway networks

John L. Van Hemert and Julie A. Dickerson*

Bioinformatics and Computational Biology, Electrical and Computer Engineering Department, Iowa State University, Ames, IA 50011, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Analysis of omics experiments generates lists of entities (genes, metabolites, etc.) selected based on specific behavior, such as changes in response to stress or other signals. Functional interpretation of these lists often uses category enrichment tests using functional annotations like Gene Ontology terms and pathway membership. This approach does not consider the connected structure of biochemical pathways or the causal directionality of events.

Results: The Omics Response Group (ORG) method, described in this work, interprets omics lists in the context of metabolic pathway and regulatory networks using a statistical model for flow within the networks. Statistical results for all response groups are visualized in a novel Pathway Flow plot. The statistical tests are based on the Erlang distribution model under the assumption of independent and identically Exponential-distributed random walk flows through pathways. As a proof of concept, we applied our method to an *Escherichia coli* transcriptomics dataset where we confirmed common knowledge of the *E.coli* transcriptional response to Lipid A deprivation. The main response is related to osmotic stress, and we were also able to detect novel responses that are supported by the literature. We also applied our method to an *Arabidopsis thaliana* expression dataset from an abscisic acid study. In both cases, conventional pathway enrichment tests detected nothing, while our approach discovered biological processes beyond the original studies.

Availability: We created a prototype for an interactive ORG web tool at http://ecoserver.vrac.iastate.edu/pathwayflow (source code is available from https://subversion.vrac.iastate.edu/Subversion/jlv/public/jlv/pathwayflow). The prototype is described along with additional figures and tables in Supplementary Material.

Contact: julied@iastate.edu

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

Received on June 10, 2011; revised on December 13, 2011; accepted on January 18, 2012

1 INTRODUCTION

Analysis of omics experiments generates lists of entities (genes, metabolites, etc.) selected based on specific behavior. These lists are then analyzed using existing functional knowledge of the entities in a list by further listing the functional annotations assigned to the members of the list. Category enrichment analysis generally refers to testing the null hypothesis that the distribution of functional annotation in the entity list is similar to the distribution of functional annotation for all entities (Barry *et al.*, 2005; Maere *et al.*, 2005; Nettleton *et al.*, 2008; Subramanian *et al.*, 2005). If the null hypothesis is rejected, one or more of the functional annotations in the entity list is either over- or underrepresented, and a general functional response is inferred for the experimental treatment. Most biologists mine a number of sets of genes from results that exhibit an expected behavior of biological interest. These sets are mapped to functional annotation to determine which functions are associated with the behavior in the experiment. For example, a gene expression profiling assay might be conducting on samples for both a stress condition and a control. The genes which are upregulated by a certain fold-change are selected to comprise the gene set of interest and all genes are mapped to zero or more Gene Ontology terms. The proportion of each term represented in the upregulated gene set is compared with its proportion in the general population of genes (Maere *et al.*, 2005). Alternative methods compare estimates of the expression distributions across treatments to mine category-related effects (Nettleton *et al.*, 2008).

Much biological knowledge is represented as networks, whether it is ontological (Ashburner *et al.*, 2000; Avraham *et al.*, 2008; Cordero *et al.*, 2009), regulatory (RegulonDB (Gama-Castro *et al.*, 2011) or metabolic [Reactome (Matthews *et al.*, 2009), KEGG (Aoki and Kanehisa, 2005; Kanehisa *et al.*, 2010; Okuda *et al.*, 2008), PathwayTools/BioCyc (Krummenacker *et al.*, 2005) and Plant Metabolic Network (Zhang *et al.*, 2010)]. This makes functional analysis much more complex than simple set comparisons, requiring more complex tools like MapMan (Rotter *et al.*, 2009; Usadel *et al.*, 2009), Array2KEGG (Kim *et al.*, 2010) or KEGG Spider (Antonov *et al.*, 2008) to name a few. A main problem is that pathway mining is treated as a category enrichment problem based on pathway membership alone. Such an approach fails to consider the interconnectivity and reactive relationships between different entities, reactions and pathways. Further, category enrichment cannot directly infer causality; if a functional annotation term is enriched in an entity list, we cannot determine whether the functional process affects the omics entities or the omics entities affect the process. Other methods like Nayak and De (2007) partition pathway networks, but stop where our method operates: discriminating pathway partitions or response groups. Our pathway mining method does not attempt to cluster or partition the network—it assumes it has already by partitioned into response groups.

With respect to comparing networks or parts of networks, many stochastic- and flow-based methods for processing networks assume

*To whom correspondence should be addressed.

Published by Oxford University Press, 2012.

the network is *undirected*. This means each edge in the network links a pair of nodes in no particular order. This works well for applications such as protein–protein interaction (Towfic *et al.*, 2010) or co-expression networks (Mao *et al.*, 2009; van Dongen, 2000). However, biochemical pathways are inherently directed networks. *Directed* networks' edges have a specific ordering; one of the nodes a directed edge connects is the *source* and the other node is the *target* and the directed edge defines a flow from one node (the *source*) to the other (the *target*). Nodes in a pathway network represent physical entities such as genes or enzymes, as well as events such as chemical reactions. Edges between nodes represent interaction (e.g. regulation or conversion) and/or participation in an event (e.g. catalysis of a reaction). Direction is necessary to indicate the direction of reactions, i.e. which participants are catabolized and which are anabolized in a particular reaction. Methods also exist for stochastic and analytical modeling of directed networks (Hoops *et al.*, 2006; Ramsey *et al.*, 2005) for the purpose of kinetic simulations. These traditional systems biology analyses are often limited by the number of experimentally derived kinetic rate constants available for a particular organism and condition (Adiamah *et al.*, 2010; Lubitz *et al.*, 2010).

Our purpose in this work is to provide a methodology for discriminating groups of entities (*Response Groups*) in a pathway network, which are highly connected in a specific direction to a *Query List* of entities using statistically sound hypothesis testing and visualization.

