BIOINFORMATICS

Vol. 30 ISMB 2014, pages i175–i184
doi:10.1093/bioinformatics/btu290

MIRA: mutual information-based reporter algorithm for metabolic networks

A. Ercument Cicek[1,*], Kathryn Roeder[1] and Gultekin Ozsoyoglu[2]

[1]Lane Center for Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA 15213 and [2]Department of Electrical Engineering and Computer Science, School of Engineering, Case Western Reserve University, Cleveland, OH, USA 44106

ABSTRACT

Motivation: Discovering the transcriptional regulatory architecture of the metabolism has been an important topic to understand the implications of transcriptional fluctuations on metabolism. The reporter algorithm (RA) was proposed to determine the hot spots in metabolic networks, around which transcriptional regulation is focused owing to a disease or a genetic perturbation. Using a z-score-based scoring scheme, RA calculates the average statistical change in the expression levels of genes that are neighbors to a target metabolite in the metabolic network. The RA approach has been used in numerous studies to analyze cellular responses to the downstream genetic changes. In this article, we propose a mutual information-based multivariate reporter algorithm (MIRA) with the goal of eliminating the following problems in detecting reporter metabolites: (i) conventional statistical methods suffer from small sample sizes, (ii) as z-score ranges from minus to plus infinity, calculating average scores can lead to canceling out opposite effects and (iii) analyzing genes one by one, then aggregating results can lead to information loss. MIRA is a multivariate and combinatorial algorithm that calculates the aggregate transcriptional response around a metabolite using mutual information. We show that MIRA's results are biologically sound, empirically significant and more reliable than RA.

Results: We apply MIRA to gene expression analysis of six knockout strains of *Escherichia coli* and show that MIRA captures the underlying metabolic dynamics of the switch from aerobic to anaerobic respiration. We also apply MIRA to an Autism Spectrum Disorder gene expression dataset. Results indicate that MIRA reports metabolites that highly overlap with recently found metabolic biomarkers in the autism literature. Overall, MIRA is a promising algorithm for detecting metabolic drug targets and understanding the relation between gene expression and metabolic activity.

Availability and implementation: The code is implemented in C# language using .NET framework. Project is available upon request.

Contact: cicek@cs.cmu.edu

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

1 INTRODUCTION

Changes with respect to environmental or genetic modifications lead to complex cellular responses. Standing on the top of the omics hierarchy, metabolomics reflects changes taking place in the transcriptome and in the genome. These responses are analyzed by researchers to discover regulatory mechanisms and dynamics of cells. Transcriptional responses of the cells, along with the corresponding metabolic alterations, have been investigated in various contexts. Some examples are plant research (Brosché *et al*., 2005; Carari *et al*., 2006), diabetes (Ferrara *et al*., 2008; Zelezniak *et al*., 2010), insulin resistance (Jans *et al*., 2011) and cancer (Schramm *et al*., 2010). Functional class-based (Gerstein and Jansen, 2000; Hughes *et al*., 2000; Karp *et al*., 2002; Pavlidis *et al*., 2002; Seshasayee *et al*., 2009) and protein–protein interaction network-based analysis of gene expression data (Chowdhury *et al*., 2010; Chuang *et al*., 2007; Goh *et al*., 2007; Ideker *et al*., 2002; Rhodes and Chinnaiyan, 2005) have been well established. Metabolic network- and metabolic pathway-driven analysis of transcriptional data have been receiving attention lately. Efforts have centered on discovering transcriptional regulation architecture of metabolic networks of organisms using genome-wide association studies (David *et al*., 2006; Ihmels *et al*., 2004; Kharchenko *et al*., 2005; Tanay *et al*., 2004). Various methods with different goals have been developed to use transcriptomic and metabolic data together in the context of a metabolic network such as (i) mining for new metabolite-gene/transcription factor relationships (Ideker *et al*., 2001; Yeang *et al*., 2006), (ii) flux balance analysis and constraint-based modeling of organisms (Covert and Palsson, 2002a, b; Shlomi *et al*., 2008) and (iii) using metabolic network topology to identify significant changes in related groups of genes (Cakir *et al*., 2006; Deo *et al*., 2010; Dinu *et al*., 2007; Hancock *et al*., 2012; Nam *et al*., 2009; Oliveira *et al*., 2008; Patil and Nielsen, 2005; Subramanian *et al*., 2005; Ulitsky and Shamir, 2009).

Network topology-based analysis of biological data is a broad research area (Ma'ayan, 2008). In the context of metabolic networks and transcriptomics, the literature so far can be divided into two subcategories. The first type of analysis uses predefined metabolic pathways as targets for transcriptional regulation and analyzes the changes in the pathways. Gene Set Enrichment Analysis (GSEA) is the first and most established analysis in this subcategory (Subramanian *et al*., 2005). Improvements have been proposed in the literature to eliminate the shortcomings of the GSEA approach (Dinu *et al*., 2007; Draghici *et al*., 2007; Hancock *et al*., 2012). The second type of analysis considers the metabolic network as a whole, and aims to find signatures or hot spots in the metabolic network that are subject to transcriptional regulation (Cakir *et al*., 2006; Nam *et al*., 2009; Patil and Nielsen, 2005; Schramm *et al*., 2010). The most established method in this group is the reporter algorithm (RA; Patil and Nielsen, 2005). Using the z-score-based method introduced before (Ideker *et al*., 2002), the algorithm aims to find metabolites around which transcriptional regulation is centered and link the complex transcriptional motives to metabolome. Various

*To whom correspondence should be addressed.

extensions and modifications to the algorithm have been published in the literature. For instance, same idea has been used to discover reporter reactions (Cakir *et al*., 2006). A similar z-score-based approach has also been used analyze the rate-limiting steps of pathways (Nam *et al*., 2009). In this article, we focus on the original RA (Patil and Nielsen, 2005) and its scoring mechanism (Ideker *et al*., 2002). Our observations, discussed next, also apply to the extensions of the RA method.

RA first maps the differential data onto the enzymes (reactions) in the metabolic network, and then calculates the *P*-values for each gene, using student's *t*-test. *P*-values are then converted to z-scores using the inverse normal cumulative distribution (θ^{-1}). Equation (1) shows how each p_i is converted to the corresponding z-score z_i. Given all samples, z-score measures how many standard deviations away a *P*-value is. It ranges from negative to positive infinity, where negative infinity corresponds to no significance at all.

$$z_i = \theta^{-1}(1 - p_i) \tag{1}$$

The z-score z_m for a metabolite *m* is the aggregation of z-scores of the *k* enzymes that are neighbors of *m* in the metabolic network (through metabolic reactions), and calculated as shown in Equation (2). The null hypothesis is that the genes adjacent to a metabolite display their normalized average response by chance. Should the significance of a metabolite need to be determined, z-score is normalized and converted back into the corresponding *P*-value.

$$z_m = \frac{1}{\sqrt{k}} \sum z_i \text{ such that enzyme i is a neighbor of metabolite m} \tag{2}$$

Although RA has been shown to be effective in several different contexts such as analysis of Type 2 Diabetes (Zelezniak *et al*., 2010), genome scale analysis of organisms (David *et al*., 2008; Usaite *et al*., 2009), genome scale analysis of cell lines (Agren *et al*., 2012), gene knockout analysis (Holm *et al*., 2010; Cimini *et al*., 2009), targeted pathway analysis (Vongsangnak *et al*., 2009), there are some shortcomings that may affect the accuracy of the algorithm. First, the z-score method uses student's *t*-test to measure the amount of change between two variables. However, the resulting *P*-value is highly dependent on the degrees of freedom. It has been shown that, owing to the small number of samples in cohorts, *P*-values may not work as intended (Cicek *et al*., 2013). Second, RA uses a univariate approach. That is, it determines the changes per gene, and does not take dependencies among genes into account. For a reaction associated with multiple genes, the gene with the highest z-score is used, and the rest are discarded (Zelezniak *et al*., 2010).

