BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 16 2007, pages 2139–2146
doi:10.1093/bioinformatics/btm319

Systems biology

Constraint-based functional similarity of metabolic genes: going beyond network topology[†]

Oleg Rokhlenko[1,*], Tomer Shlomi[2], Roded Sharan[2], Eytan Ruppin[2,3] and Ron Y. Pinter[1]

[1]Department of Computer Science, Technion, Haifa 32000, [2]School of Computer Science and [3]School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel

Received on November 21, 2006; revised and accepted on June 7, 2007

Advance Access publication June 22, 2007

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Several recent studies attempted to establish measures for the similarity between genes that are based on the topological properties of metabolic networks. However, these approaches offer only a static description of the properties of interest and offer moderate (albeit significant) correlations with pertinent experimental data.

Results: Using a constraint-based large-scale metabolic model, we present two effectively computable measures of functional gene similarity, one based on the response of the metabolic network to gene knockouts and the other based on the metabolic flux activity across a variety of growth media. We applied these measures to 750 genes comprising the metabolic network of the budding yeast. Comparing the *in silico* computed functional similarities to Gene Ontology (GO) annotations and gene expression data, we show that our computational method captures functional similarities between metabolic genes that go beyond those obtained by the topological analysis of metabolic networks alone, thus revealing dynamic characteristics of gene function. Interestingly, the measure based on the network response to different growth environments markedly outperforms the measure based on its response to gene knockouts, though both have some added synergistic value in depicting the functional relationships between metabolic genes.

Contact: olegro@cs.technion.ac.il

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

1 INTRODUCTION

Recent advances in systems biology have resulted in the reconstruction of several types of genome-scale biochemical networks—metabolic, regulatory, signaling, protein–protein interaction and more. The availability of these interaction networks in turn has stimulated the analysis of the structural, i.e. topological, properties to gain insights to the functionality of their genes. Even though recent analyses have provided valuable insights regarding this issue (Jeong *et al*., 2000; Ravasz *et al*., 2002), topological characteristics alone (as devised by e.g. Chen and Vitkup, 2006; Kharchenko *et al*., 2005) offer only a static description of the properties of interest. On the other hand, accurate prediction of dynamic cell activity using kinetic models requires detailed information on the rates of enzyme activity which is rarely available; moreover, such analysis is usually limited to small-scale networks.

Fortunately, for metabolic networks, the use of stochiometry and other sources of information can provide an added value over the topology of the underlying structure. Specifically, constraint-based models (CBMs) have emerged as a key method for studying such networks, permitting the large-scale analysis thereof. CBMs use genome-scale networks to predict steady-state metabolic activity, regardless of specific enzyme kinetics. In these models, stoichiometric, thermodynamic, flux capacity and possibly other constraints affect the space of attainable flux distributions.

In this article we employ constraint-based modeling to devise two effectively computable functional similarity measures between genes. The two measures employ large-scale *in silico* experiments, based on flux balance analysis (FBA), that can be further validated *in vitro*. Our first measure, the genetic response similarity (GRS) measure, is based on the similarity in metabolic network response to gene knockouts. The second measure, environmental response similarity (ERS), is based on similarity in the metabolic network activity across an array of various growth environments. These two measures reveal two complementary ways of defining the relation between gene *u* with its surrounding: the GRS measure defines the effect of gene *u* on its surroundings, whereas the ERS measure defines the effect of the surroundings on gene *u*.

To assess the veracity of the suggested measures, we validate them based on various biological data sources, including Gene Ontology (GO), phylogenetic profiling and gene expression measurements. The basic relation between metabolic fluxes and gene expression was previously studied and established both computationally (showing only a moderate correlation) as well as experimentally. Several studies (Bilu *et al*., 2006, Famili *et al*., 2003; Reed and Palsson, 2004; Schuster *et al*., 1999, 2002) have shown that the expression patterns of enzyme coding genes are correlated with the flux patterns predicted by FBA. In this work we extend these studies to look into ways of building upon the reported correlation between fluxes and expression, to construct efficient measures of functional similarity among metabolic genes. To this end, in contrast with

[*]To whom correspondence should be addressed.

[†]A preliminary version of this article appeared as Rokhlenko *et al*., (2006).

the previous studies, we examine the relation between fluxes and expression while concomitantly controlling for correlations caused solely by the network's topology.

Our comparison focuses on 750 metabolic genes of the yeast *Saccharomyces cerevisiae*. We find that the ERS measure outperforms topological, conservation-based and expression-based measures when testing for similarity with GO. Moreover, for many GO terms it is the only measure that succeeds to provide a significant result. On the other hand, the GRS measure shows only moderate results with only a few unique successes. We also find the correlation between model-based measures and co-expression to be statistically significant. However, we find GRS to be only moderately correlated with experimental data, whereas ERS exhibits a strong and significant correlation. Furthermore, this correlation remains so even after canceling the effect of the underlying (static) network topology. These results support the notion that a model-based ERS measure indeed captures the true functional similarity between metabolic genes.