1.1 Background terminology

Flow simulation models the movement of mass through a network as it travels via the graph's nodes and edges within constraints defined by the connectivity, edge weights and edge directionality. Flow simulation can be decomposed to a problem of connectivity matrix multiplication and used to solve a variety of problems. Flow-based metrics represent relationships (flows) between connected groups of vertices in a graph, whereas distance metrics like shortest path are limited to pair-wise comparisons. *Graphical Clustering* methods, such as MCL (Markov Clustering) (van Dongen, 2000), make use of flow simulation, where input is a network of nodes connected by undirected, weighted edges. The algorithm takes successive powers of the stochastic state transition probability matrix, with an inflation step at each iteration based on a single inflation parameter that degrades low-flowing edges until they vanish, creating a set of connected components which represent the resulting clusters. This method clusters data based on the structure of some meaningful undirected graph representing it, such as a correlation network as in (Mao *et al.*, 2009).

Graph kernels are functions that take adjacency matrices for two graphs and return as results a metric that usually compares the two networks (Vishwanathan *et al.*, 2010). A random walk kernel is a kernel that conducts operations on the input matrices, which simulate random walks along the edges of the input matrices' networks. Towfic *et al.* (2010) have used a state transition probability matrix multiplication called the random walk kernel to infer homologs from protein interaction networks. (For the necessary graph theory terminology, see Supplementary Material.)

A Response Group, defined for this work, is a subnetwork of a pathway network defined by a common function or other biologically meaningful grouping. For example, out of the entire network that represents known metabolism in *Escherichia coli*, the set of reactions, genes, enzymes and compounds that make up the Glycolysis pathway can also be a Response Group. A Response Group need not be connected; a connected subnetwork or module, like a pathway, is a special case with a pathway function-specific meaning. Other Response Group partitions might define all compounds of a certain class, or a group of pathways which are functionally related, but not directly connected to each other. *A Query List*, defined for this work, is a list of metabolic entities (genes, enzymes, compounds, etc.), which have been mined from an experimental dataset. Our goal is to determine which Response Groups in a pathway network are highly connected to the Query List as a whole, both in the forward (or downstream) and reverse (or upstream) directions.

2 METHODS

The Omics Response Group (ORG) method uses a pathway network to mine response groups, which are highly connected to a query list in a specific direction. ORG uses directed random walks, or flows, through the pathway network with stochastic matrix multiplication. The combined flow between the query list and each response group results in a single value representing flow between the query list and each response group, in both the forward (downstream) and reverse (upstream) pathway directions.

In order to test the significance of the results, we model the underlying distribution of flows in the network using the Exponential distribution and the summary metric, Θ_g, for each Response Group, g, using the Erlang Cumulative Distribution Function. For each Response Group, we bootstrap Erlang distribution parameters k_g (shape) and λ_g (rate) using Monte Carlo simulations of random query lists, and assess fit to the corresponding Erlang Probability Density Function. The results are corrected for multiple testing and visualized using a novel pathway flow plot. In brief, our method includes the following design decisions:

1. Receive as input a biochemical pathway network structure, a Query List of entities referred to by nodes in the pathway network, and a definition of Response Groups to discriminate.
2. Entities in a Query List could be any combination of genes, enzymes, chemical compounds, or reaction events in the pathway network and Response Group compartmentalization must be flexible; Response Groups can be the set of all functional pathways in the network, all reactions in the network or the set of all compound classes in the network, for example.
3. Response Groups must be able to overlap on zero or more entities; entities, both members and non-members of the Query List, must be able to be members of multiple Response Groups.
4. The set of all Response Groups need not cover the entire pathway network; not all nodes in the pathway network are guaranteed to be a member of any Response Group.

2.1 Modeling directed random walks

Flow simulation using directed graphs is possible using stochastic state transition probability matrices. Consider the right-stochastic state transition probability matrix $A_{N\times N}$ to represent a random walk on the network of a finite number of steps. In a given step in a random walk starting at node i, we must take a step somewhere, so the sum of the elements in the i-th row must equal one ($\sum A_{i\cdot} = 1$). On the other hand, if a random walk's step lands on node j, the step must have come from somewhere, so the sum $\sum A_{\cdot j}$ (a column in $A_{N\times N}$) should equal 1. If the network is not left-stochastic, this sum may be less or greater than one, neither of which make sense. We consider relationships in each direction separately, so we will avoid the non-left-stochastic contradiction by only considering random 'forward' steps from i to j on edges A_{ij}.

$$C_{N\times N} = \text{The directed adjacency matrix} \tag{1}$$

$$A_{N\times N} = \text{The right-stochastic transition prob. matrix} \tag{2}$$

$$A_{ij} = C_{ij} / \sum_{j=1}^{N} C_{ij} \tag{3}$$

$$M_{N\times N} = \sum_{s=1}^{w} A^{s}, \text{ for a random walk of } w \text{ steps} \tag{4}$$

For a state transition probability matrix, $A_{N\times N}$, the probability of transitioning from state i to state j in exactly w steps is $A^{w}{}_{ij}$. To obtain the probability of any of the steps reaching (or hitting) state j in a random walk of exactly w steps starting at state i, we take the sum of successive powers of A up to A^{w}, $M_{N\times N}$, which would be the matrix of hit rates in a random walk of length w steps [Equation (4)].

Given a biochemical pathway network represented by a weighted adjacency matrix, $C_{N\times N}$, we can right-stochastize it to fit the form of A in Equation (2) by dividing the values in C by the sum of their respective row as in Equation (3). Adjacency matrices for most networks are sparse, but as successive powers are summed, the resulting matrix quickly becomes dense and difficult to process. Fortunately, biochemical pathway networks, while sparse and not necessarily small-world networks (Koschutzki *et al.*, 2010; Park *et al.*, 2010), contain several hub nodes, which are highly connected to the rest of the network (e.g. energy-storing compounds and secondary messengers), allow relatively short random walk models (10–20 steps) to cover most of a pathway network—even across compartmental membranes via signaling and transport pathways. This is consistent with findings in (Koschutzki *et al.*, 2010). The resulting matrix M is the matrix of hit rates on random walks between nodes; M_{ij} is the hit rate at j of random walks of size w steps starting at i. Generally, we call this metric 'random walk flow'. This metric is preferred to others such as shortest path because flow considers *all possible* paths simultaneously, including those paths that include loops.