In Figure 1A, we illustrate the problem via an example. The figure shows an example from the application of the RA algorithm to compare ΔarcAΔfnr and wild-type (WT) strains of *Escherichia coli* (aerobic) in the gene expression dataset (Covert *et al*., 2004) (please see Section 2 for details). Genes (pentagons) are assigned z-scores first, and then the maximum z-score is selected and assigned as the z-score of the reaction (rectangles).

For the reaction *copragon transport via ton system*, the maximum z-score belongs to the gene *fhuE* (0.88). This scoring ignores the contribution of the gene *tonB*, as any z-score value for *tonB* in

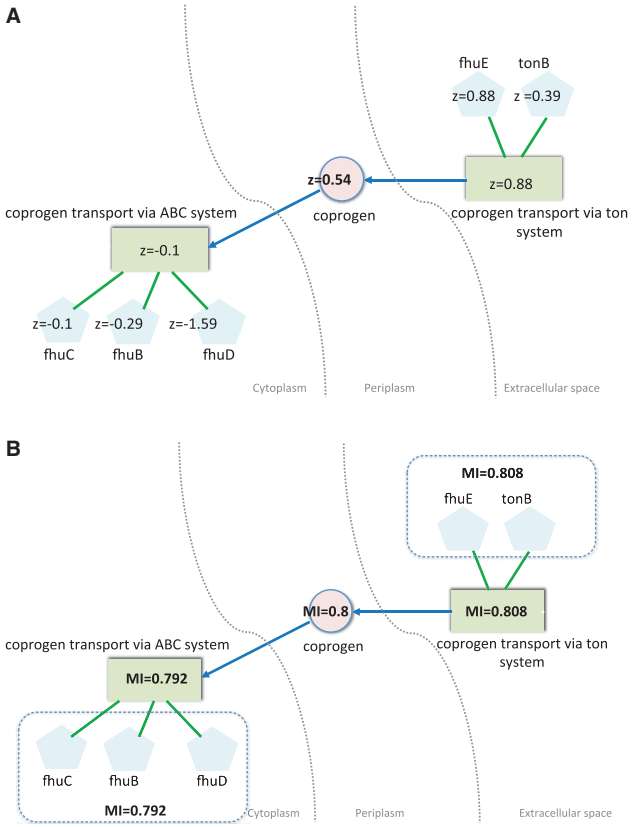

Fig. 1. Application of RA in comparison of ΔarcAΔfnr and WT strains of *E.coli* (aerobic). Rectangles represent reactions, pentagons represent genes and the circle represents the metabolite, copragon. (**A**) Reactions transfer copragon from extracellular space to periplasm, and then to cytoplasm. The maximum change (z-score) for the genes of *copragon transport via ABC system* is −0.1, and the genes of *copragon transport via ton system* is 0.88. Aggregate z-score for the metabolite *coprogen* is assigned as 0.54. (**B**) When genes of *copragon transport via ABC system* are considered together they return MI of 0.792, and MI for the genes of *copragon transport via ton system* is 0.808. Average turns out to be 0.8, which is a relatively high mutual information value. Result is different than the prediction made by RA

the range [−Infinity, 0.88) would yield the same z-score for the reaction. Finally, the method is additive in aggregating z-scores of each neighboring reaction to determine the z-score of a metabolite [as shown in Equation (2)]. However, z-score ranges from negative to positive infinity, negative infinity representing the most insignificant case. Therefore, averaging individual results introduces the problem of opposite signs cancelling each other out. In Figure 1A, *copragon* is assigned a z-score: $\frac{1}{\sqrt{2}}(-0.1 + 0.88) = 0.54$. Negative z-score (−0.1) assigned to the reaction *copragon transport via ABC system* partially cancels out the z-score of the reaction *copragon transport via ton system* (0.88). The problem would have been resolved if the scoring mechanism used, assigned zero to the most insignificant case ($P = 1$).

In this article, we present a new algorithm, called *Mutual Information-based Reporter Algorithm* (*MIRA*) that addresses the shortcomings of the original RA. MIRA is a multivariate and combinatorial algorithm that calculates the aggregate transcriptional response around a metabolite using mutual

information. Mutual Information is an information theoretic method that measures how much knowing one variable reduces the uncertainty about the other. In gene expression analysis, it has been used frequently (Butte *et al*., 2000; Chowdhury *et al*., 2010; Gupta *et al*., 2010; Steuer *et al*., 2002; Zhang *et al*., 2010). Chowdhury *et al*. show that combinatorially dysregulated subnetworks found through mutual information is predictive for cancer. In metabolomics analysis, using the combinations of metabolites and their levels, ADEMA (the algorithm for determining expected metabolite alterations) predicts expected changes of metabolite levels in a given metabolic network using mutual information (Cicek *et al*., 2013). In the context of MIRA, the two variables are (i) the genotype (e.g., case and control) and (ii) the combinations of the genes associated with a reaction and their expression levels. More specifically, MIRA performs the following tasks: (i) discretize gene expression levels using B-spline functions, (ii) calculate the mutual information between the genotype, as well as combinations of genes associated in a reaction and their discretized expression levels, and (iii) calculate the average mutual information for each metabolite using the mutual information of each neighboring reaction. Figure 2 shows the overall flow of the algorithm.

The advantages of MIRA over RA are as follows:

(1) Combinatorial mutual information works well when the sample sizes are small, and performs better than univariate significance testing.

(2) Unlike the RA, MIRA uses a multivariate method, analyzing multiple genes at a time. Therefore, it does not discard less insignificant changes unlike RA. MIRA uses measurements of individual samples instead of comparing sample means and is able to capture linear and non-linear dependencies among variables.

(3) Mutual information is bounded by zero and the minimum of the entropies of the two random variables. The most insignificant case is assigned the score zero; therefore, insignificant changes do not cancel out significant changes.

(4) MIRA has no bias toward highly connected metabolites, as it normalizes the sum of changes around a metabolite using the number of reactions instead of the square root of number of reactions as RA does [see Equations (2) and (8)].