2 METHODS

2.1 Modeling metabolism

Constraint-based modeling allows a steady-state analysis of metabolic behavior. FBA (Fell and Small, 1986; Kauffman *et al*., 2003) is a particular constraint-based method which assumes that the network is regulated to maximize or minimize a certain cellular function, which is usually taken to be the organism's growth rate. FBA has been demonstrated to be a very useful technique for the analysis of metabolic capabilities of cellular systems (Price *et al*., 2004; Varma and Palsson, 1993). It involves carrying out a steady state analysis, using the stoichiometric matrix (as defined below) for the system in question. The system is assumed to be optimized with respect to functions such as maximization of biomass production or minimization of nutrient utilization; it is solved accordingly to obtain a steady state flux distribution, which is then used to interpret the metabolic capabilities of the system.

In FBA, the constraints imposed by stoichiometry imply that for each of the M metabolites in a network the net sum of all production and consumption fluxes, weighted by their stoichiometric coefficients, is zero:

$$\sum_{j=1}^{N} S_{ij} v_j = 0, \qquad i = 1, \ldots, M \tag{1}$$

Here, S_{ij} is the element of the stoichiometric matrix S corresponding to the stoichiometric coefficient of metabolite i in reaction j. The flux v_j is the rate of reaction j at steady state, and is the j-th component of an N-dimensional flux vector v, where N is the total number of fluxes.

Additional constraints, including those pertaining to the availability of nutrients, the reversibility of reactions, or the maximal fluxes that can be supported by enzymatic pathways, can be introduced by using the bounds α and β in the following inequalities:

$$\alpha_j \leq v_j \leq \beta_j \tag{2}$$

A natural choice for an objective function in metabolic models of prokaryotes and simple eukaryotes is biomass production (Price *et al*., 2004; Varma and Palsson, 1993), as it is reasonable to hypothesize that unicellular organisms have evolved towards maximal growth performance. This process is formalized by introducing a growth flux that transforms a linear combination of fundamental metabolic precursors into biomass. The maximization of biomass production is implemented by defining an additional flux v_{gro} associated with cell growth. For this flux, the stoichiometric factors of the reactants are the experimentally known proportions c_i of metabolite precursors X_i contributing to biomass production (Price *et al*., 2004):

$$c_1 X_1 + c_2 X_2 + \cdots + c_M X_M \xrightarrow{v_{\text{gro}}} \text{Biomass} \tag{3}$$

The search for the flux vector maximizing v_{gro} under the constraints of Equations (1) and (2) is solved using linear programming.

2.2 Expression-based measure for metabolic genes

We used Rosetta's compendium dataset (Hughes *et al*., 2000) which measures expression profiles of over 6200 *S.cerevisae* ORFs across 287 deletion strains and 13 chemical conditions. In addition, the dataset contains 63 negative control measurements comparing two independent cultures of the same strain. These were used to establish individual error models for each ORF, providing not only the raw intensity and the ratio measurement values for each experimental data point, but also a P-value evaluating the significance of change in expression level. The expression-based similarity (EXBS) measure between ORFs X and Y was computed according to 1-|*Spearman_rank*(p_x,p_y)| where p_x and p_y are expression profile vectors of X and Y, respectively, and the Spearman rank was calculated as in Press *et al*. (2002).

2.3 Topology-based measure for metabolic genes

As proposed in Kharchenko *et al*. (2005), the metabolic network structure can be used to calculate the network distance between genes. Let us define a pair of directly connected metabolic genes as separated by distance 1, and the network distance between genes X and Y to be the length of the shortest path from X to Y in the metabolic network. For sake of consistency, we call this measure a *topology-based similarity* (TOBS) measure. While any metabolite can be used to establish connections between metabolic genes, the relationships established by the common metabolites and cofactors—such as ATP, water or hydrogen—are not likely to connect genes with similar metabolic functions. Hence, when compiling the metabolic network to this end, we consider a subset of metabolites which excludes the most highly connected metabolic species. An exclusion threshold was determined based on the connectivity of the resulting network. A total of the 10 most highly connected metabolites (ATP, ADP, AMP, CO_2, H, H_2O, NADP, NADPH, phosphate and diphosphate), which compose 1% of all metabolites, and their mitochondrial and external analogs were excluded. Excluding up to the top 3% of all metabolites maintains the general trends described above.

2.4 Phylogenetic profiling analysis

Ten sequenced fungal genomes (*S.cerevisiae*, *C.albicans*, *C.glabrata*, *C.neoformans*, *D.hansenii*, *E.cuniculi*, *E.gossypii*, *K.lactis*, *S.pombe*, *Y.lipolytica*) were used to construct phylogenetic profiles. The phylogenetic profile of a gene is a string of ones and zeros that encodes the presence or absence, respectively, of the gene in the corresponding genomes. We define a conservation-based similarity (COBS) measure to be the similarity between phylogenetic profiles, computed using a normalized Hamming distance (Hamming, 1950). The normalized Hamming distance measures the degree of overlap between two sets of values, x and y, and is computed as the fraction of unmatched non-zeros between **x** and **y** among all non-zeros of **x** and **y**:

$$\hat{h}(x, y) = \frac{x^T x + y^T y - 2x^T y}{n},$$

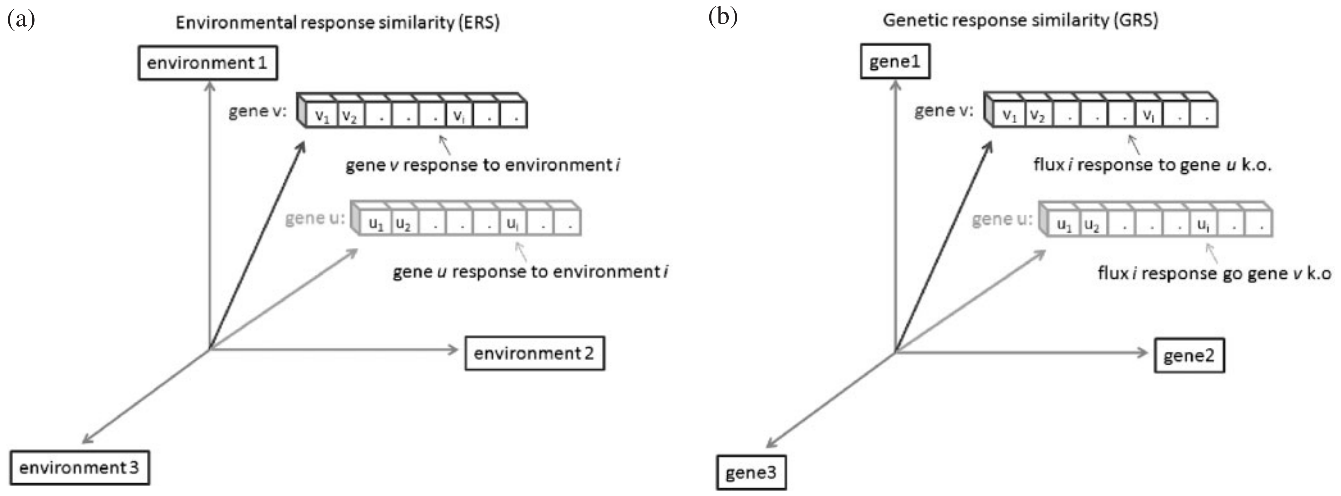
(a)
Environmental response similarity (ERS)
environment 1
gene v:
v1 v2 . . . vi . .
gene v response to environment i
gene u:
u1 u2 . . . ui . .
gene u response to environment i
environment 2
environment 3
(b)
Genetic response similarity (GRS)
gene1
gene v:
v1 v2 . . . vi . .
flux i response to gene u k.o.
gene u:
u1 u2 . . . ui . .
flux i response go gene v k.o
gene2
gene3

Fig. 1. A schematic illustration of two types of similarity measures between metabolic genes. (**a**) The ERS measure. Each element in vector $u(v)$ corresponds to the response of gene $u(v)$ to environment i. (**b**) the GRS measure. Each element in vector $u(v)$ corresponds to the response of flux i to the knockout of gene $u(v)$.

where $\|x\| = \|x\|_2^2 = \|x\|_1$ is the number of non-zeros in an n-dimensional binary vector x.

2.5 Model-based similarity measures for metabolic genes

In the context of the aforementioned motivation, we suggest two basic approaches for defining and measuring the similarity between metabolic genes: a GRS measure and an ERS measure. These are two complementary approaches, where the first reveals the effect of a genetic perturbation on the metabolic surrounding of a gene of interest, the other reveals the effect of the environmental perturbations on the gene of interest. A schematic illustration of both approaches can be seen in Figure 1.

2.5.1 Genetic response similarity Previous studies suggest that the metabolic state of an organism following genetic perturbations is close to that of the wild-type strain and does not necessarily achieve optimal growth rates as predicted by FBA (Segre *et al*., 2002; Shlomi *et al*., 2005). Specifically, regulatory on-off minimization (ROOM) was shown to successfully predict the metabolic state of knocked-out strains by minimizing the number of significant flux changes required for the wild-type strain to adapt to the gene knockout (Shlomi *et al*., 2005). Here, we use a variant of ROOM that uses the L_1 norm to minimize the number of significant flux chances (Kuepfer *et al*., 2005; Shlomi *et al*., 2005) instead of L_0 which is computationally harder. We define the GRS similarity measure between two genes as the distance between the ROOM solutions obtained for each of their knockouts. In some cases, ROOM (similarly to FBA) does not provide a unique solution, but rather a space of possible solutions. In these cases, instead of arbitrarily choosing a single ROOM solution, we define the GRS measure as the minimal distance between any two ROOM solutions for both genes. This is achieved by formulating a single optimization problem to find two ROOM solutions with minimal distance between them. The pseudocode of the procedure for computing GRS appears in Figure 2.

Throughout our study, we also examined the effect of excluding the isoenzymes from the analysis, as the model is uncapable to define which one of them is active. Notably, after exclusion of isoenzymes, the results obtained remain qualitatively similar across the entire analysis.