2.2 Summarizing flows between groups of nodes

Our goal is to determine which response groups in a pathway network are highly connected in a particular direction to a query list. Figure 1 is a hypothetical example: the query list is a list of genes that are differentially expressed under a specific condition (red-bordered nodes). The response groups could be the functional pathways, each a different color. We want to compute and test network flows between the query list nodes and each response group.

The sum of random walk hit rates between nodes in the query list and nodes in a response group summarizes the flow between the query list and response group. This is a simple matrix operation using a response group membership indicator matrix, $\Upsilon_{N\times G}$ [Equation (5)], where $\Upsilon_{ng} = 1$ if node n is a member of response group g and zero otherwise. The matrix product of the matrices $M_{N\times N}$ and $\Upsilon_{N\times G}$, $\Psi_{N\times G}$ contains the sums of flow from each node and the nodes in each response group [Equation (7)]. Similarly, the matrix product of an indicator vector, $Q_{1\times N}$ [Equation (6)] where $Q_n = 1$ if n is in the query list and zero otherwise, and $\Psi_{N\times G}$ is $\Theta_{1\times G}$, is the vector of sums of flow from the nodes in the query list to the nodes in each response group [Equation (8)].

$$\Upsilon_{ng} = \begin{cases} 1 & \text{if node } n \text{ is in response group } g \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$Q_{n} = \begin{cases} 1 & \text{if node } n \text{ is in the query list} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$\Psi_{N\times G} = M_{N\times N} \Upsilon_{N\times G} \tag{7}$$

$$\begin{aligned} \Theta_{G\times 1} &= (Q'_{N\times 1} \Psi_{N\times G})' \\ &= (Q'_{N\times 1} (M_{N\times N} \Upsilon_{N\times G}))' \end{aligned} \tag{8}$$

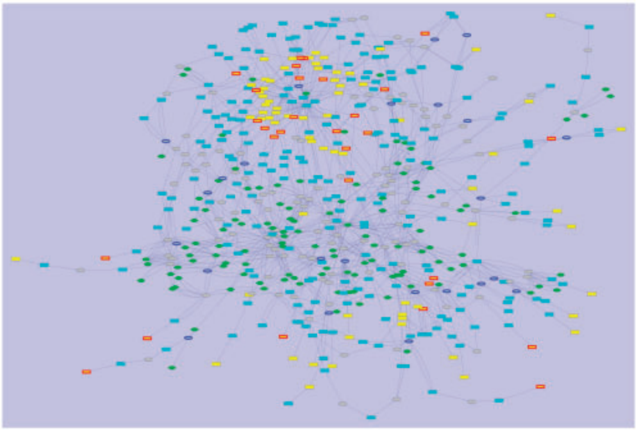

Fig. 1. A hypothetical ORG network example. Response groups are represented by different color nodes and red-bordered nodes are members of the query list. We want to determine which colors are more near the input query list than a random query list. Here, it appears the query list is generally near the yellow response group.

2.3 Modeling and testing values in Θ

After obtaining relationships for random walk flow between the query list and each response group, we want to find statistical significant flows between the query list and response groups; these are the response groups which are highly connected to the query list in a specific direction and are likely to be functionally linked to the query list. This is accomplished by a statistical test of the null hypothesis that the observed random walk flow between a response group g and the query list is equal to that of a randomly selected query list and g. There are two complicating considerations for designing such a test: Response group size and inherent connectedness.

2.3.1 Response group size Given a query list, the number of nodes and edges in each response group is variable; some response groups are larger than others. Therefore, we must account for the assumption that larger response groups are more likely to have higher random walk flow with a query list than smaller response groups. For the set of flows between the members of our query list and a given response group, we must summarize these flows to a single value which represents them as a whole. We do that with each Θ_g as it is a sum of hit rates. Using mean flows instead of sums would account for the problem because it would penalize larger response groups, i.e. Θ_g would be changed to $\frac{\Theta_g}{N_g}$, where N_g is the size of response group g. Means would also complicate the matrix operations we use to summarize the flows, and mean metrics are susceptible to outliers, which could bias our model. For this reason, we use the Erlang distribution, which has parameters that account for this problem, to model Θ_g.

2.3.2 Response group connectedness Response groups have varied connectivity with the rest of the pathway network due to its inherent structure. This can also cause bias in flow metrics where more connected response groups are more likely to have higher flows with a random query list than smaller response groups. While there may be a correlation between response group size and connectedness, it is not guaranteed, so we must account for all combinations of size and connectivity. For this reason and the fact that the distributions of flows for an arbitrary directed metabolic network are not well known, we bootstrap parameters for each Erlang-distributed Θ_g from a Monte Carlo simulation of random query lists.

2.3.3 The flow distribution underlying Θ_g We assess our assumption of the Exponential distribution of values in the M matrix [Equation (4)] using a bootstrapped sample of its values (Fig. 2). Most of these non-negative values

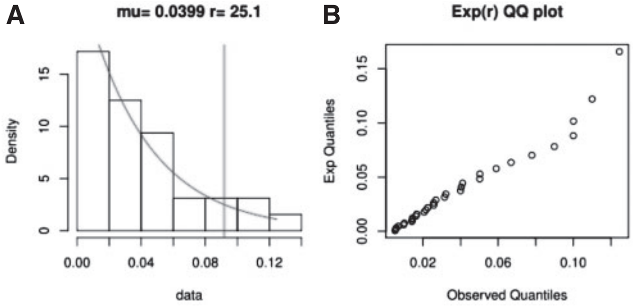

Fig. 2. General assessment of fit to an Exponential distribution for values in M on a given pathway network. The assessment includes a histogram with a fit Exponential density and vertical 95th percentile (**A**) and Quantile–Quantile (**B**) plot for all values in the matrix >0.005 from 10-step random walk simulations on the EcoCyc pathway network. We see good fit for values >0.005, indicated by good histogram fit to the curve on the left and a near-diagonal Q–Q plot on the right.

are near zero, with a skewed upper tail containing those higher random walk flow relationships. A common probability distribution with these properties is the Exponential distribution, which is often used to model waiting times for an event to occur, such as the time until a light bulb will burn out (Evans *et al.*, 2000). Plotting the bootstrapped sample from observed flows for random walks of $w=10$ steps on the EcoCyc pathway network (Keseler *et al.*, 2011) gives a good fit to an Exponential distribution for values more distant from zero.