Figure 1B shows the results obtained by MIRA for the introductory example shown in Figure 1A. Unlike the RA, which assigns a low score to the reaction *copragon transport via ABC system* indicating that there is no significant change on this enzyme, MIRA predicts relatively high mutual information indicating that when considered together genes *fhuB*, *fhuC* and *fhuD* are expressed differently. MIRA predicts similar results for both reactions and the average is found as 0.8 (max 0.98 in this test), whereas the RA assigns 0.54 to copragon (max ~13 in this test). The difference between the two algorithms stems from the fact that (i) MIRA performs a multivariate analysis compared with

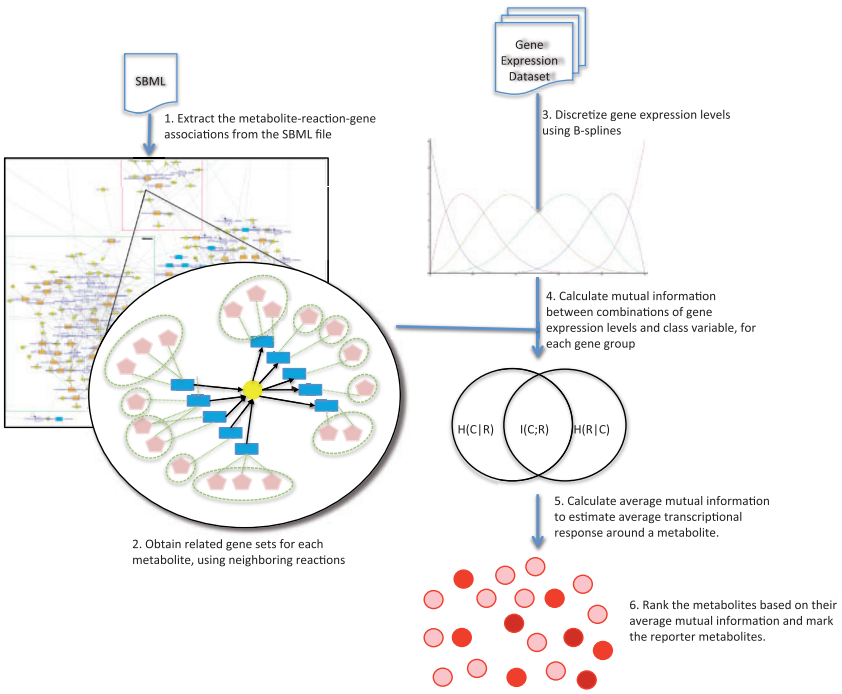

Fig. 2. Algorithm Flow. Rectangles represent reactions, pentagons represent genes and circles represent metabolites (darker red represents higher average mutual information). Algorithm starts by generating gene-reaction and reaction-metabolite associations out of the SBML file of the reconstructed metabolic network. Next, for each metabolite, gene sets are constructed based on their association with the neighboring reactions. As the third step, gene expression levels are discretized using B-splines. Fourth step consists of calculating the mutual information between the class variable and the discretized expression levels of groups of genes. After each metabolite is assigned average mutual information, based on the calculations done in Step 5, metabolites are ranked based on their average mutual information, and reporter metabolites are determined (darker red means it is a reporter metabolite)

the univariate analysis done by RA, (ii) given that there are only seven samples (3 ΔarcAΔfnr and 4 WT bacteria), mutual information is able to capture the change better than z-scores and (iii) being non-negative, mutual information does not cancel out significant effects.

To evaluate MIRA, we have analyzed six strains of *E.coli* with knockouts of transcriptional regulators in the oxygen response (ΔarcA, ΔappY, Δfnr, ΔoxyR, ΔsoxS and ΔarcAΔfnr) in aerobic and anaerobic conditions (Covert *et al.*, 2004). We have used the reconstructed metabolic network of *E.coli* (iAF1260; Feist *et al.*, 2007). We have also analyzed the autism gene expression dataset (Voineagu *et al.*, 2011) using the Recon 1 genome scale metabolic network for humans (Duarte *et al.*, 2007). For the *E.coli* dataset, we focused on the Δfnr knockout, which affects the switch between aerobic and anaerobic respiration. MIRA was able to successfully capture metabolites that were closely related to this enzyme and anaerobic respiration mechanism. For the autism dataset, MIRA predicted metabolites that have been recently discovered in the autism literature. We have also shown that MIRA has no bias toward hub metabolites unlike RA, and scoring scheme for MIRA is empirically significant.

2 MATERIALS AND METHODS

This section describes subcomponents and techniques MIRA uses to detect the hot spots in the metabolic network. Supplementary Table S1 describes the abbreviations and notations used throughout the section. Please see Supplementary Appendix A, Table S1, for a list of terms and variables and their explanations.

2.1 Constructing the network

In the context of our algorithm, the network is a hyper-graph G(V,E) where the vertex set V is the union of three entity types: metabolites, genes and reactions. The edge set E contains two types of edges: (i) edges that connect reactions to the associated genes whose expression lead to the corresponding enzyme and (ii) edges that connect metabolites to associated reactions based on producer/consumer relationships. Network information is obtained through the genome-scale reconstructed metabolic network of the organism to be analyzed using an SBML parser.

2.2 Discretizing gene expression data

To calculate mutual information between the genotype and the gene expression observations, one needs to calculate the probability of observing that profile. This is a well-studied problem in the literature (Silverman, 1986). There are three techniques that do not assume that the values come from a known distribution (which is the case for gene expression data). Kernel Density Estimation aims to measure the density of observations falling into a predetermined window using a kernel (Moon *et al.*, 1995), though it suffers from the high computational cost, and the results are dependent on the kernel length. Second technique is the histogram-based classification, which determines thresholds by dividing the domain of the variable into equal-sized chunks and classifying observations based on these thresholds. Despite the low computational cost, observations that are close to the thresholds are likely to be misclassified, as analytical methods associate an error term with each observation (Cakmak *et al.*, 2012; Cicek and Ozsoyoglu, 2012). To fix this shortcoming, B-spline functions have been used to associate probabilities with each bin determined by the histogram-based approach (Cicek *et al.*, 2013; Faith *et al.*, 2007). That is, each observation is associated with a probability to be in a bin, instead of making a binary decision to determine if it is in a given bin or not.

B-spline functions are defined by parameters M and k where M is the number of bins (chunks) that the domain is going to be divided into, and k, $1 \le k \le M$, is the number of bins an observation can be assigned to (e.g. $k = 1$ is equivalent to histogram-based binning). Based on M and k, each B-spline curve i, $1 \le i \le M$, is assigned a basis vector, the so-called knot vector t_i, defined as in Equation (3), and then the curve i is defined as a function of the neighboring curves, as shown in Equations (4) and (5).

$$t_i = \begin{cases} 0, & i < k \\ i - k + 1, & k \le i \le M \\ M - k + 2, & M < i \end{cases} \tag{3}$$

$$B_{i,1}(\S(\mathrm{gn})) = \begin{cases} 1, & t_i \le \S(\mathrm{gn}) < t_{i+1} \\ 0, & otherwise \end{cases} \tag{4}$$

$$B_{i,k}(\S(\mathrm{gn})) = B_{i,k-1}(\S(\mathrm{gn}))\left[\frac{\S(\mathrm{gn}) - t_i}{t_{i+k-1} - t_i}\right] + B_{i+1,k}((\S(\mathrm{gn}))\left[\frac{t_{i+k} - \S(\mathrm{gn})}{t_{i+k} - t_{i+1}}\right] \tag{5}$$

Each measurement (e.g. expression level for gene gn for person x) is assigned to a bin i with probability $B_{i,k}(\S(\mathrm{gn}))$, where §(gn) is a function that normalizes the value of the measurement for gn using the maximum and the minimum values observed for gn in the dataset.

2.3 Calculation of mutual information

After expression values are calculated, mutual information is calculated. Given two random variables X and Y, mutual information $I(X;Y)$ measures how much knowing one reduces the uncertainty about the other.

In the context of MIRA, the first variable is the binary class variable C [e.g. control versus variable) and the second variable $B(G(r))$ is the *binned* measurements of a group of genes $G(r)$ that are associated with a reaction r. For instance, for the example shown in Figure 2B, genes *fhuE* and *tonB* are grouped based on their association with the reaction *coprogen transport via ton system (ctts)*. Assuming we use 2 bins (up and down) then, $C = \{WT, \Delta arcA\Delta fnr\}$ and $B(G(ctts)) = \{$*fhuE up & tonB up, fhuE up & tonB down, fhuE down & tonB up, fhuE down & tonB down*$\}$. $I(C;B(G(ctts)))$ is found to be 0.808.

