Molecular Systems Biology 5; Article number 263; doi:10.1038/msb.2009.22
Citation: ***Molecular Systems Biology*** **5:263**

www.molecularsystemsbiology.com

molecular systems biology

REPORT

Predicting metabolic biomarkers of human inborn errors of metabolism

Tomer Shlomi[1,4,*], Moran N Cabili[2,4,*] and Eytan Ruppin[2,3,*]

[1] Department of Computer Science, Technion—Israel Institute of Technology, Haifa, Israel, [2] School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel and [3] Department of Physiology and Pharmacology, School of Medicine, Tel-Aviv University, Tel-Aviv, Israel
[4] These authors contributed equally to this work
* Corresponding authors. T Shlomi, Department of Computer Science, Technion—Israel Institute of Technology, Haifa 32000, Israel. Tel.: +972 4 829 4356; Fax: +972 4 829 3900; E-mail: tomersh@cs.technion.ac.il or MN Cabili, School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. Tel.: +972 545 860 060; Fax: +972 3 640 9357; E-mail: natalym@post.tau.ac.il or E Ruppin, School of Computer Science and School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel. Tel.: +972 3 640 6528; Fax: +972 3 640 9357; E-mail: ruppin@post.tau.ac.il

Received 24.7.08; accepted 25.2.09

Early diagnosis of inborn errors of metabolism is commonly performed through biofluid metabolomics, which detects specific metabolic biomarkers whose concentration is altered due to genomic mutations. The identification of new biomarkers is of major importance to biomedical research and is usually performed through data mining of metabolomic data. After the recent publication of the genome-scale network model of human metabolism, we present a novel computational approach for systematically predicting metabolic biomarkers in stochiometric metabolic models. Applying the method to predict biomarkers for disruptions of red-blood cell metabolism demonstrates a marked correlation with altered metabolic concentrations inferred through kinetic model simulations. Applying the method to the genome-scale human model reveals a set of 233 metabolites whose concentration is predicted to be either elevated or reduced as a result of 176 possible dysfunctional enzymes. The method's predictions are shown to significantly correlate with known disease biomarkers and to predict many novel potential biomarkers. Using this method to prioritize metabolite measurement experiments to identify new biomarkers can provide an order of a 10-fold increase in biomarker detection performance.
Molecular Systems Biology 28 April 2009; doi:10.1038/msb.2009.22
Subject Categories: cellular metabolism; molecular biology of disease
Keywords: constraint-based modeling; disease biomarkers; human metabolism; inborn error of metabolism

Introduction

The study of genetic metabolic disorders originated in the early 1900s with Sir Archibald Garrod's discovery of the first inborn errors of metabolism (IEM), alkaptonuria, pentosuria, cystinuria, and albinism (Vangala and Tonelli, 2007). IEM are caused by alterations of specific metabolic reactions and a few hundreds of IEM affecting about 1 in every 5000 born babies are currently characterized (Lanpher *et al*, 2006). Fortunately, recent advances in metabolomic approaches enable an expanded newborn screening that improves early diagnosis and treatment in numerous IEM disorders (Lanpher *et al*, 2006).

Metabolomics, the study of the complete repertoire of small molecules in cells, tissues and biological fluids, represents a major and rapidly evolving research field in systems biology. It has been fueled by the development of experimental platforms such as gas chromatography and liquid chromatography-based mass spectrometry that are capable of accurately measuring hundreds of small molecules in biological samples (Kaddurah-Daouk *et al*, 2008). These methods promise to substantially advance our understanding of disease pathophysiology and advance the discovery of new diagnostic biomarkers for disease. Such metabolic biomarkers denote sets of metabolites that show a consistent change in concentration during a disease state and are hence effective diagnostic means. Metabolomics offers several advantages over genomics and proteomics as a tool for diagnosing and understanding disease (Vangala and Tonelli, 2007): (1) Metabolic

biomarkers can be measured noninvasively in biofluids such as plasma, urine and feces in a rather straightforward manner (Seegmiller, 1968). (2) Changes in metabolite levels reflect the actual metabolic state of a tissue and its histology, translating the genotype and environmental factors into the phenotype (Nicholson *et al*, 2002) (Urbanczyk-Wochniak *et al*, 2003). (3) There is a relatively small number of biomarkers (~2500–3000). Several IEM have already been characterized through the identification of clinical metabolic biomarkers that can explain the pathological phenotype (Ito *et al*, 2000). However, with the expected surge in the scope and quality of metabolomic measurements, metabolomics is destined to play an even more central role in the near future as an efficient diagnostic tool and as a safety evaluator of drug candidates.