The Exponential distribution has many useful properties. One is that the sum of k independent and identically distributed Exponential random variables with rate parameter λ follows the Erlang distribution with shape parameter k and rate parameter λ. The Erlang distribution is a special case of the Gamma distribution where the shape parameter is an integer (Evans *et al.*, 2000), which is the number of summed Exponential random variables. This makes the Erlang distribution well-suited to model our null hypothesis that the Query List is unrelated to a Response Group. Since the matrix multiplication in Equation (8) actually sums the values in M for a given query list in each response group, we can assume that the values in the Θ vectors each follow a different Erlang distribution with same shape parameter equal to the size of the query list and rate parameter equal to the inverse of the mean of all values in M [Equations (9–12)].

$$\lambda = \left(\frac{\sum M}{N^2}\right)^{-1} \tag{9}$$

$$= \frac{N^2}{\sum M} \tag{10}$$

$$M_{ij} \sim Exp(\lambda) \tag{11}$$

$$\Theta_g \sim Erlang(k, \lambda_g) \text{ , where } k \text{ is the query list size} \tag{12}$$

Assessing the Erlang-based model involves a Monte Carlo simulation, which repeatedly draws, with replacement, a random query list of k entities out of the pathway network and compute Θ each draw, building a multivariate (in the number of response groups) sampling distribution for Θ. For a given response group, we then fit an Erlang distribution to the simulated results using the convenient Erlang distribution property that the rate parameter equals the ratio of the shape parameter to the mean. With this ternary relationship, we can estimate the rate parameter by taking the ratio of the shape parameter (k) to the mean of the Monte Carlo simulation. As with the Exponential distribution above, for a given query list size and response group, we can then inspect fit by plotting the histogram of the Monte Carlo values with the density of the fit Erlang distribution as well as creating a Quantile–Quantile plot for each Θ_g (Fig. 3).

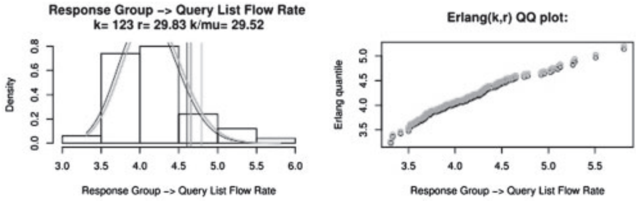

Fig. 3. Erlang assessments for Θ_g on two pathway response groups after random walk simulations of $w=10$ steps on the EcoCyc pathway network and 100 Monte Carlo simulations of flow rates with a query list of size $k=123$. Showing good fit to an Erlang density function is the assessment for an arbitrarily selected pathway response group, g, the putrescine degradation II pathway. Such an assessment can be done for each response group.

2.3.4 Hypothesis testing After computing our vector of Erlang-based test statistics, Θ, we can define a null hypothesis to test for each value in Θ, where each value represents flow between the query list and a response group. We stated earlier that the goal is to test the case where there is no flow relationship between a response group g and the query list, so our null hypothesis, H_o, is that the unknown true rate parameter, λ_g^*, equals the Monte Carlo-estimated λ_g, which can be interpreted as the rate parameter for flows between unrelated query lists and response groups [Equation (13)].

The alternative hypothesis should reflect a high flow rate between the query list and the response group g; a random query list drawn from the set of nodes, which are biologically linked to response group g would follow an Erlang distribution with the same shape parameter, k, but a larger rate parameter, λ_g. Therefore, the alternative hypothesis is the upper-tail in Equation (14).

And we reject H_0 if the observed Θ_g falls above the $(1-\alpha)$ percentile of the Erlang distribution with shape k and rate λ_g, where α is a selected Type I Error Rate, which is the rate at which the null hypothesis is rejected incorrectly [Equation (15)].

$$H_0 \ : \ \lambda_g^* = \lambda_g \tag{13}$$

$$H_A \ : \ \lambda_g^* > \lambda_g \tag{14}$$

$$\text{Rej. } H_0 \text{ if } P_{Erl_{k,\lambda_g}}\left(X > \Theta_g\right) \le \alpha \tag{15}$$

When multiple hypothesis tests are conducted simultaneously, the family-wise error rate (FWER) is inflated by the number of tests; e.g. if we conduct 10 tests, each with $\alpha=0.01$, each test has a 1% probability of making a Type I Error, but the overall probability of making a Type I Error is the sum of each α, or $0.01 \times 10 = 0.1$. The *multiple testing* problem has been a focus for mining high-throughput data, such as microarrays because families of tests are conducted on thousands of genes in this field creating strong demand for clever correction methods. The most straightforward and conservative correction, Bonferronni (Holm, 1979), simply uses a corrected α value for tests equal to the original desired Type I Error Rate divided by the number of tests, $\alpha' = \frac{\alpha}{m}$, where m is the number of tests. Several more complex methods exist which focus on the false discovery rate and estimate parameters for the specific distribution that P-values follow for microarray experiments (Fodor *et al.*, 2007; Storey, 2003; Storey *et al.*, 2004; Storey and Tibshirani, 2003), where P-values are uniformly distributed between zero and one with a spike near zero containing the relatively large set of genes perturbed by the experiment. However, P-values for response groups are not always expected to follow such a distribution because there are often only a few significant response groups in one of our analyses. For this reason, we discretionarily use Bonferroni correction to correct for multiple testing where an independent test is conducted for each response group.

