BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i178–i185
DOI: 10.1093/bioinformatics/bth930

Filling gaps in a metabolic network using expression information

Peter Kharchenko[†], Dennis Vitkup[†] and George M. Church[*]

Department of Genetics, Harvard Medical School, 77 Louis Pasteur Avenue Boston, MA 02115, USA

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: The metabolic models of both newly sequenced and well-studied organisms contain reactions for which the enzymes have not been identified yet. We present a computational approach for identifying genes encoding such missing metabolic enzymes in a partially reconstructed metabolic network.

Results: The metabolic expression placement (MEP) method relies on the coexpression properties of the metabolic network and is complementary to the sequence homology and genome context methods that are currently being used to identify missing metabolic genes. The MEP algorithm predicts over 20% of all known *Saccharomyces cerevisiae* metabolic enzyme-encoding genes within the top 50 out of 5594 candidates for their enzymatic function, and 70% of metabolic genes whose expression level has been significantly perturbed across the conditions of the expression dataset used.

Availability: Freely available (in Supplementary information).

Contact: g1m1c1@arep.med.harvard.edu

Supplementary information: Available at the following URL http://arep.med.harvard.edu/kharchenko/mep/supplements.html

1 INTRODUCTION

With a growing number of completely sequenced genomes, increasing attention has been devoted to understanding the functional coordination of individual genes in complex biological processes. Information about numerous relationships among genes and gene products is represented at various levels by protein–protein interaction, regulatory and metabolic networks. A combination of experimental and computational techniques is being used in large-scale efforts to reconstruct such networks (Karp, 1998; Uetz *et al*., 2000; Ito *et al*., 2001; Forster *et al*., 2003; Kaern *et al*., 2003). Metabolism currently presents a particularly suitable target for computational analysis. The metabolic pathways are well characterized, and while metabolic capabilities of various organisms can differ, reconstruction efforts benefit from a conserved nature of the underlying biochemical reactions and abundance of metabolic enzymes in multiple species. Computational reconstruction of metabolic networks typically uses genomic information to associate genes with enzymatic functions, thereby identifying the metabolic pathways encoded by the organism. The most common approach seeks to identify the genes responsible for a particular metabolic function by establishing sequence homology to functionally characterized proteins in other species. Similar techniques have been used extensively for general genome annotation, and a comprehensive set of resources has been developed for that purpose (Tatusov *et al*., 1997). Although the sequence homology-based methods have been remarkably successful overall, they fail to assign functions to a considerable fraction of genes (31-80%) in completely sequenced genomes (Iliopoulos *et al*., 2001), and have been known to produce imprecise or incorrect annotations (Devos and Valencia, 2001; Iliopoulos *et al*., 2003).

In many cases, while there exists sufficient biological evidence to believe that a given pathway is present in an organism, one or more enzymes responsible for the critical reaction steps cannot be identified via sequence homology methods alone. One possibility is that the gene encoding such enzymatic function is not present in a given organism, and the reaction is either bypassed or catalyzed by some other means. Another possibility is that the corresponding enzymes are encoded by genes with little or no sequence similarity to known orthologs as a consequence of convergent evolution or a horizontal transfer from a distant organism (Yanai *et al*., 2002; Kunin and Ouzounis, 2003). The identification of the enzymes catalyzing individual metabolic reactions in a well-characterized or nearly complete metabolic network has been referred to as the ‘missing genes problem’ (Osterman and Overbeek, 2003). In contrast to the traditional problem of gene annotation, where a functional description is assigned to a given gene, the missing gene problem assigns a gene to a specific metabolic function.

In addition to sequence homology, various types of genomic evidence have been used to identify the missing metabolic genes in several organisms (Bobik and Rasche, 2001; Bishop *et al*., 2002). Functional coupling to known genes has been

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

inferred by analyzing gene clustering on the chromosome (Overbeek *et al*., 1999), by monitoring protein fusion events (Marcotte *et al*., 1999), and by studying gene co-occurrence profiles across multiple species (Pellegrini *et al*., 1999). A combination of such techniques, commonly referred to as genome context analysis, is an integral part of the available metabolic reconstruction tools (Overbeek *et al*., 2000; Osterman and Overbeek, 2003). Despite these advances, missing genes still remain abound in metabolic models of recently sequenced and even well-studied organisms (Bono *et al*., 2003; Forster *et al*., 2003).