The recent publication of the first Human metabolic network model (Duarte *et al*, 2007), along with the detailed documentation of all known IEM in the OMIM database (McKusick, 2007), have given us an opportunity to systematically predict potential metabolic disease biomarkers on a large scale. Herewith, we propose a new computational approach that predicts, for each metabolic gene, a set of metabolites that are expected to show a concentration change in biofluids after its knockout. The method is based on the constraint-based modeling (CBM) approach, which is commonly used to predict metabolic phenotypes in microorganisms (Price *et al*, 2004) and specifically the effects of gene knockouts (Segre *et al*, 2002; Stelling *et al*, 2002; Shlomi *et al*, 2005). Recently, CBM has been used to predict human tissue-specific metabolism (Duarte *et al*, 2007; Shlomi *et al*, 2008). Our approach differs from earlier attempts to identify disease biomarkers, which mostly use data mining techniques that analyze metabolomics data taken from healthy and diseased subjects (Wagner *et al*, 2004; Yang *et al*, 2004; Kenny *et al*, 2005). As it is model based, the current approach permits the prediction of large sets of diagnostic biomarker patterns for many disorders, laying down a computational parallel to the upcoming advances in metabolomic measurements in biofluids. As a first validation of our method, we apply it to predict changes in metabolite concentrations due to dysfunctional enzymes in red-blood cell (RBC) metabolism, whose dynamic behavior can be reliably simulated through a kinetic model (Jamshidi *et al*, 2001). Then, the method is applied to the genome-scale human metabolic network model of Duarte *et al*, and its performance is comprehensively evaluated based on various sets of known biomarker extracted from different databases.

Results and discussion

A constraint-based approach for predicting metabolic biomarkers

We present a new computational approach for systematically predicting the pattern of metabolic biomarkers characterizing each metabolic disorder whose causative gene is included in the human metabolic network model (Duarte *et al*, 2007). Let a *boundary metabolite* denote a metabolite that is known to be taken-up or secreted between the intracellular and extracellular compartments (as indicated in the network model). Let an *exchange interval* denote a possible range of uptake and secretion fluxes of a given boundary exchange interval. For each metabolic disorder and each boundary metabolite, we predict its exchange interval between human tissues and biofluids, for both healthy and disease cases (Materials and methods). This exchange interval is computed through a CBM method called flux variability analysis (FVA) (Mahadevan and Schilling, 2003), which accounts for the entire space of feasible flux states that satisfy mass-balance stoichiometric constraints and reaction directionality constraints (embedded in the model of (Duarte *et al*, 2007)). For the healthy case, the exchange interval is computed while the reactions affected by the disease are constrained to be active, whereas for the disease case, they are constrained to be inactive. By comparing the predicted exchange interval between the healthy state and the disease state for each boundary metabolite, one can determine whether the pertaining boundary metabolite concentration in biofluids (termed *biomarker*) is expected to be *elevated*, *reduced* or *unchanged* (see Materials and methods). If the predicted changes are marked such that there is no overlap between the exchange intervals of the healthy case and the disease case, the predicted biomarker change is considered to be *highly confident*.