2.5.2 Environment response similarity This measure aims to capture the similarity betweens the patterns of flux activity of two genes across a variety of growth media. To this end, we follow and extend the approach of Bilu *et al*. (2006), which studied the relation between the flux ranges of different reactions/genes and their regulation and conservation. Specifically, we compute genes' activities across 100 randomly generated growth media, employing flux variability analysis (Mahadevan and Schilling, 2003; Reed and Palsson, 2004): for each reaction we computed the maximal and minimal flux values attainable in the space of optimal flux distributions for growth conditions simulating 100 different growth media. Random growth media were generated by setting limiting values to the uptake reactions independently at random. With probability 0.5, the maximal uptake rate was set to 0, i.e. only excretion was allowed. Otherwise, uptake rate was limited to a value chosen uniformly at random in the range [0.01, 5], at a resolution of 0.01. A similar sampling method was used in Almaas *et al*. (2005). In addition, to ensure sufficient variability between media, we switched between aerobic and anaerobic growth media with probability 0.5.

For each generated growth medium, we predicted which of the reactions are active, i.e. carry a non-zero metabolic flux (namely either its maximum or minimum flux values are different than zero). Active genes were denoted by '0' and non-active ones by '1'. This way we created for each gene a binary vector of its activity across a series of generated media. We define the ERS measure as the normalized Hamming distance (see Section 2.4) between two binary vectors reflecting metabolic genes' activity. The pseudo-code of the entire procedure is presented in Figure 3.

3 RESULTS

Functional similarity between genes is commonly inferred based on similarity in expression patterns across conditions (Eisen *et al*., 1998; Tavazoie *et al*., 1999). Following this paradigm, we define the ERS measure between gene pairs as the similarity[1] in their predicted flux activity patterns across multiple growth environments (Fig. 1a and the Methods section).

[1] For sake of clarity and for being consistent we use the similarity notion instead of distance for all the measures presented in this study.

```
Procedure ComputeGRS
 Output: $results$ - matrix $num\_genes \times num\_genes$
         containing the distance between metabolic genes.
 Run FBA to maximize biomass (growth rate), obtain wild type
 flux distribution $w$;
 for each gene $g$ do
     $A \leftarrow$ set of reactions associated with the deleted gene g;
     Compute knock-out flux distribution $v_g$ and a minimal
     distance from the wild-type distribution $l_g$ by solving the
     following LP problem:
     $\min \|v_g - w\|_{L1}$
     $s.t. \quad S \cdot v_g = 0; \quad v_{min} \leq v_g \leq v_{max};$
       $v[ko1] = 0, ko1 \in A;$
 end
 for each gene $g_1$ do
     for each gene $g_2 \neq g_1$ do
         $A_1 \leftarrow$ set of reactions associated with the gene $g_1$;
         $A_2 \leftarrow$ set of reactions associated with the gene $g_2$;
         results[g1][g2] = $dist$ where $dist$ is an objective
         function of the following LP problem:
         $\min \|v_{g1} - v_{g2}\|_{L1}$
         $s.t.$
           $S \cdot v_{g1} = 0; \quad v_{min} \leq v_{g1} \leq v_{max};$
             $v_{g1}[ko1] = 0, ko1 \in A_1;$
           $S \cdot v_{g2} = 0; \quad v_{min} \leq v_{g2} \leq v_{max};$
             $v_{g2}[ko2] = 0, ko2 \in A_2;$
           $\|w - v_{g1}\|_{L1} = l_{g1}; \quad \|w - v_{g2}\|_{L1} = l_{g2};$
     end
 end
```

Fig. 2. The procedure for computing the GRS measure. First, for each knocked-out gene the flux distribution is computed over the remaining fluxes. Then for each pair of genes the minimal distance (under L_1 norm) between the corresponding flux distributions is computed. When solving LP problems, S is a stoichiometric matrix and $v_{\min}$, $v_{\max}$ limit nutrient uptake and define the reactions' irreversibility.

```
Procedure ComputeERS($N$)
 Input: $N$ - the number of required media.
 Output: $results$ - matrix $num\_genes \times num\_genes$
         containing the distance between metabolic genes.
 for k=1..N do
     for each external flux $f$ do
         with probability 0.5, set $f = 0$;
         otherwise $f$ receives a random value chosen uniformly in
         the range [0.01, 5];
     end
     Run FBA to maximize biomass (growth rate), obtain
     $wild\_growth\_rate$;
     Add constraint: biomass $\geq 0.9 * wild\_growth\_rate$;
     for i=1..$num\_fluxes$ do
         Run FBA to maximize flux $i$, obtain $i_{max}$;
         Run FBA to minimize flux $i$, obtain $i_{min}$;
     end
     for each gene $g$ do
         if for one of its related fluxes $i_{max} = i_{min} = 0$ then
             $activity\_vec$[g][k] = 1;
         else
             $activity\_vec$[g][k] = 0;
         end
     end
     for each gene g1 do
         for each gene g2 $\neq$ g1 do
             results[g1][g2] =
             Hamming_distance($activity\_vec$[g1],$activity\_vec$[g2]);
         end
     end
 end
```

Fig. 3. The procedure for computing the ERS measure. For each simulated medium flux variability analysis is applied in order to create an activity profile for each gene. Then the distance between the computed profiles is calculated.

Cellular response to a gene knockout involves rerouting of metabolic flux through alternative pathways and the utilization of isoenzymes (Emmerling *et al.*, 2002). We hypothesize that similar metabolic responses to gene knockouts may provide evidence for similar metabolic functionality between genes. Based on this hypothesis, we define the GRS measure between gene pairs as the similarity in the metabolic response following their knockout (Fig. 1b and the Methods section).