Similarity of gene expression profiles has been used extensively to assign genes to general functional categories (DeRisi *et al*., 1997; Wen *et al*., 1998; Tavazoie *et al*., 1999; Wu *et al*., 2002). However, prediction of specific gene function from expression information alone has, so far, not been possible. This suggests that an application of expression analysis to the missing genes problem would require additional sources of information. For example, a recent study by van Noort *et al*. (2003), showed that increased prediction specificity can be attained by considering expression conservation across multiple species. Here, we illustrate that specific functional predictions can be made by considering expression similarity together with the structural information of the metabolic network. Our recent work (P.Kharchenko, G.M.Church and D.Vitkup, submitted for publication) has demonstrated a prevalence of local co-expression of genes in the metabolic network. This suggests that the expression profiles of the metabolic genes contain information about their precise location in the metabolic network, and therefore, their enzymatic function.

Here we present an automated method, metabolic expression placement (MEP), for selecting candidate genes for an unassigned enzymatic function, based on the gene expression data and structure of the partially reconstructed metabolic network. The method does not directly rely on genomic data, providing an assessment complementary to the genome context analysis techniques currently being used to search for missing metabolic genes. The effectiveness of the method is demonstrated by restitution of known yeast metabolic enzymes. We analyze factors determining the quality of the predictions, and suggest strategies for targeting specific enzymatic functions using this approach.

2 METHODS

2.1 Metabolic dependency graph, network distance and neighborhoods

As in our previous work (P.Kharchenko, G.M.Church and D.Vitkup, submitted for publication), metabolism was represented in a form of a metabolic gene dependency graph. Nodes of the graph correspond to metabolic genes, the edges to dependencies established by metabolic reactions (Fig. 1). Dependencies between genes were established according to the following definition:a metabolic gene X is dependent on a gene Y if and only if there exists a metabolite that is (1) produced by a reaction catalyzed by the product of gene X; and (2) consumed by a reaction catalyzed by the product of gene Y.

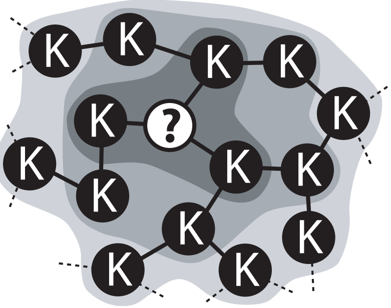

Fig. 1. A schematic illustration of the MEP approach. The missing metabolic gene (designated by the question mark) is surrounded by known metabolic genes (marked with the letter ‘K’). The network neighborhood formed by the known genes, consists of three neighborhood layers of increasing radii, as indicated by the background color. The MEP algorithm uses combined expression of the network neighborhood to identify candidates for the missing metabolic genes.

The metabolic gene dependency graph is then used to calculate network distance between the genes. We define a pair of directly dependent metabolic genes X and Y to be separated by a distance 1; similarly, the distance between X and Z, given dependencies between $X \rightarrow Y$ and $Y \rightarrow Z$, is 2 and so on. Network distance is always symmetric, so the distance from X to Z is equal to the distance from Z to X. In general, we define the metabolic network distance between genes X and Y as a length of the shortest path from X to Y on the undirected metabolic dependency graph. A manually curated metabolic network model of the *Saccharomyces cerevisiae* (Forster *et al*., 2003) was used to construct a comprehensive metabolic dependency graph, consisting of 1172 reactions on 786 metabolic species. While any metabolite can be used to deduce gene dependencies, the relationships established by common cofactors, such as ATP are not likely to connect genes with similar metabolic functions. In compiling a global metabolic dependency graph we considered all metabolites, excluding the following highly connected metabolic species: ATP, ADP, AMP, CO_2, CoA, glutamate, H, NAD, NADH, NADP, NADPH, NH_3, orthophosphate and pyrophosphate.

The metabolic neighborhood of radius R around a gene X is defined as the set of all genes separated by network distance of R or less from gene X. The i-th layer of the network neighborhood is defined as a set of genes that are precisely distance i from gene X.

2.2 Distances between gene expression profiles

We have used Rosetta’s ‘compendium’ dataset (Hughes *et al*., 2000) as the source of gene expression information. This dataset measures expression profiles of over 6200 yeast open reading frames (ORFs) across 287 deletions and 13 chemical perturbations. The expression distance measure between ORFs X and Y is calculated as $1 - |\mathrm{corr}(p_x, p_y)|$, where $\mathrm{corr}(p_x, p_y)$ is the Spearman's rank correlation (Press *et al.*, 2002) between expression profile vectors of X and Y.