An illustrative example of the predicted biomarker changes' ranges and their underlying rationale for the healthy state and some disease state is depicted in Figures 1A and B. The predicted exchange intervals of metabolite *M1* (*M2*) suggest that its extracellular concentration is elevated (reduced) in the disease case. The disjoint exchange intervals obtained for the healthy case and the disease case for both *M1* and *M2* render these predictions as highly confident. The exchange intervals of metabolite *M6* (*M4*) suggest that their extracellular concentrations are elevated (reduced) in the disease case. Examining, for example, the exchange interval of metabolite *M6* shows that in the healthy case, *M6* can be either taken-up from biofluids or secreted in a lower rate (as some of it is required in the healthy state; Supplementary Figure 1). In the disease case, *M6* (synthesized through *M5*) can only be secreted to biofluids. It should be noted that mass-balance stoichiometric constraints that play an important role in determining the exchange intervals of different metabolites and are accounted for by the CBM method (and as will be shown, play an important role in determining biomarker changes in addition to the network topology) are not depicted in this kind of illustration.

Validating the biomarker prediction method through a small-scale kinetic model of RBC metabolism

As a first validation of our method, we applied it to predict metabolic biomarkers for enzyme deficiencies in human erythrocytes, for which a detailed kinetic model (Jamshidi *et al*, 2001) is readily available to simulate the dynamic metabolic behavior after enzymatic perturbations. This kinetic model consists of four basic classical pathways: glycolysis, the pentose pathway, adenosine nucleotide metabolism, and the Rapoport-Leubering shunt, accounting for 43 metabolites, 43 internal reactions, and 12 primary exchange reactions. We applied this model to predict changes in extracellular metabolite concentrations after a disruption to 43 enzyme-