We applied the proposed computational measures to the metabolic network model of *S.cerevisiae* by Duarte *et al.* (2004). The model consists of 1060 metabolites and 1149 reactions (accounting for 750 genes). The obtained ERS and GRS measures were found to be significantly correlated (R^2=0.53, P-value = 3.2 × 10^{-3}), testifying that indeed both measures capture the same overall signal. The remaining analysis provides evidence that these measures are indeed indicative of functional similarity and outperform the strictly topological measures.

3.1 Validating the similarity measures based on GO

To assess the accuracy of the ERS and GRS measures, we compared them to the GO functional annotations. Specifically, we expect an accurate similarity measure to have relatively high values for genes that are annotated with the same GO term, and low values for genes in different terms. In our analysis, we used all non-redundant (i.e. containing different genes) GO terms of sizes between 5 and 100. The overlap between these gene sets is quite low, as shown in Figure 1 of the Supplementary Material. For each such GO term, we computed the average distance between all genes annotated with this term. To assess the statistical significance of the average distance, we compared it to average distances obtained for 10 000 random sets of genes, whose annotations were randomly shuffled while preserving the overall annotation distribution, obtaining an empirical P-value. The resulting P-values were further corrected for multiple testing of the many annotations via the false discovery rate procedure (Benjamini and Hochberg, 1995). The averaged similarity measures and the corresponding P-values are shown in the Supplementary Material.

We define a GO term to be *consistent* under some similarity measure if the resulting P-value (after FDR correction) for this term under this similarity measure is significant ($\leq$0.05). Our results show that 86.5 and 18.7% of the GO terms are consistent under the ERS and GRS measures, respectively (Fig. 4). Interestingly, although the ERS provides better results

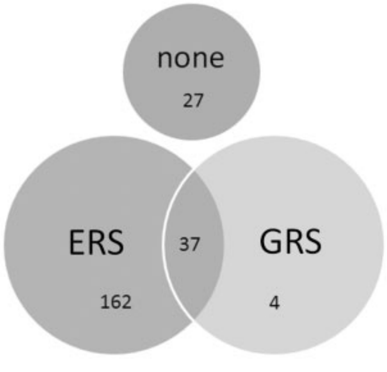

Fig. 4. A Venn diagram displaying the consistency of GO terms under the ERS and GRS measures.

overall, in some cases, only the GRS measure truly captures the functional similarity within some GO terms. For example, GRS finds the GO term 42 724 corresponding to thiamin and derivative biosynthetic process to be consistent, while ERS does not. The results suggest that GRS outperforms ERS only for small GO terms (of size 5) where ERS does not receive a *P*-value significant enough to define a GO term as consistent. One putative reason for this may be the noisyness of the ERS measure, due to the large number of genes that tend to be active across many growth media. Comparing the ERS measure with other commonly used measures of functional similarity (Fig. 5 and the Methods section), we find that the ERS measure outperforms both: the EXBS and COBS measures, which obtain only 29.1 and 12.6% of consistent GO terms, respectively. Moreover, 82 and 96.5% of consistent terms found by EXBS and COBS measures were also found consistent by the ERS method. Using an alternative similarity measure (the Jaccard coefficient) between phylogenetic profiles provided similar results. Additionally, we tested the COBS measure with a different set of phylogenetic profiles, consisting of 17 higher eukaryotes (from NCBI's HomoloGene's database). Using this dataset, only 5.6% of the GO terms were found to be consistent under the COBS measure, testifying that conservation coherency is indeed much stronger among the closely related yeast genomes. As a next step, we compared the accuracy of the ERS and GRS measures to a measure obtained by considering only the topology of the network—TOBS (Fig. 5). We find that the ERS measure outperforms TOBS with 86.5% against only 63.9% rate of discovering consistent GO terms. ERS covers 82 and 96.5% while TOBS covers only 74 and 3.5% of consistent GO terms found by EXBS and COBS, respectively.

To gain further insights as to why ERS outperforms the simpler topological measure, it is illustrative to examine the GO term 0006696 which corresponds to the process of ergosterol biosynthesis, for example. As shown in Figure 6, ergosterol biosynthesis is carried out through a long, chain-like pathway, and hence the average distance between genes annotated with this term is significantly high (with a topological similarity *P*-value of 0.35). On the other hand, since these genes form an unbranched linear pathway, mass-balance constraint determines that all genes should either be coherently active or non-active. In this case, both the ERS and GRS measures show significant high similarity scores with *P*-values of 9.99×10^{-5} and 1.7×10^{-3}, respectively. We note however that for this specific example, the expression similarity term is also relatively high, with a *P*-value of 1.8×10^{-3}.