2.4 Reversing directionality

The previous formulation results in random walk flow summarizations from the query list to response groups, i.e. pathways which the query list

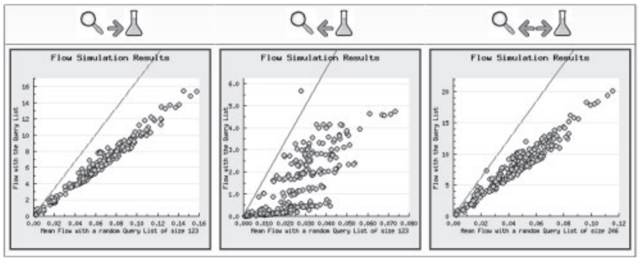

Fig. 4. Example PathwayFlow plots for *E.coli* pathways. Results are visualized with a plot of the response groups for each direction and the total. The Y-axes are Θ, $\Theta^{(rev)}$ and $\Theta^{(tot)}$, respectively. The X-axes are the inverses of λ, $\lambda^{(rev)}$ and $\lambda^{(tot)}$, respectively, which are also the expected values of the Θ's for the respective Monte Carlo simulations. Points near the origin are response groups that are small and/or poorly connected to the rest of the network. Points high and far to the right are large and highly connected (hub-like) response groups. The diagonal lines mark the null hypothesis rejection cutoff, given the confidence level, correction and λ value (X-coordinate).

regulate. The question of what is regulating the query list, or signaling its members to change behavior, is often equally or even more interesting. We can reverse direction rearranging the matrix multiplication to compute a different, but analogous set of metrics, statistics and hypothesis test [Supplementary Equations (S1)–(S7)] resulting in another vector of flow rates for each response group, $\Theta^{(rev)}$ and hypothesis test [Supplementary Equations (S8)–(S10)]. The difference is that these reversed flow rates represent flow summaries from the response groups to the query list.

In order to compute reverse flows, we reinitialized the random walk rate matrix, A, as $A^{(rev)}$ by left-stochastizing the adjacency matrix [Supplementary Equation (S1)] because reverse direction focuses on backtracking the directed graph using arrival probabilities, which are represented by columns in $A^{(rev)}$. Again, the alternative hypothesis is in the upper tail of Supplementary Equation (S9).

In addition to the two sets of analogous flow metrics, we can also compute a third set [Supplementary Equations (S11)–(S16)], which reflects a sum of the first two. This might be used as an in initial investigation or in cases where directionality in the pathway networks is not well known. If we sum the flow rates in both directions at the Ψ step, we get flow rate summaries between the query list and each response group overall [in both directions, $\Theta^{(tot)}$ in Supplementary Equation (S11)]. Like before, the alternative hypothesis is in the upper tail of Supplementary Equation (S15). Assessments in Supplementary Figures S5 and S6 show that our assumptions are accurate for these directions as well.

2.5 Visualizing results with PathwayFlow plots

Results for all response groups can be visualized simultaneously for manually investigation and selection. We plot a data point for each response group, g, in a 2D space where the X-axis is the inverse of the fit rate parameter, λ, and the Y-axis is the observed flow statistic, Θ_g. We can then plot a one-tailed Erlang confidence interval boundary line marking the lowest Y-axis locations a response group can exist and have a significantly high flow with the query list (Fig. 4).

3 EXAMPLES

For real data examples, we applied the ORG method to publicly available microarray datasets for *E.coli* and *A.thaliana* (Figure 5). We chose these model organisms because there are relatively rich pathway networks available from EcoCyc (Keseler *et al.*, 2011) and AraCyc (Rhee *et al.*, 2005; Zhang *et al.*, 2005). Both datasets pertain to well-studied processes and we show that our results are more complete than existing methods and are consistent with the literature (Batchelor *et al.*, 2005; Loui *et al.*, 2009; Raetz *et al.*, 2006).

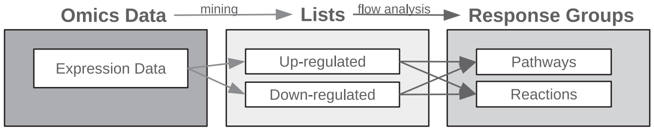

Fig. 5. From the data, we mined lists of genes. For each of those lists, we performed our flow analysis for two different types of response groups.

Genes, protein products, metabolites, and reaction and regulation events were grouped into response groups representing pathways found in the EcoCyc version 15.0 and AraCyc version 7 metabolic networks. For each experiment, two different query lists were mined using GEO's T-test data analysis tool. Single-tailed tests at the 90% confidence level for 'control < treatment' and 'treatment > control' created a query list for upregulated genes and downregulated genes, respectively. These lists are not actually genes, but probeset identification numbers that do not exist in our reference pathway networks. We mapped the probesets IDs to corresponding EcoCyc and AraCyc genes according to common locus IDs.

3.1 Lipid A inhibition in *E.coli*

The data for this use case comes from the GEO (Barrett *et al.*, 2009) dataset accession GDS3597 by Zhu *et al.* (2009), who investigated transcriptional regulation by FabR of the fatty acid biosynthesis genes fabA and fabB in the presence of endogenous and/or exogenous unsaturated fatty acids. Among other factors in their experiment, gene expression was measured in a control and treatment with CHIR-090, an antibiotic that inhibits the biosynthesis of Lipid A (Barb *et al.*, 2007). Lipid A is the anchor by which lipopolysaccharides attach to the outer membrane of Gram-negative bacteria, which provide much of the cell's structural stability and are also recognized by immune systems.

In all, 123 upregulated EcoCyc genes were identified from the list of probesets whose expression is induced when lipid A synthesis was inhibited. Our method results in three different PathwayFlow plots: one for each of the forward direction, reverse direction and total, respectively [Supplementary Fig. S4(a)]. We used Bonferroni correction at the 95% confidence level and the red cutoff lines are drawn accordingly. Pathways that fall above the red lines have significantly high flow with the query list and are listed below with P-values. The superpathway of KDO$_2$-lipid A biosynthesis is the only pathway that is a significant successor (downstream in the directed pathway network) to our query list of upregulated genes, with a $P < 0.0001$. This is the expected result when the cells are unable to produce the lipid A they require for membrane structure; they are increasing their efforts to produce more lipid A. The CpxAR Two-Component Signal Transduction System is the only significant predecessor (upstream) pathway to our query list of upregulated genes, with a $P < 0.0001$. This is a signaling system which senses cell envelope stress (Wolfe *et al.*, 2008), which is also expected because we can interpret our results as evidence for CpxAR signaling the increased expression of the genes in our query list. The CpxAr system responds to cell envelope stress and regulates transcription of the porin genes ompF and ompC, and a loss of function mutation in cpxAr can result in increased transcription of ompC and decreased transcription of ompF (Batchelor *et al.*, 2005). Pathway enrichment analysis using hypergeometric tests detected nothing.