2.3 Cost functions

Gene candidates for catalyzing a particular unassigned metabolic reaction are identified using the MEP algorithm. Given a node L in the metabolic dependency graph, and a set of candidate ORFs, the MEP algorithm ranks the list of the candidate genes; the first ORF being the most probable candidate for a metabolic function described by L, and the last ORF being the least probable candidate. The ordering is determined by a cost function, which evaluates the similarity of expression profile of each candidate ORF with each member of the metabolic neighborhood of location L (Fig. 1). Two different types of cost functions were tested:

$$F(x) = \frac{1}{|N|} \sum_{i=1}^{R} \sum_{g \in N_i} \frac{w_i}{d(x,g)^p}, \qquad \text{(type1)}$$

$$F(x) = \sum_{i=1}^{R} w_i \left\langle \frac{1}{d(x,g)^p} \right\rangle_{g \in N_i}, \qquad \text{(type2)}$$

where x is the candidate gene, R is the network neighborhood radius, N is a neighborhood of radius R around the metabolic location L, $|N|$ is the total number of genes in the neighborhood, N_i is the set of genes in the i-th layer of the network neighborhood, $d(x, g)$ is the expression distance between genes x and g, $\vec{w}$ is a vector of the neighborhood layer weights and p is a positive power factor.

2.4 Performance assessment through self-ranking

To assess the performance of the method, a self-ranking test was conducted by running the MEP algorithm on known enzyme-encoding genes. The candidate set comprises all non-metabolic ORFs plus the gene being tested—a total of 5594 ORFs. The self-rank of a known metabolic gene is its rank in the candidate gene list ordered using the cost function. The self-rank can range anywhere from 1 to 5594. A self-rank of 1 indicates that the gene originally assigned to the metabolic reaction was determined to be the top candidate. The overall performance of the algorithm is quantified by calculating the fraction of well-ranked genes, i.e. the fraction of known metabolic enzyme-encoding genes that rank among the top K candidates, where K is chosen according to the desired stringency.

2.5 Neighborhood layer weight optimization, and error estimation

The optimal weight vector $\vec{w}$ was determined by minimizing the log sum of the self-ranks of known metabolic enzymes. Minimization was performed using the Nelder–Mead simplex algorithm (Nelder, 1965).

Confidence intervals on the presented results (weight values and self-ranks) were estimated by a non-parametric bootstrap method (Efron and Tibshirani, 1993). The set of known metabolic genes was sampled (with replacement) to obtain a list of genes of the same size. The weight optimization procedure was run on the sampled list of genes and the self-ranking performance of all known metabolic genes was measured using the optimal weights determined for each sample. This sampling/optimization/evaluation procedure was repeated 1000 times. The 95% confidence intervals on various self-rank statistics, and the weights themselves were then estimated from the resulting distributions using the percentage method.

2.6 Expression variability

The variability v_g of an individual gene g was measured as the number of experiments in which the expression level of the ORF g has changed with probability >0.9. The network neighborhood variability was calculated as a weighted sum of the mean gene variability of the individual neighborhood layers:

$$v_N = \sum_{i=1}^{R} \left(\frac{w_i}{N_i} \sum_{g \in N_i} v_g \right).$$

3 RESULTS AND DISCUSSION

3.1 Local coexpression in the *S.cerevisiae* metabolic network

Our approach for identifying missing metabolic genes relies on the relationships between structure of the metabolic network and the coexpression properties of metabolic enzymes. In order to treat such relationships quantitatively, we establish an abstract representation of metabolism and a formal definition of network distance between the enzymes. Existing knowledge of metabolism is used to determine a set of metabolic dependencies among the enzymes, resulting in a metabolic gene dependency graph (see Methods section). The network distance between two genes is then calculated as a shortest path between the corresponding nodes on this graph.