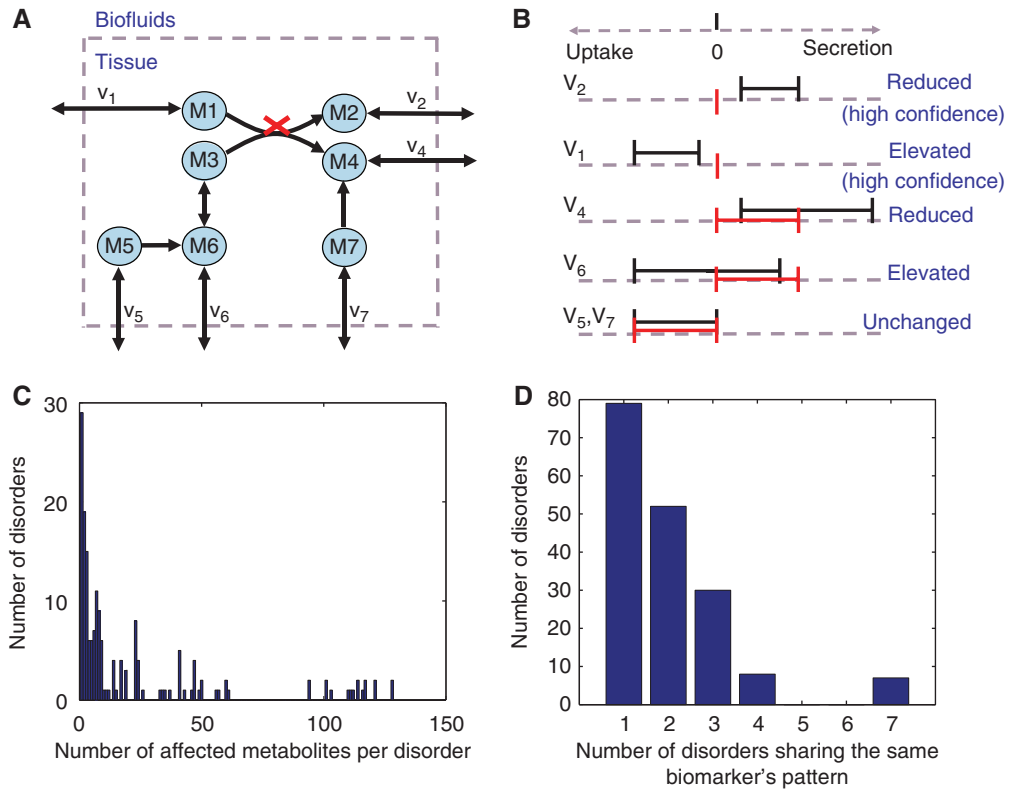

Figure 1 An illustrative example of the prediction of biomarker concentration changes. (**A**) Circular nodes represent metabolites, solid edges represent reactions. The disease causing reaction is marked with a red cross. Out of six boundary metabolites in this network, only four metabolites (*M1*, *M2*, *M4*, *M6*) are predicted to show a concentration change when the disease causing reaction is inactivated. (**B**) Concentration change predictions based on exchange interval comparisons: the healthy state and disease state exchange intervals are colored black and red, respectively. Positive flux values represent metabolite secretion, whereas negative values represent metabolite up-take. For example, the concentration of *M2* (associated with *V2*) is predicted to be reduced with high confidence due to a substantial change in the exchange interval. Similarly, *M1* is predicted to be elevated with high confidence. *M4* is reduced in the disease state as it must be secreted in the healthy case but is only potentially secreted in the disease case. The concentration level of *M5* and *M7* is predicted to be unchanged between the healthy case and the disease case. (**C**) The distribution of the number of the predicted alterations among the 176 disorders analyzed. (**D**) The distribution of predicted biomarker alteration patterns that are jointly shared by a number of disorders. As shown, various disorders tend to have different sets of biomarkers (the histogram is skewed to the left).

catalyzed reactions in the model (i.e. simulating changes in the steady-state behavior by iteratively solving the set of differential equations specified in the model; Jamshidi *et al*, 2001; see Materials and methods). These simulations resulted in a set of 156 metabolic biomarkers whose extracellular concentration is either elevated or reduced in a perturbed steady-state behavior.

Earlier studies have shown that CBM of RBC metabolism correctly captures various aspects of the metabolic behavior simulated through a kinetic model (Wiback and Palsson, 2002; Durmus Tekir *et al*, 2006). Applying our constraint-based method for the RBC model to predict changes in extracellular metabolites after the knockouts of model reactions resulted in a set of 85 biomarker predictions (see Materials and methods). These biomarker predictions are significantly correlated with the kinetic simulation results (hypergeometric P-value $< 7 \times 10^{-11}$, comparing the predicted accuracy with a random model), obtaining a precision of 0.73 (fraction of the predicted biomarkers that are correct) and recall of 0.40 (fraction of the biomarkers that are correctly predicted; Table I). This result testifies to the ability of our method to correctly identify alterations in extracellular metabolite concentrations, relying solely on reaction stoichiometry and directionality data.

Table I Prediction accuracy of the biomarker prediction method based on a comparison with predictions obtained through kinetic model of red-blood cell metabolism, and based on comparisons with various experimental datasets

	Precision	Recall	*P*-value
RBC Kinetic Model	0.73	0.40	$< 7 \times 10^{-11}$
OMIM: Automatic	0.37	0.27	$< 7 \times 10^{-12}$
OMIM: Manual	0.76	0.56	$< 4 \times 10^{-13}$
Ramedis/HMDB	0.41	0.1	$< 5 \times 10^{-5}$

A detailed comparison of the predictions versus the known biomarkers is available in Supplementary Dataset 1.

Large-scale prediction and validation of biomarkers for human metabolic disorders

We applied our method to the human metabolic network model (Duarte *et al*, 2007) to predict biomarker changes for 304 metabolic disorders (documented in the OMIM database) whose causative genes are included in the model (see Materials and methods). The analysis resulted in a total of 3912 predictions of biomarkers' changes involving 233 boundary metabolites (whose concentration is predicted to change in at least a single disease), and 176 diseases (for which at least a single biomarker change is predicted). Out of all

biomarker alteration predictions, 19% are with high confidence. A high fraction of the disorders (42%) are predicted to have very few biomarker changes (<6), whereas up to 61% of the disorders are predicted to have <10 biomarkers (Figure 1C).

The various disorders tend to have different sets of biomarkers (e.g. only in a very few cases, the same set of biomarkers correspond to more than three diseases; Figure 1D), suggesting that large-scale biofluids' metabolomics may be effectively used for the diagnosis of metabolic disorders. Notably, the majority of the predicted biomarkers (129 metabolites out of 175 for which this information is available) are known to be present in the blood and urine and are hence readily available media for metabolomic analysis (Human Metabolome Database (HMDB); Wishart *et al*, 2007; Supplementary Figure 2).