In all, 81 downregulated EcoCyc genes were identified from the list of probesets switched to lower expression when lipid A synthesis was inhibited, and pathway response groups are plotted in Supplementary Figure S4(b). When conservatively using Bonferroni correction, two pathways were significant, but some pathways are plotted very near the significance cutoff line. We adjusted the confidence level to 99% and skipped Bonferroni correction to be slightly less conservative with our Erlang tests. The TorSR and ZraSR Two-Component Signal Transduction Systems are the significant successor (forward direction) pathways to the query list. The TorSR system regulates use of Trimethylamine *N*-oxide (TMAO), which is both an osmoprotectant and alternative electron acceptor during anaerobic respiration (Ansaldi *et al.*, 2000). The ZraSR system senses toxic levels of zinc and lead in the periplasm. The CpxAr Two-Component Signal Transduction System and Acetoacetate Degradation to Acetyl CoA pathways are the significant predecessors (reverse direction) to the query list. Recall that the CpxAr system also appeared in the significant predecessor pathways of the upregulated query list, which might be explained by a feedback loop. In this case, the CpxAr system has an extremely low P-value, so if we adjust the confidence level for the Erlang test, it will not drop out of either the upregulated or the downregulated pathways. If we adjust the confidence level for the T-tests used to generate the up- and downregulated gene lists, we checked whether the Erlang test results change. After entering query lists based on T-test at the 95% confidence level, results for the upregulated pathways in both forward and reverse directions as well as downregulated successors (forward direction) remained constant, while downregulated predecessors (reverse direction) changed from CpxAr to the DpiAB Two-Component Signal Transduction System, which regulates citrate fermentation genes. The DpiAB system is also known to interrupt chromosome duplication in the SOS response (Yamamoto *et al.*, 2008). Acetoacetate degradation feeds carbon energy into the TCA cycle (Pauli and Overath, 1972) and genes for this production are negatively regulated by ArcA. Pathway enrichment analysis using hypergeometric tests and the MRPP method (Nettleton *et al.*, 2008) detected nothing.

In summary, lipid A inhibition causes a breakdown of the cell's structure and osmotic stress, which the cell senses and responds with several different methods of action. First, it activates the genes to produce both the inhibited lipid A (Table 1{1}) and the KDO (Table 1{2}) that the lipid A should be anchoring to the cell membrane. It also shifts priorities away from growth (Table 1{6,9}), toxin sensing in the periplasm (Table 1{5,9}) and osmoprotectant production (Table 1{7}). OmpR activation is increased because both OmpC and OmpF porins production require it (Table 1{3}), but since the promoter for ompf has higher affinity for OmpR-P than the promoter for ompc, ompF transcription is specifically decreased using a separate mechanism (Table 1{4}) so that only OmpC porins are produced. Most of these inferences are consistent with the literature, and we can hypothesize that the cell knows that the osmotic stress is caused by structural insufficiencies and not by a severe change in solute concentrations, so it chooses not to produce osmoprotectant. See Supplementary Figure S10 for a flowchart of the transcriptional response.

Table 1. Interpretation of different flow simulations and tests, confirmed by the literature

	Successors (forward)	Predecessors (reverse)
Up	**Activated by the query list**	**Activate the query list**
	{1} KDO$_2$-lipid A biosynthesis	{4} CpxAR signalling
	{2} Arabinose-5-phosphate isomerase	{5} nitrate and nitrite sensors
	{3} OmpR phosporylation	{6} ArcAB
Down	**De-activated by the query list**	**De-activate the query list**
	{7} TorSR signalling	{9} DpiAB signalling
	{8} ZraSR signalling	

Numbers in { } are directly referenced in Section 4.

3.2 ABA signaling in *A.thaliana*

Expression data from GEO dataset accession GDS2730 by Kuhn *et al.* (2008) was used for a second example. In this study, the role of abh1 in abscisic acid (ABA) signaling was investigated. ABA signaling is known to be involved in many plant processes including but not limited to seed development and abiotic stress response (Kuhn *et al.*, 2008). The two query lists contain genes up- and downregulated under ABA treatment. AraCyc 7 lacks gene regulatory data, meaning gene predecessors cannot be mined as they can be in EcoCyc. Instead, we can only discriminate downstream of the gene query lists.

In all, 259 probesets were identified as upregulated and mapped to AraCyc genes. The pathways galactosylcyclitol biosynthesis, ajugose biosynthesis II (galactinol-independent) and ajugose biosynthesis I (galactinol-dependent) produced significant Erlang tests (all $P < 0.0001$). These pathways are all related to stachyose synthesis, which is known to be the dominant sugar found in the 'resurrection plant', *Craterostigma plantagineum* and can play a role in stress response (Bartels and Salamini, 2001). Pathway enrichment analysis using hypergeometric tests and the MRPP method (Nettleton *et al.*, 2008) detected nothing.

Like the *E.coli* example, we took a subset of data from a previous study and used our method to discover relevant Response Groups different from the processes investigated in the original study, but consistent with the literature.