We have previously shown that the expression profiles of enzymes appearing near each other in the metabolic network tend to be more similar than expected by chance (P.Kharchenko, G.M.Church and D.Vitkup, submitted for publication). We find that the average distance between expression profiles of metabolic genes increases monotonically with their separation in the metabolic network (Fig. 2). The fact that mean expression distance of the genes adjacent in the network is significantly smaller than at any other metabolic separation might suggest that the adjacent genes of a given enzyme can be identified by simply selecting those genes with the highest expression similarity. To test this hypothesis, we have calculated the coexpression ranks of the adjacent

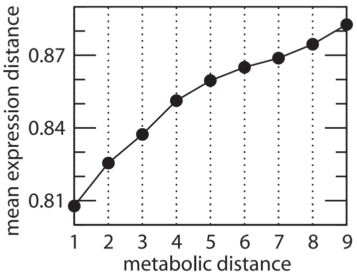

Fig. 2. Local coexpression in the metabolic network. Mean expression distance is shown as a function of the network distance between metabolic genes.

genes for every known metabolic enzyme. The adjacent genes, closest in terms of their expression, have an average rank of 759 (out of 6206 ORFs). The genes farthest in terms of expression rank, on an average, at 4405; and the mean rank of all adjacent genes is 2530. The distribution of the adjacent gene ranks is shown in Figure 3a. Only 8.2% of the adjacent enzymes appear within 50 most similarly expressed ORFs (Fig. 3b). The enzymes with the highest similarity of expression profiles are located, on an average, at a metabolic distance of 3.65, which is comparable with the mean metabolic distance in the network (4.47).

3.2 Metabolic expression placement: strategy and validation approach

These data reveal the need for a more sophisticated strategy to identify a significant fraction of missing metabolic genes from expression information. We sought to identify the metabolic enzymes by considering similarity between their expression profile and the combined expression of the surrounding metabolic network neighborhood. The MEP algorithm evaluates each candidate gene using a cost function that measures the correspondence of its expression profile to that of members of the individual layers of the network neighborhood (Fig. 1). Given a set of potential candidates, the MEP algorithm sorts them according the result of the cost function evaluation. The method relies on partial reconstruction of the metabolic network to identify metabolic neighborhood of the enzyme in question, and thus will be effective only if a substantial part of metabolism is already known. In other words, the algorithm is designed to fill the gaps in the metabolic network, and not to reconstruct the entire network *de novo*.

To determine an informative network neighborhood radius, we find a maximum metabolic distance up to which one can still observe statistically significant coexpression levels. Beyond such a distance, the coexpression ceases to be significant enough to warrant inclusion by the MEP algorithm. The significant coexpression radius of our network is between 3 and 4 (P.Kharchenko, G.M.Church and D.Vitkup, submitted for publication); calculations presented in this work use a network neighborhood of radius 3.

The co-expression structure within the network neighborhood is taken into account by assigning different weights to the contributions of individual neighborhood layers. As mean expression distance increases with metabolic separation (Fig. 2), contributions of neighborhood layers with lower network distance should be given more weight. Indeed, the values of the weight vector determined by the self-rank optimization procedure (see Methods section) follow this trend. The cost function is a weighted sum of the contributions of the individual neighborhood layers. The type 1 and type 2 cost functions differ in the way the size of the layer is taken into an account. The type 2 functions normalize the neighborhood layer contributions by the number of genes in the layer, while the type 1 functions do not. These functional types implement the basic rational of the method. More elaborate functional forms, for example neighbor ‘voting’, may perform better.

In order to evaluate the overall performance of the MEP algorithm, and to determine parameters such as the optimal type of cost function and the neighborhood weight vector, we use a test that quantifies the ability to predict the identity of known metabolic enzymes. The test measures a self-rank of every known enzyme-encoding gene, which is its rank among the set of all non-metabolic genes, in an ordering determined by the cost function (see Methods section). A perfect prediction algorithm would return a self-rank of 1 for every known metabolic gene, and a random prediction would result in a uniform distribution of self-ranks from 1 to 5594 (total number of ORFs that are not known to be metabolic enzymes). The self-rank test is designed to simulate MEP algorithm performance in trying to identify a missing metabolic gene. In a practical application of the method, however, it would be prudent to reduce the number of potential ORF candidates by eliminating from consideration all ORFs that are known not to be involved in a given metabolic activity, or in metabolism in general.

3.3 Parameter optimization and performance

To select an optimal value of the power factor p, we monitored performance of the algorithm on the entire set of known metabolic enzymes. We find that the mean self-rank values improve with increasing values of p, reaching a constant level around $p = 10$ (Figure 1 in the supplementary materials). The performance of the type 1 cost functions proved to be better than of type 2. The calculations presented in this paper use type 1 cost function with the power factor $p = 15$.