The goal of calculating $I(C;B(G(r)))$ is to learn how much knowing discretized gene expression levels of related genes reduces the uncertainty on the genotype. In other words, we find out whether a reaction r and its corresponding genes are predictive on the genotype. If the expression levels of these genes (when considered together) are different with respect to the class variable, then we obtain a high mutual information value. Equation (6) specifies the mutual information formula.

$$I(C; B(G(r))) = \sum_{bg \in B(G(r))} \sum_{c \in C} p(bg, c) * \lg\left(\frac{p(bg, c)}{p(bg)p(c)}\right) \tag{6}$$

$p(bg)$ is calculated as shown in Equation (7) and k is a constant input to the algorithm. That is, given a dataset D, probability of observing g (e.g. *fhuE up* and *tonB up*) is the multiplication of spline values for each gene (e.g. *fhuE* and *tonB*) to be in the corresponding bins (e.g. *up* and *up* in this case), summed and averaged over all individuals in D. This calculation assumes that gene expression values are independent of each other. $p(bg,c)$ is calculated similarly and $p(c)$ is constant in the dataset.

$$p(bg) = \frac{\sum_D \prod_{gn \in G(r)} B_{bin\,of\,gn\,bg,k}(\S(\mathrm{gn}))}{|\mathrm{D}|} \tag{7}$$

We define the aggregate transcriptional regulation around a metabolite m as the *average mutual information* of the consumer and producer reactions of m. Given that $R(m)$ is the set of neighboring reactions, then average mutual information I_m for metabolite m is defined as in Equation (8).

$$I_m = \frac{1}{|R(m)|}\sum_{r\in R(m)} I(C; B(G(r))) \quad (8)$$

MIRA fits a beta distribution to I_m values given the interval. Then, $I_m \sim \text{Beta}(\alpha, \beta)$ where α and β are learned from the sample. Finally, all metabolites with $P<0.05$ are picked as reporter metabolites.

2.4 Datasets and experimental design

To test the performance of MIRA and compare it with RA, we considered two resources. First, we used mRNA expression profiles of six *E.coli* strains with knockouts of transcriptional regulators of the oxygen response (ΔarcA, ΔappY, Δfnr, ΔoxyR, ΔsoxS and ΔarcAΔfnr) published and released in Covert *et al.* (2004). For each strain they obtained measurements in aerobic and anaerobic conditions. Consequently, we used 12 datasets for the knockouts and 12 for the WT. We compared the expression profiles of selected knockouts against the control, to obtain reporter metabolites using each respective method. We did not perform cross-condition comparisons. For instance, we compared knockout measurements under aerobic conditions with WT measurements under aerobic conditions only. Second, we ran tests on the autism spectrum disorder (ASD) brain gene expression dataset (Voineagu *et al.*, 2011). The dataset contains gene expression levels for 8858 genes for 58 human cortex samples (29 ASD and 29 controls).

We implemented MIRA in C# language using ET Framework 4.0. Tests were run on a Dell PowerEdge R710 Server with two Intel® Xeon® quad processors and 48 GB main memory. The server runs on Windows Server 2008 operating system.

For *E.coli* knockout tests, we used the genome scale-reconstructed metabolic network model of *E.coli*, iAF1260 (Feist *et al.*, 2007). The model contains 1972 metabolites, 2382 reactions, 1261 genes and 3 compartments. We mapped the measured genes to the corresponding reactions in the metabolic network, as annotated in the model. After the mapping, we obtained 1643 metabolites, to which there was at least one reaction associated with a gene measured in the dataset. Only the obtained 1643 metabolites were considered in the tests. For the Autism dataset, we used the Recon1 genome-scale metabolic network for humans (Duarte *et al.*, 2007). The model consists of 3188 metabolites, 3742 reactions, 1499 genes and 8 compartments. Mapping the genes measured to the network resulted in a metabolite set of size 2331 for further consideration. For both datasets, we used $M = 6$ and $k = 4$ to discretize measurements using B-splines. Please see Supplementary Appendix B for time requirements.

3 RESULTS

3.1 Comparing reporter metabolite sets of MIRA and RA

To compare the performances of MIRA and RA, we obtained reporter metabolites using both algorithms. First, we investigated how similar two sets are using the Jaccard distance (JD). JD is complementary to the Jaccard index, which is the ratio of shared items between the two sets, A and B, and the union of the items in two sets. Jaccard index is denoted as $J(A,B)$. JD, $J_\delta(A,B)$, is equivalent to $1 - J(A,B)$. Two algorithms yield different sets of reporter metabolites for *E.coli* knockouts. For the knockouts ΔappY (anaerobic) and ΔOxyR (aerobic), the sets of reporter metabolites are totally distinct (JD = 1) and the smallest JD obtained in these tests is 0.85. For the Autism dataset, JD is 0.9.

3.2 Robustness against hub metabolites

Metabolic networks are known to be scale free (Albert, 2005), that is, their degree distribution follows the power law. Our results show that the reporter metabolites found by RA are usually highly connected metabolites known as hubs or common metabolites, which are rare in scale-free networks. For instance, in the test for Δfnr (aerobic), RA marks H+, H_2O, ATP, ADP, Phosphate, Diphosphate and CO_2 as the top 7 reporter metabolites, which participate in many reactions and are associated with many genes in the metabolic network. To test if RA has a bias toward hub metabolites, we calculated the average gene connectivity and reaction connectivity of the reporter metabolites for each algorithm. We also formed a random set of metabolites where the size of the random set is randomly chosen as a number between the sizes of reporter sets produced by MIRA and RA. We repeated the random set selection 10 000 times and found the average connectivity for the random set. Figure 3A shows the average gene connectivity, and Figure 3B shows the average reaction connectivity of the reporter metabolites for MIRA, RA and the random set for *E.coli* knockout tests. Results show that RA favors highly connected metabolites, whereas reporter metabolites found by MIRA have a closer degree distribution to the random set, and does not have such a bias. This also applies to reporter metabolites found in the Autism dataset by RA (Supplementary Table S5). The intuitive reason for this difference is that MIRA averages the mutual information found for each reaction, whereas RA divides the sum by the square root of the number of reactions around a metabolite.

3.3 Interpreting reporter metabolites for Δfnr (anaerobic) dataset

When *E.coli* has no O_2 as the final electron acceptor, it switches to anaerobic respiration and uses electron donating dehydrogenases and accepting reductases on the membrane. Fumarate Nitrate Reductase (fnr) is a transcriptional regulator that regulates 100+ genes and nitrate/fumarate reduction in response to the switch from aerobic to anaerobic respiration in *E.coli*. Fnr has an iron-sulfur cluster [4Fe–4S] that senses the presence of oxygen and becomes inactivated when oxidized in the presence of oxygen. It can also be converted into a disulfide form by glutathione or thioredoxin when inactive (Daruwala and Meganathan, 1991).

Figure 4 shows a simplified depiction of the dynamics in anaerobic respiration with respect to fnr (Keseler *et al.*, 2013; MetaCyc; Unden and Bongaerts, 1997; UniProt). Although there are many types of dehydrogenases and reductases, we draw them in two groups for the sake of simplicity: (i) hydrophilic side toward periplasm and (ii) hydrophilic side toward cytoplasm. Electron donors like G3P, formate, lactate, NADH, H_2 are oxidized by the hydrogenases, and electrons are transported using menaquinone to the reductases. Focusing on nitrate reductase, this electron is transferred to protoheme, then to Fe-S cluster and, finally, to molybdenum (Mo). It is used to reduce nitrate to nitrite. Fnr stimulates the expression of this enzyme. In the presence of O_2, fnr is oxidized and inactivated. As stated above, glutathione and thioredoxin act as electron donors to activate fnr.