To systematically validate the predictions, we extracted biomarker data for all metabolic disorders documented in the OMIM database, whose causing genes are included in the model of Duarte *et al* (by automatic parsing of disease description texts in the OMIM database; see Materials and methods). Comparing this dataset with our predictions showed a highly significant correlation (hypergeometric P-value$<7\times10^{-12}$), though with precision and recall levels lower than those described above versus the kinetic RBC model (Table I). Still, this correlation is quite remarkable, considering the erroneous nature of the simple text-mining method that underlies this validation dataset.

To derive a more refined biomarker dataset for validation, we manually extracted biomarker data from the OMIM database for a set of 17 inborn errors of amino-acid metabolism (see Materials and methods). This manual curation process permits a finer tuned resolution of problems arising in the validation data, for example, from differences in metabolite naming conventions. Furthermore, it allows one to extract data on the precise specific enzymatic reactions that are affected in each metabolic disorder, in cases where mutations may disrupt the activity of multifunctional genes. A comparison of this dataset to the model predictions again demonstrated a fairly accurate level of prediction (hypergeometric P-value$<4\times10^{-13}$, precision=0.76, recall=0.56; Figure 2), close to that shown above versus the kinetic RBC model (Table I).

To further validate our prediction method, we extracted biomarker data for a set of 29 rare metabolic disorders from the Rare Metabolic Disease database (Ramedis; Töpel *et al*, 2006), recording clinical measurements of metabolite levels in biofluids, and from HMDB (Wishart *et al*, 2007; see Materials and methods). Comparing the set of predicted biomarkers alterations to this clinical dataset exhibited a highly statistically significant overlap (hypergeometric P-value$<5\times10^{-5}$), with moderate precision (0.41) but rather low recall (0.1) (Table I). This lower accuracy level may result in part from the lower quality of the pertaining clinical data; the latter is prone to the influence of several nondisorder-specific factors (e.g. the medical treatment the patient was subject to, or his nutritional state). Indeed, cross-referencing these clinical biomarkers with those reported by OMIM shows a rather low overlap between the two different kinds of data sources (50% of the biomarker-diseases associations marked in OMIM are found in