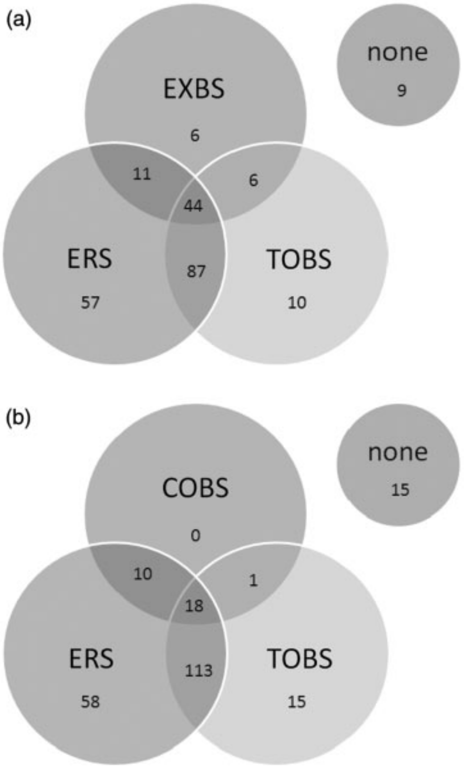

Fig. 5. A Venn diagram presenting the consistency of GO terms under ERS and TOBS versus the EXBS **(a)** and COBS **(b)** measures.

Other cases where the topological similarity measure fails to identify true functional similarities relate to the identification and removal of currency metabolites. The removal of currency metabolites (which are hubs in the network) is essential for the topological similarity measure to make any sense. Without the removal of these metabolites, the average distance between two genes is as low as 1.78 and only 1.3% of the GO terms are identified as consistent. However, the removal of currency metabolites may cause functionally related genes to be relatively far. For example, the genes annotated as involved in GO term 15 698, corresponding to inorganic anion transport dissociate into four densely connected clusters in the network if the currency metabolite *inorganic phosphate* is removed.

3.2 Validating the similarity measures based on gene expression data

Similarity in gene expression patterns across multiple conditions is commonly used as indication of functional similarity (Eisen *et al*., 1998; Tavazoie *et al*., 1999). Specifically, this paradigm is further strengthened in the context of metabolic genes, whose expression is adjusted ‘just-in-time’ according to metabolic demands (Zaslaver *et al.,* 2004). Notably, although similarity in expression is believed to be indicative of functional similarity, a comparison between the two only reveals a moderate correlation (Sevilla *et al*., 2005), with this claim further supported in the results shown in Figure 5.

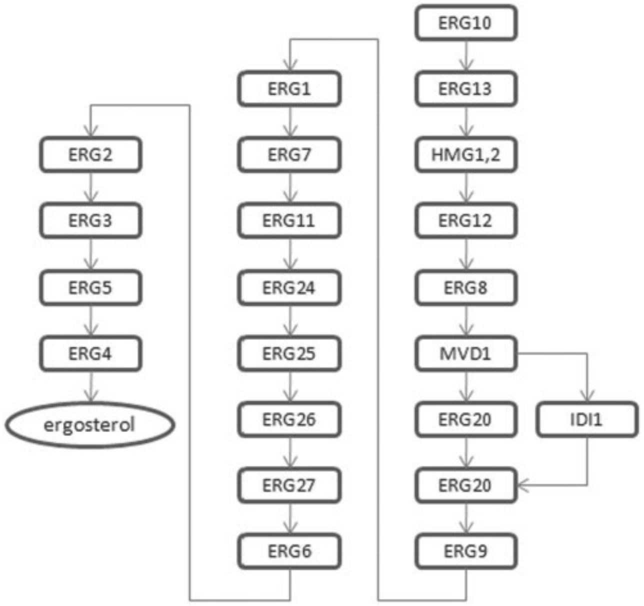

Fig. 6. The ergosterol biosynthesis pathway. Each node (rectangle) represents an enzyme [except for the last one (ellipse) representing the final product—ergosterol]. Each edge represents a metabolite which is produced by one enzyme and consumed by the following one in the pathway. Since ergosterol biosynthesis is carried out through a long, chain-like pathway, the average distance between genes annotated with this term is significantly high, while mass-balance constraint determines that all genes should either be coherently active or non-active. Thus ERS outperforms the TOBS topological measure.

Measuring the correlation between the GRS and EXBS measures we observed (see Fig.7a) a moderate correlation (R^2=0.38 with a *P*-value of 2.1×10^{-2}). As for the ERS measure, we observe (see Fig. 7a) that it exhibits a strong correlation with the expression similarity (R^2=0.94 with a *P*-value of 1×10^{-9}). The correlations were obtained using a linear binning procedure (Sevilla *et al*., 2005) which averages one measure values over uniform intervals of the second measure. We note that our results regarding the correlation between ERS and expression similarity are in agreement with previous findings (Bilu *et al*., 2006; Famili *et al*., 2003; Reed and Palsson, 2004; Schuster *et al*., 1999, 2002).

Measuring the topological similarity measure and expression similarity showed a weaker, but still strong correlation of R^2=0.78 (*P*-value = 6.6×10^{-5}), demonstrating that genes closer to each other in the metabolic network tend to have, on average, higher level of co-expression (Fig. 7b), in agreement with the previous findings of Kharchenko *et al*. (2005).

Finally, we tested whether the ERS measure is advantageous over the TOBS measure, using a partial correlation test (Kendall and Stuart, 1969). The partial correlation method quantifies the correlation between two variables whilst eliminating the effects of another variable on this relationship, namely network distance in our case. Our results show a significant partial correlation ($R^2 = 0.65$, with a *P*-value of 3.8×10^{-6}) between ERS and similarity in expression levels. This result further supports the claim that the ERS similarity measure better captures the true functional similarity between genes compared to the TOBS topological measure. Furthermore, this result reaffirms FBA’s ability to accurately predict metabolic behavior across multiple conditions.