The neighborhood weight vector $\vec{w}$ was determined by minimizing the log sum of self-ranks of known metabolic enzymes (see Methods section). Because such self-ranks are also used to evaluate the overall performance of the algorithm, the effects of weight vector optimization should be taken into account. All the results presented below are reported together with the 95% confidence intervals determined by non-parametric bootstrap based on the 1000 random samples of known metabolic genes (see Methods section). Mean

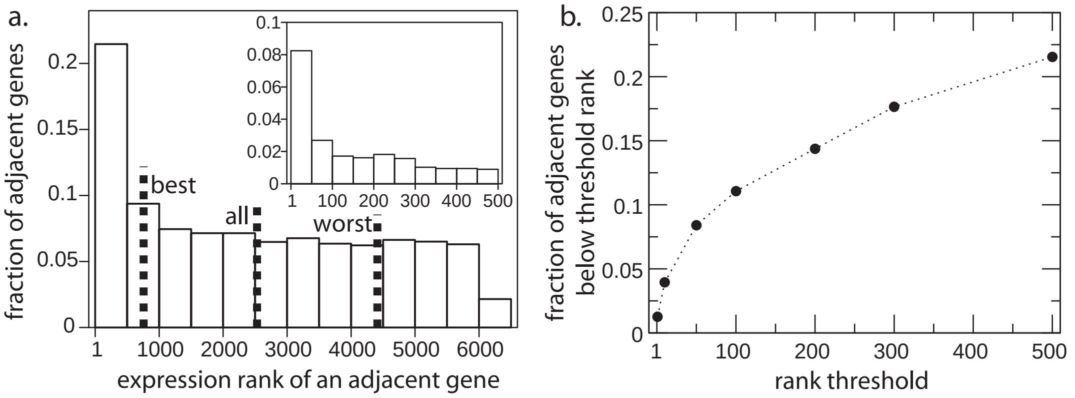

Fig. 3. Coexpression of genes adjacent in the metabolic network. (**a**) Expression ranks of metabolically adjacent genes. For each metabolic gene, the remaining yeast ORFs were ordered according to their expression distance to the metabolic gene. The ordering was used to calculate the rank of the metabolically adjacent genes. Distribution of these ranks is shown. Mean rank of all adjacent genes is 2530 (dashed line, marked ‘all’). Mean rank of the closest adjacent gene is 759 (dashed line, marked ‘best’); furthest adjacent genes rank, on an average, at 4405 (‘worst’). The inset shows the same distribution, on a different scale. (**b**) Fractions of the metabolically adjacent genes ranked within different thresholds.

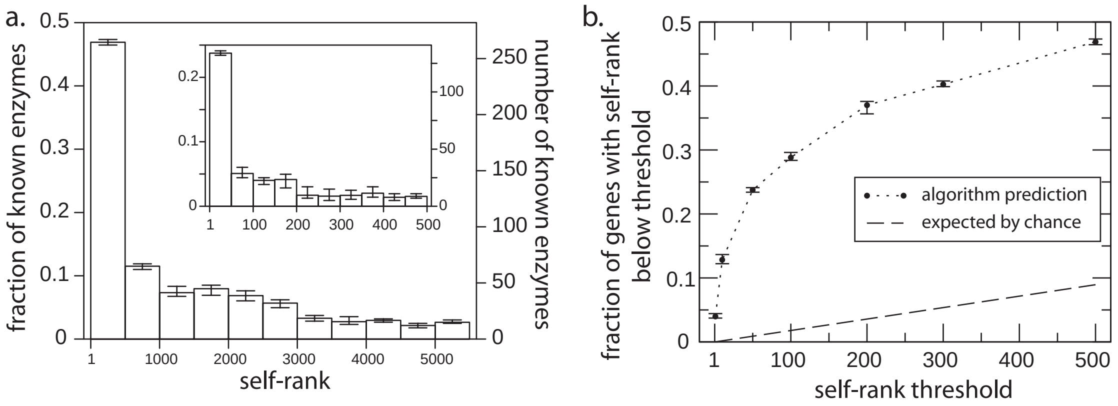

Fig. 4. Validation of known *S.cerevisiae* metabolic genes. (**a**) Distribution of self-ranks for known yeast metabolic genes, as predicted by the placement algorithm. Error bars correspond to the 95% confidence intervals determined by non-parametric bootstrap. (**b**) Fraction of known yeast metabolic genes that were self-ranked within different thresholds. For comparison, the fraction expected from a random candidate ordering is shown by the dashed line.

value of the neighborhood weight vector $\vec{w}$, describing the relative contributions of neighborhood layers was found to be (0.892, 0.106, 0.00126), with the 95% confidence intervals of $(0.803 \leftrightarrow 0.970, 0.0251 \leftrightarrow 0.177, 4.54{\cdot}10^{-5} \leftrightarrow 2.63{\cdot}10^{-3})$. The small value of w_3 indicates that the expression behavior at the distance 3 is no longer informative, even though on an average, it is still significantly more correlated than the rest of the network.