In Supplementary Appendix C, Table S2 lists the reporter metabolites found by MIRA and Table S3 lists reporter metabolites

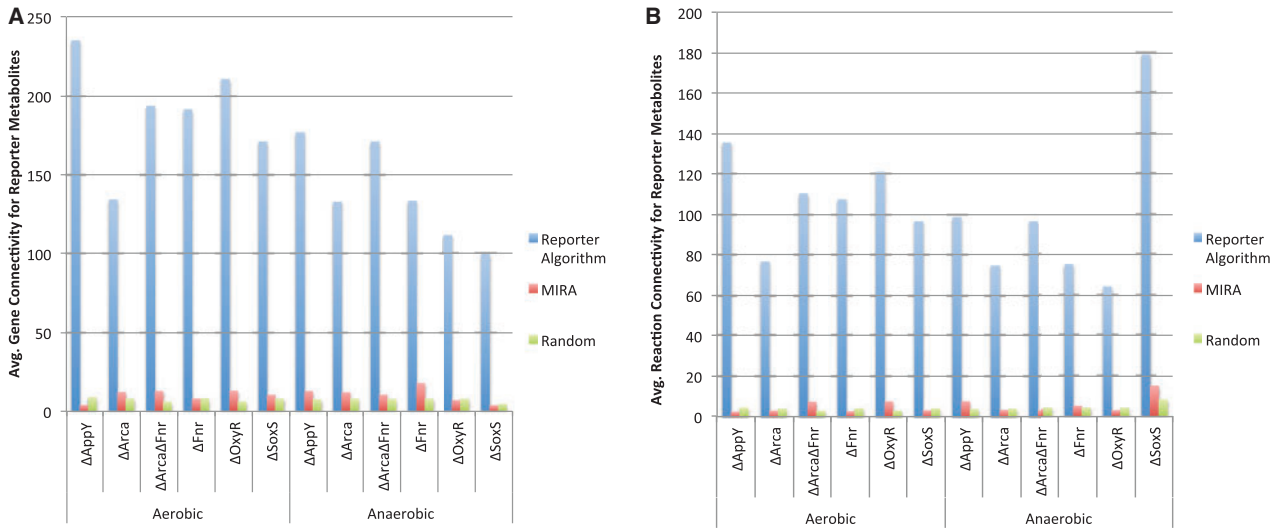

Fig. 3. Application of Average gene and reaction connectivity of the reporter metabolites. Panel (**A**) shows average number of genes associated with the reporter metabolites found by RA and MIRA in *E.coli* knockout tests. Random set reports the average number of genes connected randomly chosen metabolites of the same size as the original sets (repeated 10 000 times and averaged). Panel (**B**) shows the same results for the average number of reactions associated with metabolites

found by RA based on Δfnr knockout under anaerobic condition. Previous description is based on the data obtained from UniProt Database, which summarizes the key concepts and metabolites related to fnr. The metabolites reported by MIRA highly overlap with this definition. The second reporter metabolite protoheme (heme b) is an important metabolite in the heme-biosynthesis pathway, and is the first electron acceptor in nitrate reductase complex (Metacyc). As mentioned above, glutathione and thioredoxin are key agents to convert the enzyme, and MIRA detects them as reporter metabolites [glutaredoxin (reduced/oxidized) and thioredoxin (reduced/oxidized)]. Nitrite is one of the direct products of nitrate reductase, and it is also reported as a reporter metabolite.

Aside from the metabolites in UniProt's definition, MIRA found dimethyl sulfide/sulfoxide (DMSO) and trimethylamine/trimethylamine n-oxide (TMAO) as reporter metabolites. As Figure 4 shows, *E.coli* uses *N*- and *S*- oxides as the terminal electron acceptors (Daruwala and Meganathan, 1991).

When RA is considered, the top metabolite picked by MIRA is not a reporter metabolite in RA's list. Tungstate is known as a direct inhibitor of nitrate reductase as well as TMAO and DMSO reductases, which are also reporter metabolites (Prins *et al.*, 1980). Similarly, *tripeptide murein units* [short name for *two linked disacharide tripeptide murein units (uncrosslinked middle of chain)*] is also not picked by RA and might seem unrelated at first. Murein units constitute bacterial cell walls. Membrane-bound lytic murein transglycosylase is the enzyme that degrades mureins, and the gene responsible to transcribe this enzyme is dniR (mltD). A mutant of this enzyme is known to be defective in producing nitrite reductase. Nitrate and nitrite also stimulate the expression of dniR (Kajie *et al.*, 1991).

Most striking difference in the reporter metabolites found by RA is that the first seven metabolites are H+, H2O, ADP, P, ATP, NAD, NADH (NADP, NADPH, O_2, CDP, UDP, dADP, dCDP, GDP, GTP, CoA are also listed). These are highly connected metabolites, which take place in many biochemical activities in the cell, and therefore, it is hard to link them to the effect of the knockout in this test. MIRA puts these metabolites within (695 964] in the ranking of 1643 metabolites considered. Also RA lists many metabolites from the central metabolism: Glycerophosphoglycerol, pyruvate, glycerol, F6P, succinate d-glucose and acetyl-CoA. Similarly, these metabolites can be considered as general owing to the diverse functionality of the pathway they are in. Having that said, RA detects some metabolites that are not picked by MIRA and are relevant. For instance, sulfate and molybdate are incorporated into fnr and nitrate reductase (Tavares *et al.*, 2006). Menaquinone 8, menaquinol 8 and 2-demethylmenaquinol 8, link dehydrogenases and reductases/oxidases in the electron transport chains, and hence are directly related to fnr enzyme as shown in Figure 4 (Unden and Bongaerts, 1997). Although ubiquinone-8/ubiquinol-8 play an important role in aerobic respiration, they are listed higher than menaquinone 8/menaquinol 8.

In conclusion, the results suggest that (i) MIRA has no bias toward hub metabolites, and successfully downplays their importance, and (ii) MIRA yields reporter metabolites, which are in close proximity to the enzyme and are relevant with respect to the literature.

3.4 Interpreting reporter metabolites for the autism dataset

ASD is a developmental genetic disorder that causes social interaction abnormalities, communication deficiencies and repetitive behavior. Although the disease has a genetic origin, metabolic implications have been studied widely in the literature (Boccuto *et al.*, 2013; Emond *et al.*, 2013; Yap *et al.*, 2010). Running MIRA on the gene expression dataset provided by Voineagu *et al.*, 2011, has resulted in 52 reporter metabolites as shown in Supplementary Appendix C, Table S4. Reporter metabolites found by RA is listed in Supplementary Appendix C, Table S5.