4 DISCUSSION

We took a subset of expression data from work investigating fab genes' roles in *E.coli* cell membrane homeostasis and uncovered new insights supported by the literature. We clearly saw activity relevant to the cell's boundary (envelope and periplasm), which is consistent with our understanding of the utility of lipid A. Likewise, the set of upregulated *A.thaliana* genes under ABA exposure tested significantly connected to stress-related responses. We compared results to two category enrichment testing methods where pathway membership according to respective BioCyc databases was used to assign categories: hypergeometric tests, a simple approach which tests for category overrepresentation in the query list, and the MRPP method which tests multivariate distributions of expression across categories and has been shown to perform better than other methods (Nettleton *et al.*, 2008). Our ORG method detected pathways which fit existing literature while these methods detected none, which illustrates the main problem we set out to ameliorate. Category enrichment remains a common way to mine pathway databases for functional insights into lists of differential expression genes. However, pathways are not categories—they are subnetworks of a larger, interconnected metabolic and regulatory system. A query list can be functionally related to a pathway without containing any members of the pathway, so we proposed computing the query list–pathway relationship using flow metrics and then simultaneously plotting results for all pathways with our pathway flow plot. Our method can be generalized to mine user-defined response groups, which might be pathway subnetworks or other subnetworks that may exhibit any or no connectivity. For example, we also applied our method to mine individual reactions instead of entire pathways for the *E.coli* data and results were also consistent with the previous work (Supplementary Material). Existing category enrichment methods are not applicable in this case because it relies on *membership* in the response groups by many of the membership of the query list, which is by definition rare for individual reactions.

While our method does not directly infer new pathway models, it uses existing knowledge about pathways to generate hypotheses about the members of a query list. Therefore, adding hypothesized relationships to pathway models can strengthen the results. It uses a biochemical pathway network structure, a Query List of entities, and a set of Response Groups as input, in order to discriminate Response Groups that are highly connected to the Query List from those that are not. Entities in a Query List could be any combination of genes, enzymes, chemical compounds or reaction events in the pathway network while Response Groups can be the set of all functional pathways in the network, all reactions in the network or any other potentially overlapping or incomplete entity grouping such as the set of all compound classes in the network, for example. It can visualize statistical hypothesis test results for decision support and discretionary test parameter adjustments, and the hypothesis test accounts for both Response Group size and inherent connectivity with the rest of the network. Our method is distinctly different from kinetic simulation, which uses metabolic networks and experimentally derived rate constants to predict relative concentrations of all entities.

Further, our tools integrate with the MCL suite of tools (van Dongen, 2000), which includes a preprocessing step to remove cycles. We can run our method with or without this step. We compared both results on our test data and saw little difference. This makes sense because our method computes ran- dom walk hit rates across all possible paths; cycles contribute to these rates, but their contribution is dwarfed by that of all other paths.

The main weakness of the ORG method is sensitivity to missing information from the pathway network, i.e. nodes and relationships are missing from the network. If the function of an entity in a query list is not well understood, the best we can do with our method is assume 'guilt by association' and infer its involvement in the response groups we associate with the well-understood entities in the query list. Methods exist for filling gaps, but they require either human decision making, extra experimental data or flux data (Orth and Palsson, 2010). We considered pathway and regulatory network gap-filling and *de novo* construction beyond the scope of this article so that we could present a methodology for mining existing biological knowledge. Another weakness is that reverse flow results are only possible when the pathway network contains an adequate amount of gene–regulatory relationships, which are represented by edges and flows into genes. Cycles and feedback loops might create ambiguity between significant successor and predecessor response groups, though we did not find significant differences when removing all loops.

As a proof of concept, we created a web-based implementation of our method (See Supplementary material) and applied it to omics data from a simple *E.coli* microarray dataset, verified the results with the literature and generated new hypotheses. For this dataset, we also compared our method to hypergeometric test-based enrichment analysis. When correcting for multiple testing, enrichment testing detects no significant response groups. When (incorrectly) omitting multiple testing correction, enrichment tests produce lists of response groups not associated with the known behavior of the dataset, while our method produced very similar results whether or not we correct for multiple testing, indicating that our method is also robust to multiple testing. Our method outperforms pathway enrichment testing because our method considers pathway network connectivity rather than direct pathway or group membership; it can detect relationships between a query list and response group even if the members of the query list are not members of the response group. It also has the ability to detect directionality in those relationships, which enrichment analysis lacks.

Future work includes application to more diverse omics datasets, which include compounds and enzymes as well as more diverse collections of response groups such as transcription factor families and reaction classes. Our tool is compatible with output from the Markov Clustering software (van Dongen, 2000) and we intend to investigate Response Groups defined by graphical clusters mined from large metabolic networks. Lastly, the web tool is to be fully integrated with the PLEXdb.org (Shen *et al.*, 2005) website.

ACKNOWLEDGEMENTS

The authors thank Basil Nikolau, Roger Wise, Gustavo MacIntosh, and Peng Liu for their valuable feedback and advice in this work.

Funding: National Science Foundation programs DBI (grant number 0604755) and EEC (grant number 0813570). Per Q12, change to Section heading.

Conflict of Interest: none declared.

REFERENCES

Adiamah,D.A. *et al.* (2010) Streamlining the construction of large-scale dynamic models using generic kinetic equations. *Bioinformatics*, **26**, 1324–1331.

Ansaldi,M. *et al.* (2000) The torr high-affinity binding site plays a key role in both torr autoregulation and torcad operon expression in escherichia coli. *J. Bacteriol.*, **182**, 961–966.

Antonov,A.V. *et al.* (2008) Kegg spider: interpretation of genomics data in the context of the global gene metabolic network. *Genome Biol.*, **9**, 11.

Aoki,K.F. and Kanehisa,M. (2005) Using the KEGG database resource. *Curr. Protocols Bioinformatics*, **Chapter 1**, Unit 1.12.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Avraham,S. *et al.* (2008) The plant ontology database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Res.*, **36**, D449–D454.

Barb,A.W. *et al.* (2007) Inhibition of lipid A biosynthesis as the primary mechanism of CHIR-090 antibiotic activity in Escherichia coli. *Biochemistry*, **46**, 3793–3802.

Barrett,T. *et al.* (2009) Ncbi geo: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.

Barry,W.T. *et al.* (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.

Bartels,D. and Salamini,F. (2001) Desiccation tolerance in the resurrection plant craterostigma plantagineum. a contribution to the study of drought tolerance at the molecular level. *Plant Physiol.*, **127**, 1346–1353.