The distribution of self-ranks generated by the MEP algorithm is shown in Figure 4a. The distribution exhibits a large peak at low ranks, indicating that the algorithm is capable of predicting a substantial fraction of known metabolic enzymes. Out of 564 known metabolic genes, the method identifies 23 enzymes (with the 95% confidence interval of 21 ↔ 25) as top candidates (self-rank of 1) for their enzymatic function; 72 enzymes (95% confidence within 69 ↔ 77) are identified to be within the top 10 candidates, compared with 1 enzyme expected by chance; 134 enzymes (95% confidence within 132 ↔ 136) are identified within the top 50 candidates, with only 5 enzymes expected by chance. The fraction of metabolic genes ranking within different self-rank thresholds is shown in Figure 4b.

3.4 The performance impact of the expression variability

The shape of the self-rank distribution at high ranks is relatively flat (Fig. 4a), suggesting that for a large fraction of known enzymes evaluation of the cost function was not at

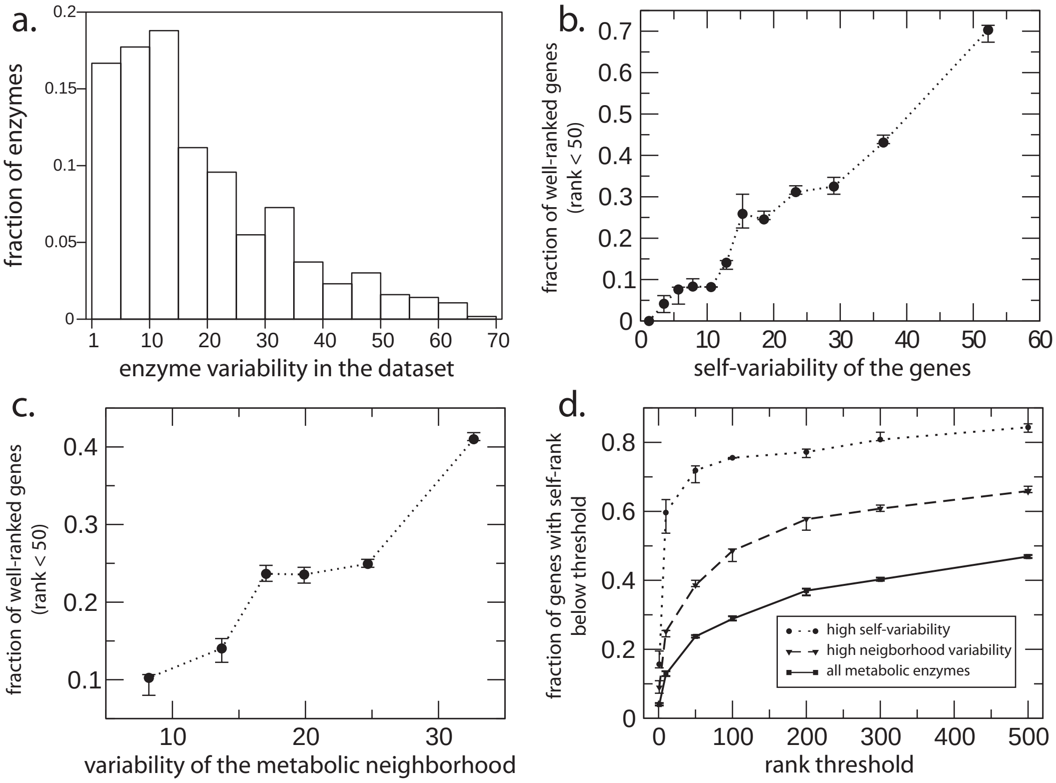

Fig. 5. Effects of expression variability. (**a**) Distribution of the individual enzyme variability in the Rosetta compendium dataset. (**b**) Dependency of the MEP algorithm's predictive ability on the variability of the target gene. Known metabolic enzymes were binned according to their variability in groups of 49 genes, and the fraction of the enzymes predicted by the placement algorithm (within top 50 candidates) is shown for every bin. (**c**) Dependence on the variability of the network neighborhood. Known enzymes were binned according to variability of their network neighborhood into groups of 98 genes, and a fraction of genes predicted to be within the top 50 candidates is shown. (**d**) Predictive ability of the algorithm is compared for all known enzymes, for the enzymes with high network neighborhood variability (above 30), and for the enzymes with high self-variability (above 45). The fraction of predicted enzymes is shown for different self-rank thresholds. Error bars correspond to the 95% confidence intervals determined by the non-parametric bootstrap.