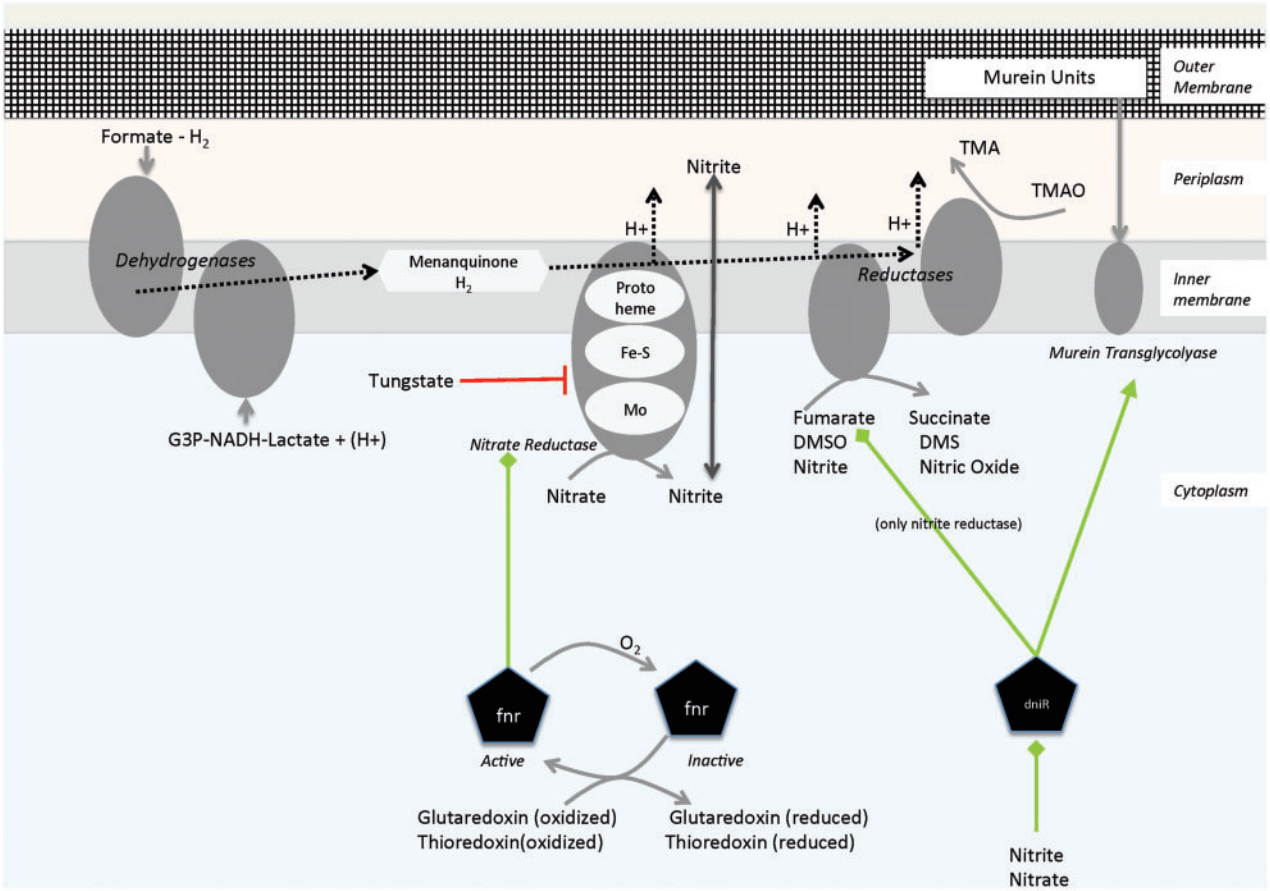

Fig. 4. Anaerobic respiration of *E.coli* with respect to fnr and reporter metabolites found by MIRA. Anaerobic respiration of *E.coli* couples electron donors to electron acceptors via dehydrogenases and reductases on the inner membrane. There are many types of dehydrogenases and reductases; however, only two types are shown: hydrophilic side toward cytosol and toward periplasm. Sizes and shapes of the proteins are not drawn to scale. Only nitrate reductase is shown in more detail and separately. Formate, lactate, NADH, H_2 G3P are electron donors and lead to reduction of acceptors like nitrate, nitrite, DMSO, TMSO and fumarate. Menanquinone acts as a mediator between dehydrogenases and reductases. In the case of nitrate reductase, electron is transferred through protoheme, Fe-S cluster and Molybdenum to reduce nitrate to nitrate. Fnr activates this enzyme, and tungstate is a well-known inhibitor. Fnr is inactivated by oxygen and can be reactivated by agents like glutaredoxin and thioredoxin. Murein units constitute the cell wall, and the enzyme that degrades murein units is transcribed by dniR gene. It is known that dniR regulates nitrite reductase and it is stimulated but nitrite and nitrite

The first and the eighth metabolites located as reporter by MIRA are protein-linked serine/threonine residue and protein-linked asparagine residue at glycosylation sites. Glycosylation is a process that attaches glycans to proteins as a post-translational modification in the secretion pathway. Secretion pathway is known as an important factor in brain development, and DIA1R mutation leads to ASD and mental retardation (Aziz *et al.*, 2011). More specifically, it is reported in the literature that 7 glycosylation-related genes are affected by copy number variations (CNVs) in autism patients (van der Zwaag *et al.*, 2009; Pinto *et al.*, 2010). We also observe glycans such as glycophosphatidylinositol later in the reporter list.

Glucuronidation is an important process for detoxification of most xenobiotics by making such components more water-soluble and less toxic by attaching glucuronate to the substrates (Stein *et al.*, 2011). The second metabolite d-glucorate and the third metabolite d-glucurono-6,3-lactone (d-glucurone), located by MIRA, are direct precursors of glucuronate. Stein *et al.* (2011) reports evidence on lower glucuronidation levels in children ASD, which may explain d-glucurono-6,3-lactone being a stress point in the metabolic network. L-Arabinose is also in this pathway and is located as a reporter by MIRA.

Histone n6-methyl-l-lysine, protein n6,n6-dimethyl-l-lysine, peptdyl-l-lysine and protein n6,n6,n6-trimethyl-l-lysine, all located by MIRA, belong to lysine degradation pathway. These metabolites are centered on the path that consumes lysine and produces carnitine. Celestino-Soper *et al.* have revealed that dysregulation of carnitine metabolism may be important in non-dysmorphic autism. The results show that a deletion in TMLHE gene on this pathway has a significant correlation with ASD (Celestino-Soper *et al.*, 2012). In a recent work, Frye *et al.* (2013) links the abnormalities in acyl-carnitine levels and autism.

Lipoylprotein, lipoamide, dihydrolipolprotein and dihydrolipoamide, all located by MIRA, are four metabolites that are in

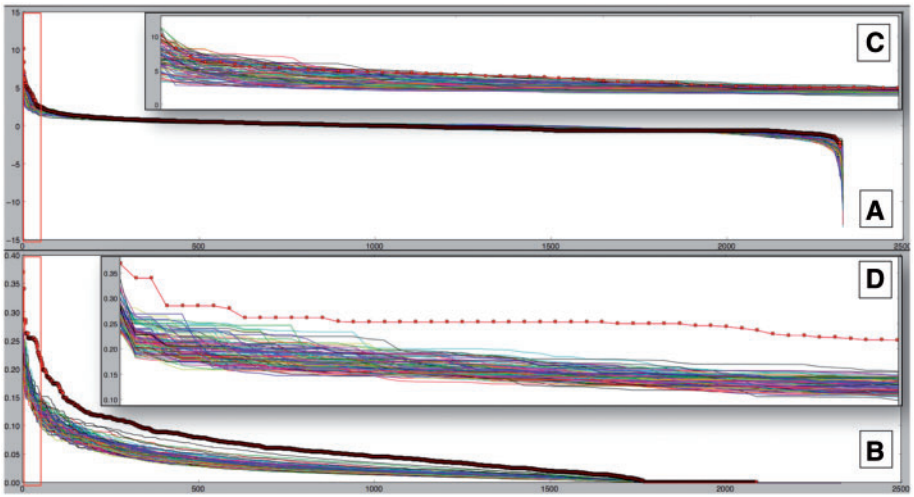

Fig. 5. Empirical testing of the significance of the scoring schemes. Panel (**A**) shows series of z-scores obtained for each random set, and for the original dataset (shown as the red line with large circles) by RA. Scores are sorted in descending order. Panel (**B**) shows the $I_{mr}s$ for the same data obtained by MIRA. Panels (**C**) and (**D**) show close-ups for the first 50 metabolites for Panels (A) and (B), respectively

the glycine cleavage pathway. MIRA reports four metabolites out of six in this pathway and points to an alteration in this process. A recent work by Yu *et al.* (2013) confirms that a mutation in AMT gene is also associated with ASD. This mutation leads to a deficiency in glycine cleavage system. In addition to this, the metabolic profiling done by Yap *et al.* (2010) shows significant differences in glycine levels in the urine of autistic children.