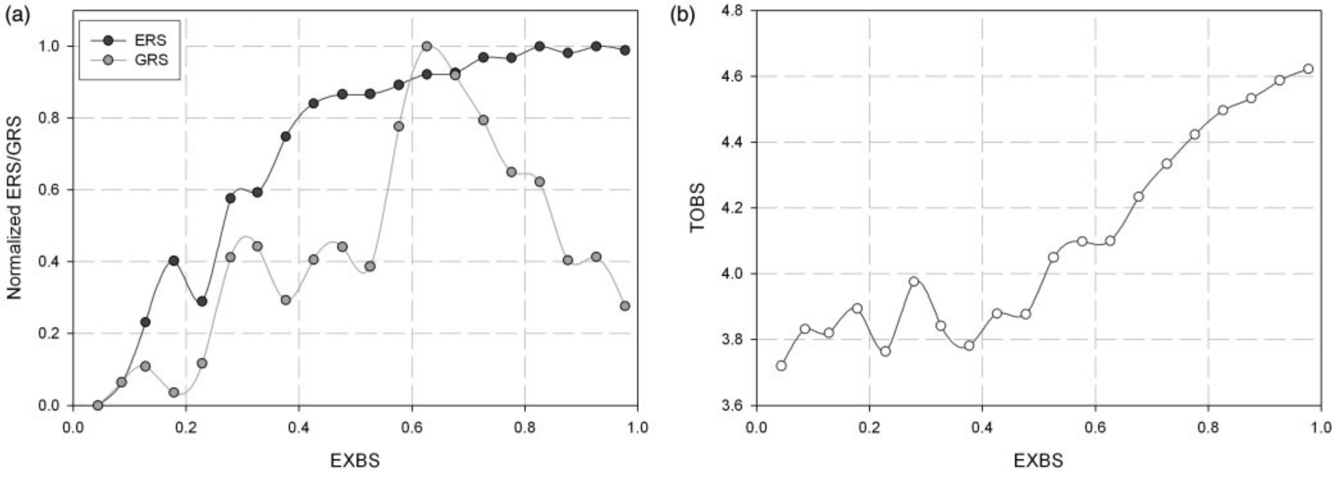

Fig. 7. Correlation between co-expression levels (EXBS) and model-based (GRS and ERS) or topology-based (TOBS) measures. The correlation is obtained by dividing the EXBS axis into uniform intervals and averaging the corresponding values of GRS, ERS and TOBS in each interval. (**a**) GRS/ERS measures. (**b**) TOBS measure.

4 DISCUSSION

This article shows that metabolic network-based similarity measures between genes can go beyond previous measures that are based solely on network topology. We applied two schemes to compute this similarity: the GRS scheme and the ERS scheme. While the former shows a fairly moderate correlation with the experimental results as well as a pretty modest ability for explicating GO terms, the latter provides a strong, statistically significant measure. One possible explanation of this behavior may be that the ERS studies probe the natural wild type across a variety of media, whereas the GRS method does it in less natural strains and in a sole media. Another reason may be the more cumbersome computational method used in the GRS case, which is likely to add significant noise to the results obtained.

Furthermore, when examining the correlation with co-expression levels, one can observe that the GRS measure shows a certain decline as levels of EXBS approach 1. We believe that this phenomenon is driven by the nature of the GRS measure which is based on an underlying process of rerouting the metabolic fluxes through isoenzymes and alternative pathways. Recently, Kafri *et al.* (2005) have shown that in yeast most duplicate-associated backups involve genes that—on average—are not strongly co-expressed.

Notably, one cannot expect to find a 100% accuracy in finding consistent GO terms under the model-based measures as well as an absolute correlation between the model-based measures and gene co-expression. In essence, the fluxes predicted by the ERS measure across various growth media reflect a 'wishful thinking' of an ideal system whose regulatory apparatus has developed with the sole optimization objective of maximizing growth. In this sense, the high levels of consistent terms (80–90%) and the high levels of correlations (0.8–0.9) found with the ERS measure in this study are truly striking.

Similarity in gene expression patterns across multiple conditions is commonly used as an indication of functional similarity. However, our results show that in many cases genes that are annotated with the same GO term are not expression coherent. Specifically, we find that only ~30% of the GO terms are composed of genes which are expression coherent. This lack of expression coherency may be the result of the complex interplay ongoing between metabolic and hierarchical regulation (ter Kuile and Westerhoff, 2002). Remarkably, the ERS and GRS measures show significant high similarity values for 62.6 and 13% of the GO terms that are not expression coherent, showing their advantage over this traditional similarity measure.

One important problem that can be addressed in this context is that of functional prediction of gene annotation. It is well known that sequence similarity predicts rather well GO function annotations but fails to predict GO process annotation. In a similar vein, we computed the correlation between GO function and process annotations and sequence similarity of metabolic genes, using the measure of semantic similarity introduced by Resnik 1995. We observed a significant correlation between sequence similarity and GO functional annotations (R^2=0.95, P-value = 2.3×10^{-5}), while for process annotation the correlation was very low and insignificant (R^2=0.4, P-value = 0.2). Hence, quite obviously there is much room for new approaches for process annotation. Our study suggests that model-based, topology-based and expression measures can contribute to the GO process annotation in a synergistic manner, with ERS having the largest potential contribution. Nevertheless, the goal of the method presented is not to provide functional annotation of new, unannotated genes, but rather to explore the functional relations between genes across the network, showing quite a few novel and interesting insights.