all informative of their metabolic role. In other words, no correspondence could be detected between the expression profile of the corresponding gene and the expression of its network neighborhood. While this is expected from the enzymes whose metabolic activity is regulated primarily through post-transcriptional mechanisms, a failure to detect expression correspondence is more likely to be reflective of the relatively uninformative expression dataset. For example, if a certain region of the metabolism has not been perturbed in any of the conditions comprising the dataset, no information can be gained on the coexpression properties of genes within that region.

To analyze the effects related to the information content of the expression dataset, we introduce a gene variability measure, equal to the number of experiments in which a given gene has been perturbed with high probability (see Methods section). Low variability score indicates that a gene has been perturbed rarely, if at all, in the given expression dataset, and therefore, there is little information available to compare its expression profile with the profiles of other genes. An expression variability distribution of the metabolic genes, presented in Figure 5a, shows that the majority of the metabolic genes are significantly perturbed (with P-value below 0.1) in less than 19 out of 300 conditions measured by the Rosetta dataset, while some metabolic genes (5%) are significantly perturbed in over 50 conditions.

As expected, the ability of the MEP algorithm to identify the metabolic enzymes depends considerably on the variability of the corresponding gene in the expression dataset (Fig. 5b). While the probability of identifying a poorly perturbed metabolic gene within the top 50 candidates is below 10%, the probability of identifying a highly perturbed metabolic gene is $\sim$70%. This demonstrates that the proposed algorithm can be very effective in identifying metabolic enzymes given an informative dataset.

In a realistic application of the method, one would require a way of assessing the potential performance of the algorithm without knowing the true identity of the metabolic gene.

Because candidate genes are ranked by matching their expression profiles with that of the metabolic network neighborhood, the quality of the prediction should be related to the variability of genes in the neighborhood. Indeed, we find that increased network neighborhood variability (see Methods section) also accounts for an improvement in the self-ranking performance (Fig. 5c), although the magnitude of the improvement is smaller than for the variability of the target gene. The probability of identifying a metabolic enzyme within the top 50 candidates is ∼15% for the neighborhoods with poor variability, and ∼40% for the highly variable neighborhoods. The performance of the algorithm in identifying well-perturbed metabolic genes, genes with high variability of the metabolic neighborhood, and all known metabolic genes is compared in Figure 5d. This indicates that performance of the proposed method can be improved by increasing the amount of relevant expression information, for example by utilizing an expression dataset with conditions that specifically address the functions performed by the metabolic neighborhood of the target gene of interest.

4 CONCLUSIONS

With the rapid accumulation of completely sequenced genomes, there has been an increased reliance on computational methods to reconstruct metabolic networks. Sequence similarity and genome context techniques have proved highly effective at assigning enzymatic functions to genes in newly sequenced organisms. The emphasis has, therefore, shifted towards the problem of completing metabolic models by identifying the enzymes missing from the metabolic network (Osterman and Overbeek, 2003). Here, we present a strategy for identifying the missing genes by combining gene expression analysis with the structural information provided by partial reconstruction of the metabolic network.

The MEP algorithm was validated using a recently published *S.cerevisiae* metabolic network (Forster *et al*., 2003). We show, that a substantial fraction (>20%) of known enzymes can be predicted within the top 50 out of 5594 candidates for their enzymatic function. The predictive power of the method critically depends on the amount of information provided by the expression dataset. We show that the method is capable of predicting (within the top 50 candidates) ∼70% of the metabolic enzymes that have been significantly perturbed across multiple conditions of the Rosetta compendium dataset. It is important to emphasize that our approach should be used in conjunction with the array of genome context analysis techniques currently being used to identify the missing metabolic genes.