Starting with stearidonyl coenzyme A, MIRA detects 17 intermediates of fatty acid synthesis pathway as reporter metabolites. Fatty acid metabolism and autism have been associated in the literature. Richardson and Ross point to the growing evidence on the relation between neurodegenerative diseases and fatty acid abnormalities (Richardson and Ross, 2000). Among more recent works, Tamiji and Crawford state that children with autism show higher rates of lipid metabolism than controls (Tamiji and Crawford, 2010). El-Ansary *et al.* (2011) report increase in most of the saturated fatty acids in a cohort of 52 autism patients.

In comparison, none of the aforementioned reporter metabolites are located by RA, but common metabolites are highly ranked as in *E.coli* tests.

In summary, reporter metabolites picked by MIRA for autism are backed by literature. Results show that, although some predictions (e.g. glycine cleavage deficiency) are not obvious targets, the MIRA method was able to predict them. The literature on the relation between autism and (i) glucuronation (Stein *et al.*, 2011), (ii) lysine degradation and carnitine metabolism (Celestino-Soper *et al.*, 2012; Frye *et al.*, 2013) and (iii) glycine cleavage (Yu *et al.*, 2013) did not exist at the time of the gene expression dataset was published. Hence, MIRA shows promising prospect for discovering new metabolic targets.

3.5 Empirically testing the significance of scoring schemes used by MIRA and RA

To assess the significance of the scores calculated by both algorithms and assess the reliability of the rankings, we used an empirical significance testing using the following method. We (i) shuffled the labels of the individuals in the autism dataset 100 times to obtain 100 random datasets, (ii) ran MIRA and RA on these datasets, as well as on the original data, and (iii) sorted and plotted the scores in descending order for all 101 instances. Figure 5A shows the scores for RA, and Figure 5B shows the results for MIRA. Big red circles represent the results found on the original dataset. Figures 5C and 5D show a close-up for the first 50 metabolites in the ranking. MIRA's results for the original dataset dominate the random curves and, hence, suggest an empirically significant result. On the other hand, RA's output for original data follows a similar pattern with the random datasets.

4 CONCLUSION

Metabolic networks have received significant attention in the past decade. This advancement has led to the investigation of various genetic diseases and their metabolism with the use of the reconstructed genome scale metabolic networks. One application is to find the regulatory architecture of the metabolic network using the underlying transcriptome. The RA finds the metabolic hot spots around which transcriptional regulation is centered. In this article, we developed a novel method, called MIRA, that uses a combinatorial approach and mutual information to find reporter metabolites. Our approach addresses the shortcomings of the existing RA algorithm. More specifically, it is robust against small sample sizes, uses a multivariate approach instead of a univariate one and does not cancel out significant changes in expression levels with non-significant ones. Our results show that (i) MIRA has no bias on picking hub metabolites as reporter metabolites, (ii) reporter metabolites found by MIRA are biologically sound and are supported by literature even for a complex disease like Autism, (iii) MIRA captures the effects of a knockout of the Fnr gene in *E.coli* successfully and (iv) MIRA provides empirically significant results, which supports the fact that it captures the underlying biological phenomenon.

Funding: This research has been supported by the National Science Foundation grants DBI 0743705, DBI 0849956, CRI 0551603 and by the National Institute of Health grant GM088823. A. Ercument Cicek has also been supported by Ray and Stephanie Lane Fellowship.

Conflict of Interest: none declared.

REFERENCES

Agren,R. *et al.* (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput. Biol.*, **8**, e1002518.

Albert,R. (2005) Scale-free networks in cell biology. *J. Cell Sci.*, **118**, 4947–4957.

Aziz,A. *et al.* (2011) DIA1R is an X-linked gene related to deleted in autism-1. *PLoS One*, **6**, e14547.

Boccuto,L. *et al.* (2013) Decreased tryptophan metabolism in patients with autism spectrum disorders. *Mol. Autism*, **4**, 16.

Brosché,M. *et al.* (2005) Gene expression and metabolite profiling of Populus euphratica growing in the Negev desert. *Genome Biol.*, **6**, R101.

Butte,A.J. and Kohane,I.S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, **5**, 418–429.

Cakir,T. *et al.* (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. *Mol. Syst. Biol.*, **2**, 50.

Cakmak,A. *et al.* (2012) A new metabolomics analysis technique: steady state metabolic network dynamics analysis. *J. Bioinform. Comput. Biol.*, **10**, 1240003.

Carrari,F. *et al.* (2006) Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior. *Plant Physiol.*, **142**, 1380–1396.

Caspi,R. *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **42**, D459–D471.

Celestino-Soper,P.B.S. *et al.* (2012) A common X-linked inborn error of carnitine biosynthesis may be a risk factor for nondysmorphic autism. *Proc. Natl Acad. Sci. USA*, **109**, 7974–7981.

Chowdhury,S.A. *et al.* (2010) Subnetwork state functions define dysregulated subnetworks in cancer. *J. Comput. Biol.*, **18**, 263–281.

Chuang,H-Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Cicek,A.E. and Ozsoyoglu,G. (2012) Observation conflict resolution in steady state metabolic network dynamics analysis. *J. Bioinform. Comput. Biol.*, **10**, 1240004.

Cicek,A.E. *et al.* (2013) ADEMA: an algorithm to determine expected metabolite level alterations using mutual information. *PLoS Comput. Biol.*, **9**, e1002859.

Cimini,D. *et al.* (2009) Global transcriptional response of *Saccharomyces cerevisiae* to the deletion of SDH3. *BMC Syst. Biol.*, **3**, 17.

Covert,M.W. and Palsson,B.Ø. (2002a) Constraints-based models: regulation of gene expression reduces the steady-state solution space. *J. Theor. Biol.*, **221**, 309–325.

Covert,M.W. and Palsson,B.Ø. (2002b) Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. Biol. Chem.*, **277**, 28058–28064.

Covert,M.W. *et al.* (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, **429**, 92–96.

Daruwala,R. and Meganathan,R. (1991) Dimethyl sulfoxide reductase is not required for trimethylamine N-oxide reduction in *Escherichia coli*. *FEMS Microbiol. Lett.*, **83**, 255–259.

Daub,C.O. *et al.* (2004) Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, **5**, 118.

David,H. *et al.* (2006) Metabolic network driven analysis of genome-wide transcription data from *Aspergillus nidulans*. *Genome Biol.*, **7**, R108.

David,H. *et al.* (2008) Analysis of Aspergillus nidulans metabolism at the genome-scale. *BMC Genomics*, **9**, 163.