Finally, it is pertinent to consider the role of genomic and annotation information used in the reconstruction of the metabolic networks that are at the basis of our approach. We believe that one of the main ideas underlying the study of networks in systems biology is that one may find emergent network properties, i.e. new phenomena that were not explicit when constructing the network from its basic building blocks. The same idea is applied in this work: although genomic and annotation information have been used during the reconstruction of the metabolic network, our model is further based on considerable additional information, including the intrinsic network topology, the reactions stochiometry, the growth media and the mass balance and biomass maximization assumptions. All these transcribe together in a complex manner to reveal additional and different functional roles/annotations of the genes involved, as testified to by the results we report in this article. Specifically, one can note that our approach is essentially different than using similarity that is solely computed based on GO annotations (known also as semantic similarity). First, the latter is based on the partitioning of genes to groups/terms while this partition does not explicitly exist in the metabolic network. Furthermore, as we show by comparing to expression and conservation data, the functional similarity measures presented in this article outperform the metabolic network TOBs measure which is obviously closely related to the GO annotation.

ACKNOWLEDGEMENTS

R.S. was supported by an Alon Fellowship. T.S. is supported by an Eshkol Fellowship from the Israeli Ministry of Science. E.R.'s research is supported by grants from the Israeli Science Foundation (ISF), the German Israeli Research Fund (GIF) and the Yeshayahu Horowitz Center for Complexity Science.

Conflict of Interest: none declared.

REFERENCES

Almaas,E. *et al.* (2005) The activity reaction core and plasticity of metabolic networks. *PLoS Comput. Biol.*, **1**, e68.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.

Bilu,Y. *et al.* (2006) Conservation of expression and sequence of metabolic genes is reflected by activity across metabolic states. *PLoS Comput. Biol*, **2**, e106.

Chen,L. and Vitkup,D. (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol.*, **7**, R17.

Duarte,N. *et al.* (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.*, **14**, 1298–1309.

Eisen,M. B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Emmerling,M. *et al.* (2002) Metabolic flux responses to pyruvate kinase knockout in Escherichia coli. *J. Bacteriol.*, **184**, 152–164.

Famili,I. *et al.* (2003) Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl Acad. Sci. USA*, **100**, 13134–13139.

Fell,D. and Small,J. (1986) Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem. J.*, **238**, 781–786.

Hamming,R. W. (1950) Error detecting and error correcting codes. *Bell Syst. Technical J.*, **26**, 147–160.

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

Jeong,H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kafri,R. *et al.* (2005) Transcription control reprogramming in genetic backup circuits. *Nat. Genet.*, **37**, 295–299.

Kauffman,K. *et al.* (2003) Advances in flux balance analysis. *Curr. Opin. Biotechnol.*, **14**, 491–496.

Kendall,M. G. and Stuart,A. (1969) *The Advanced Theory of Statistics*. Vol. 1, Charles Griffin, London.

Kharchenko,P. *et al.* (2005) Expression dynamics of a cellular metabolic network. *Mol. Syst. Biol.*, **1**, E1–E6.

Kuepfer,L. *et al.* (2005) Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome Res.*, **15**, 1421–1430.

Mahadevan,R. and Schilling,C. (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.*, **5**, 264–276.

Press,W. H. *et al.* (2002) *Numerical Recipes in C++: The Art of Scientific Computing*, Cambridge University Press, Cambridge.

Price,N. D. *et al.* (2004) Genome-scale Models of Microbial Cells: Evaluating the consequences of constraints. *Nat. Rev. Microbiol.*, **2**, 886–897.

Ravasz,E. *et al.* (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Reed,J. and Palsson,B. (2004) Genome-scale in silico models of e. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Res.*, **14**, 1797–1805.

Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy. *IJCAI*, 448–453.

Rokhlenko,O. *et al.* (2006) Flux-based vs. topology-based similarity of metabolic genes. *In WABI 2006, LNBI 4175*, 274–285.

Schuster,S. *et al.* (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.

Schuster,S. *et al.* (2002) Use of network analysis of metabolic systems in bioengineering. *Bioprocess Biosyst. Eng.*, **24**, 363–372.

Segre,D. *et al.* (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl Acad. Sci. USA*, **99**, 15112–15117.

Sevilla,J. L. *et al.* (2005) Correlation between gene expression and go semantic similarity. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **2**, 330–338.

Shlomi,T. *et al.* (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci. USA*, **102**, 7695–7700.

Tavazoie,S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

ter Kuile,B. H. and Westerhoff,H. V. (2002) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.*, **500**, 169–171.

Varma,A. and Palsson,B. (1993) Metabolic capabilities of Escherichia coli: II. Optimal growth patterns. *J. Theor. Biol.*, **165**, 503–522.

Zaslaver,A. *et al.* (2004) Justin-time transcription program in metabolic pathways. *Nat. Genet.*, **36**, 486–491.