The approach also represents an example of how expression information can be used to make predictions on the level of a specific gene function. It will be interesting to test the predictive ability of this algorithm by combining it with other new expression analysis techniques (van Noort *et al*., 2003). MEP uses available structure of the metabolic network to enhance the predictive capability of the expression data. A similar strategy can also be applied to gene context methods. For example, it will be interesting to evaluate candidate genes for a missing enzymatic function by analyzing genome co-occurrence profiles of the enzyme-encoding genes in the metabolic network neighborhood, or by evaluating clustering of these genes on the chromosome. As improvements in sequence homology and genome context methods enhance reconstruction of metabolic networks, the approaches that directly take into account network structure will become increasingly important.

REFERENCES

Bishop,A.C., Xu,J., Johnson,R.C., Schimmel,P. and de Cricy-Lagard,V. (2002) Identification of the tRNA-dihydrouridine synthase family. *J. Biol. Chem.*, **277**, 25090–25095.

Bobik,T.A. and Rasche,M.E. (2001) Identification of the human methylmalonyl-CoA racemase gene based on the analysis of prokaryotic gene arrangements. Implications for decoding the human genome. *J. Biol. Chem.*, **276**, 37194–37198.

Bono,H., Nikaido,I., Kasukawa,Y., Hayashizaki,Y. and Okazaki,Y. (2003) Comprehensive analysis of the mouse metabolome based on the transcriptome. *Genome Res.*, **13**, 1345–1349.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Devos,D. and Valencia,A. (2001) Intrinsic errors in genome annotation. *Trends Genet.*, **17**, 429–431.

Efron,B. and Tibshirani,R.J. (1993) *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Forster,J., Famili,I. Fu,P., Palsson,B.O. and Nielsen,J. (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.*, **13**, 244–253.

Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A.,Coffey,E., Dai,H., He,Y.D. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126.

Iliopoulos,I., Tsoka,S. Andrade,M.A., Enright,A.J., Carroll,M., Poullet,P., Promponas,V., Liakopoulos,T., Palaios,G., Pasquier,C. *et al*. (2003) Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics*, **19**, 717–726.

Iliopoulos,I., Tsoka,S. Andrade,M.A., Janssen,P., Audit,B., Tramontano,A., Valencia,A., Leroy,C., Sander,C. and Ouzounis,C.A. (2001) Genome sequences and great expectations. *Genome Biol.*, **2**, Interactions0001.

Ito, T., Chiba,T., Ozawa,R., Yoshida,M., Mattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci., USA*, **98**, 4569–4574.

Kaern, M., Blake,W.J. and Collins,J.J. (2003) The engineering of gene regulatory networks. *Annu. Rev. Biomed. Eng.*, **5**, 179–206.

Karp,P.D. (1998) Metabolic databases. *Trends Biochem. Sci.*, **23**, 114–116.

Kunin,V. and Ouzounis,C.A. (2003) The balance of driving forces during genome evolution in prokaryotes. *Genome Res.*, **13**, 1589–1594.

Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Einsenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

Nelder, A. and Mead,R. (1965) A simplex method for function minimization. *Comput. J.*, **7**, 308–313.

Osterman,A. and Overbeek,R. (2003) Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.*, **7**, 238–251.

Overbeek,R., Fonstein,M. D'souza,M., Push,G.D. and Maltseze,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci., USA*, **96**, 2896–2901.

Overbeek,R., Larsen,N., Pusch,G.D., D'souza,M., Selkov,E.,Jr, Kyrpides,N., Fonstein,M., Maltsev,N. and Selkov, E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**, 123–125.

Pellegrini,M., Marcotte,E.M., Thompson, M.J., Einsenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci., USA*, **96**, 4285–4288.

Press,W.H., Teukolsky,S.A., Wetterling,W.T. and Flannery,B.P. (eds) (2002) *Numerical Recipes in C++: the Art of Scientific Computing*. Cambridge University Press, Cambridge, MA.

Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M. Pochart,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae. Nature*, **403**, 623–627.

van Noort,V., Snel,B. and Huynen,M.A. (2003) Predicting gene function by conserved co-expression. *Trends Genet.*, **19**, 238–242.

Wen,X., Fuhrman,S., Michaels,G.S., Carr,D.B., Smith,S., Barker,J.L. and Somogyi,R. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci., USA*, **95**, 334–339.

Wu,L.F., Hughes,T.R., Daviarwala,A.P., Robinson,M.D., Stoughton,R. and Altschuler,S.J. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.*, **31**, 255–265.

Yanai,I., Wolf,Y.I. and Koonin,E.V. (2002) Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol.*, **3**, RESEARCH0024.