Batchelor,E. *et al.* (2005) The Escherichia coli cpxa-cpxr envelope stress response system regulates expression of the porins ompf and ompc. *J. Bacteriol.*, **187**, 5723–5731.

Cordero,F. *et al.* (2009) *Ontology-Driven Co-clustering of Gene Expression Data*, Vol. 5883. Springer, Berlin, pp. 426–435.

Evans,M. *et al.* (2000) *Statistical Distributions*, 3rd edn. Chapter 12: Erlang Distribution. Wiley, Hoboken, New Jersey.

Fodor,A.A. *et al.* (2007) Towards the uniform distribution of null p-values on affymetrix microarrays. *Genome Biol.*, **8**, R69.

Gama-Castro,S. *et al.* (2011) RegulonDB version 7.0: transcriptional regulation of escherichia coli k-12 integrated within genetic sensory response units (Gensor units). *Nucleic Acids Res.*, **39**, D98–D105.

Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.

Hoops,S. *et al.* (2006) COPASI-a COmplex PAthway SImulator. *Bioinformatics*, **22**, 3067–3074.

Kanehisa,M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Keseler,I. *et al.* (2011) EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res.*, **39**, D583–D590.

Kim,J.S. *et al.* (2010) Array2kegg: Web-based tool of kegg pathway analysis for gene expression profile. *Biochip J.*, **4**, 134–140.

Koschutzki,D. *et al.* (2010) Structural analysis of metabolic networks based on flux centrality. *J. Theor. Biol.*, **265**, 261–269.

Krummenacker,M. *et al.* (2005) Querying and computing with biocyc databases. *Bioinformatics*, **21**, 3454–3455.

Kuhn,J.M. *et al.* (2008) mRNA cap binding proteins: effects on abscisic acid signal transduction, mRNA processing, and microarray analyses. *Curr. Topics Microbiol. Immunol.*, **326**, 139–150.

Loui,C. *et al.* (2009) Role of the arcab two-component system in the resistance of escherichia coli to reactive oxygen stress. *BMC Microbiol.*, **9**, 183.

Lubitz,T. *et al.* (2010) Parameter balancing in kinetic models of cell metabolism. *J. Phys. Chem. B*, **114**, 16298–16303.

Maere,S. *et al.* (2005) Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Mao,L.Y. *et al.* (2009) Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics*, **10**, 346.

Matthews,L. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes *Nucleic Acids Research*, **37**(Database issue), D619–622. PMID: 18981052.

Nayak,L. and De,R.K. (2007) An algorithm for modularization of MAPK and calcium signaling pathways: comparative analysis among different species. *J. Biomed. Informat.*, **40**, 726–749.

Nettleton,D. *et al.* (2008) Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*, **24**, 192–201.

Okuda,S. *et al.* (2008) KEGG atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.*, **36**, W423–W426.

Orth,J.D. and Palsson,B. (2010) Systematizing the generation of missing metabolic knowledge. *Biotechnol. Bioeng.*, **107**, 403–412.

Park,C.Y. *et al.* (2010) Simultaneous genome-wide inference of physical, genetic, regulatory, and functional pathway components. *Plos Comput. Biol.*, **6**, e1001009.

Pauli,G. and Overath,P. (1972) Ato operon - a highly inducible system for acetoacetate and butyrate degradation in escherichia coli. *Eur. J. Biochem.*, **29**, 553.

Raetz,C.R.H. *et al.* (2006) Kdo(2)-lipid a of escherichia coli, a defined endotoxin that activates macrophages via tlr-4. *J. Lipid Res.*, **47**, 1097–1111.

Ramsey,S. *et al.* (2005) Dizzy: stochastic simulation of large-scale genetic regulatory networks. *J. Bioinformatics Comput. Biol.*, **3**, 415–436.

Rhee,S.Y. *et al.* (2005) AraCyc: overview of an arabidopsis metabolism database and its applications for plant research. In: Saito,K. *et al.*, (eds), *Plant Metabolomics*, Vol. 57, Springer-Verlag, Berlin/Heidelberg, pp. 141–154.

Rotter,A. *et al.* (2009) Gene expression profiling in susceptible interaction of grapevine with its fungal pathogen eutypa lata: extending mapman ontology for grapevine. *BMC Plant Biol.*, **9**, 104.

Shen,L.H. *et al.* (2005) Barleybase - an expression profiling database for plant genornics. *Nucleic Acids Res.*, **33**, D614–D618.

Storey,J.D. (2003) The positive false discovery rate: a bayesian interpretation and the q-value. *Ann. Stat.*, **31**, 2013–2035.

Storey,J.D. *et al.* (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **66**, 187–205.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proce. Natl Acad. Sci. USA*, **102**, 15545–15550.

Towfic,F. *et al.* (2010) Detection of gene orthology from gene co-expression and protein interaction networks. *BMC Bioinformatics*, **11**, S7.

Usadel,B. *et al.* (2009) A guide to using mapman to visualize and compare omics data in plants: a case study in the crop species, maize. *Plant Cell Environ.*, **32**, 1211–1229.

van Dongen,S. (2000) *Graph Clustering by Flow Simulation*. PhD thesis. University of Utrecht, Utrecht, The Netherlands.

Vishwanathan,S.V.N. *et al.* (2010) Graph kernels. *J. Mach. Learn. Res.*, **11**, 1201–1242.

Wolfe,A.J. *et al.* (2008) Signal integration by the two-component signal transduction response regulator cpxr. *J. Bacteriol.*, **190**, 2314–2322.

Yamamoto,K. *et al.* (2008) Anaerobic regulation of citrate fermentation by CitAB in Escherichia coli. *Biosci. Biotechnol. Biochem.*, **72**, 3011–3014.

Zhang,P. *et al.* (2010) Creation of a genome-wide metabolic pathway database for populus trichocarpa using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol.*, **153**, 1479–1491.

Zhang,P *et al.* (2005) MetaCyc and AraCyc. metabolic pathway databases for plant research. *Plant Physiology*, **138**, 27—37.

Zhu,K. *et al.* (2009) Transcriptional regulation of membrane lipid homeostasis in escherichia coli. *J. Biol. Chem.*, **284**, 34880–34888.