Deo,R.C. *et al.* (2010) Interpreting metabolomic profiles using unbiased pathway models. *PLoS Comput. Biol.*, **6**, e1000692.

Dinu,I. *et al.* (2007) Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, **8**, 242.

Draghici,S. *et al.* (2007) A systems biology approach for pathway level analysis. *Genome Res.*, **17**, 1537–1545.

Duarte,N.C. *et al.* (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl Acad. Sci. USA*, **104**, 1777–1782.

Emond,P. *et al.* (2013) GC-MS-based urine metabolic profiling of autism spectrum disorders. *Anal. Bioanal. Chem.*, **405**, 5291–5300.

El-Ansary,A.K. *et al.* (2011) Plasma fatty acids as diagnostic markers in autistic patients from Saudi Arabia. *Lipids Health Dis.*, **10**, 62.

Faith,J.J. *et al.* (2007) Large-scale mapping and validation of *escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.

Feist,A.M. *et al.* (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 121.

Ferrara,C.T. *et al.* (2008) Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. *PLoS Genet.*, **4**, e1000034.

Frye,R.E. *et al.* (2013) Unique acyl-carnitine profiles are potential biomarkers for acquired mitochondrial disease in autism spectrum disorder. *Transl. Psychiatry.*, **3**, e220.

Gerstein,M. and Jansen,R. (2000) The current excitement in bioinformatics—analysis of whole-genome expression data: how does it relate to protein structure and function? *Curr. Opin. Struct. Biol.*, **10**, 574–584.

Goh,K.I. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Gupta,N. and Aggarwal,S. (2010) MIB: using mutual information for biclustering high dimensional data. *Pattern Recognit.*, **43**, 2692–2697.

Hancock,T. *et al.* (2012) Identifying neighborhoods of coordinated gene expression and metabolite profiles. *PLoS One*, **7**, e31345.

Holm,A.K. *et al.* (2010) Metabolic and transcriptional response to cofactor perturbations in *Escherichia coli*. *J. Biol. Chem.*, **285**, 17498–17506.

Hughes,T.R. and et al. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Ideker,T. and 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**), 233–240.

Ihmels,J. *et al.* (2004) Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat. Biotechnol.*, **22**, 86–92.

Jans,A. *et al.* (2011) Transcriptional metabolic inflexibility in skeletal muscle among individuals with increasing insulin resistance. *Obesity*, **19**, 2158–2166.

Kajie,S.I. *et al.* (1991) Molecular cloning and DNA sequence of dniR, a gene affecting anaerobic expression of the *Escherichia coli* hexaheme nitrite reductase. *FEMS Microbiol. Lett.*, **83**, 205–211.

Karp,P.D. *et al.* (2002) The pathway tools software. *Bioinformatics*, **18** (**Suppl 1**), S225–S232.

Keseler,I.M. *et al.* (2013) EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.*, **41**, 605–612.

Kharchenko,P. *et al.* (2005) Expression dynamics of a cellular metabolic network. *Mol. Syst. Biol.*, **1,** 2005.0016.

Ma'ayan,A. (2008) Network integration and graph analysis in mammalian molecular systems biology. *IET Syst. Biol.*, **2**, 206–221.

Moon,Y.I. *et al.* (1995) Estimation of mutual information using kernel density estimators. *Phys. Rev. E*, **52**, 2318–2321.

Nam,H. *et al.* (2009) Computational identification of altered metabolism using gene expression and metabolic pathways. *Biotechnol. Bioeng.*, **103**, 835–843.

Oliveira,A.P. *et al.* (2008) Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. *BMC Syst. Biol.*, **2**, 17.

Patil,K.R. and Nielsen,J. (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl Acad. Sci. USA*, **102**, 2685–2689.

Pavlidis,P. *et al.* (2002) Exploring gene expression data with class scores. *Pac. Symp. Biocomput.*, **2002**, 474–485.

Pinto,D. *et al.* (2010) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**, 368–372.

Prins,R.A. *et al.* (1980) Inhibition of nitrate reduction in some rumen bacteria by tungstate. *Appl. Environ. Microbiol.*, **40**, 163.

Rhodes,D.R. and Chinnaiyan,A.M. (2005) Integrative analysis of the cancer transcriptome. *Nat. Genet.*, **37**, S31–S37.

Richardson,A.J. and Ross,M.A. (2000) Fatty acid metabolism in neurodevelopmental disorder: a new perspective on associations between attention-deficit/

hyperactivity disorder, dyslexia, dyspraxia and the autistic spectrum. *Prostaglandins Leukot Essent. Fatty Acids*, **63**, 1–9.
Schramm,G. *et al.* (2010) Analyzing the regulation of metabolic pathways in human breast cancer. *BMC Med. Genomics*, **3**, 39.
Seshasayee,A.S. *et al.* (2009) Principles of transcriptional regulation and evolution of the metabolic system in *E. coli*. *Genome Res.*, **19**, 79–91.
Shlomi,T. *et al.* (2008) Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, **26**, 1003–1010.
Silverman,B.W. (1986) *Density Estimation For Statistics And Data Analysis*. Chapman and Hall, London.
Stein,T.P. *et al.* (2011) Autism and phthalate metabolite glucuronidation. *J. Autism Dev. Disord.*, **43**, 2677–2685.
Steuer,R. *et al.* (2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, **18** (**Suppl 2**), 231–240.
Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
Tanay,A. *et al.* (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci USA*, **101**, 2981–2986.
Tamiji,J. and Crawford,D.A. (2010) The neurobiology of lipid metabolism in autism spectrum disorders. *Neurosignals*, **18**, 98–112.
Tavares,P. *et al.* (2006) Metalloenzymes of the denitrification pathway. *J. Inorg. Biochem.*, **100**, 2087–2100.
Ulitsky,I. and Shamir,R. (2009) Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics*, **25**, 1158–1164.
Unden,G. and Bongaerts,J. (1997) Alternative respiratory pathways of *Escherichia coli*: energetics and transcriptional regulation in response to electron acceptors. *Biochim. Biophys. Acta*, **1320**, 217–234.
The UniProt Consortium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.
Usaite,R. *et al.* (2009) Reconstruction of the yeast Snf1 kinase regulatory network reveals its role as a global energy regulator. *Molecular Syst. Biol.*, **5**, 319.
van der Zwaag,B. *et al.* (2009) Gene-network analysis identifies susceptibility genes related to glycobiology in autism. *PLoS One*, **4**, e5324.
Venelli,A. (2010) Efficient entropy estimation for mutual information analysis using B-splines. *Lect. Notes Comput. Sci.*, **6033**, 17–30.
Voineagu,I. *et al.* (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, **474**, 380–384.
Vongsangnak,W. *et al.* (2009) Genome-wide analysis of maltose utilization and regulation in aspergilla. *Microbiology*, **155**, 3893–3902.
Yap,I.K. *et al.* (2010) Urinary metabolic phenotyping differentiates children with autism from their unaffected siblings and age-matched controls. *J. Proteome Res.*, **9**, 2996–3004.
Yeang,C.H. *et al.* (2006) A joint model of regulatory and metabolic networks. *BMC Bioinformatics*, **7**, 332.
Yu,T.W. *et al.* (2013) Using whole-exome sequencing to identify inherited causes of autism. *Neuron*, **77**, 259–273.
Zelezniak,A. *et al.* (2010) Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes. *PLoS Comput. Biol.*, **6**, e1000729.
Zhang,H. *et al.* (2010) MIClique: an algorithm to identify differentially coexpressed disease gene subset from microarray data. *Biomed. Res. Int.*, **2009**, 642524.