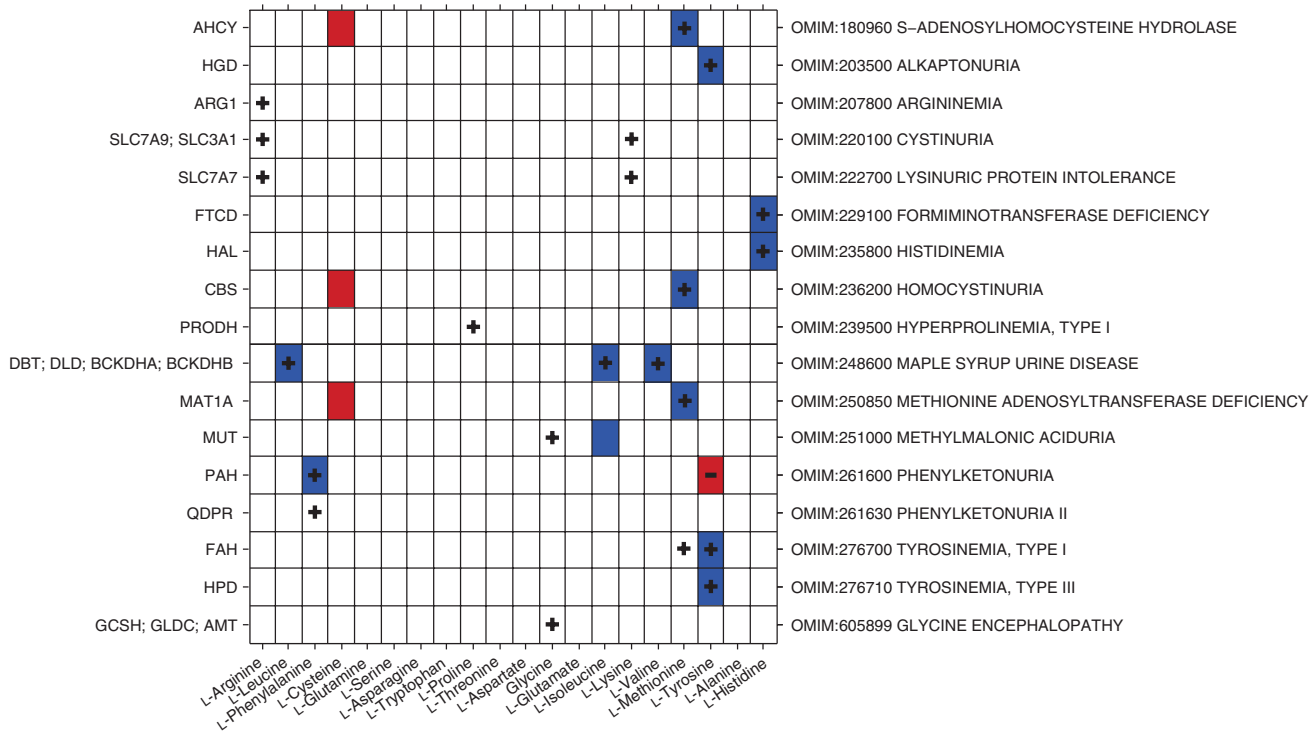

Figure 2 Prediction of amino-acid biomarkers for a set of amino-acid metabolic disorders. Rows represent metabolic disorders and columns represent amino acids. The causative gene's name is indicated on the left. Blue and red entries represent biomarkers that are predicted by our method to be elevated or reduced, respectively. Table entries marked in '+' or '−' represent elevation or reduction in the metabolite's concentration in biofluids according to OMIM, respectively.

Ramedis-HMDB, but only 19% of the associations marked in Ramedis-HMDB are found in OMIM). Notably, the agreement between the different pertaining databases is of the order of their agreement with the model's predictions.

In-depth inspection of candidate biomarker predictions

To explore the added value of our biomarker prediction method over naïve visual inspection of the network topology, we manually inspected the corresponding network regions affected by errors in amino-acid metabolism. We found that although in a few cases it might have been possible to correctly predict biomarkers just by observing the perturbed pathway in a topological map representation of the human metabolic network, in other cases such a simple topologically based inference of a biomarker alteration would fail and lead to false predictions. Thus, indeed, a method, accounting for the network dependencies between pathways along with stoichiometry and reaction directionality constraints is of value. For instance, in methionine adenosyltransferase deficiency (OMIM: 250850), the potential elevation of methionine in biofluids can be predicted in a straightforward manner as the only reaction that catabolizes methionine is inactivated in this disorder. However, in the cases of homocystinuria, hypermethioninemia, and tyrosinemia that we discuss next, the network topology alone fails to identify the correct biomarkers.

An example of a correctly predicted biomarker that is difficult to infer simply by observing the network topology occurs in the case of homocystinuria, caused by the deficient activity of Cystathionine β-synthase (CBS; E.C 4.2.1.22; converting homocysteine and serine to cysteine; Figure 3). In the healthy case, when CBS is functional, methionine is taken-up from biofluids and is eventually converted to cysteine by series of enzymes that includes CBS. Our method predicts that the biofluids' concentration of methionine is elevated (with high confidence) in homocystinuria as reported in OMIM, and that the concentration of cysteine is reduced in the extracellular as reported in (Lee and Briddon, 2007) (Figure 2). In this case, inferring the elevated extracellular concentration of methionine simply based on the network topology is impossible. This is because of the existence of an alternative cyclic pathway that metabolizes methionine in homocystinuria (the methionine salvage pathway; Cellarier *et al*, 2003), but in fact, cannot change the methionine extracellular concentration due to mass-balance constraints. A similar scenario pertains to the predicted elevated extracellular concentration of methionine in hypermethioninemia caused by *S*-adenosylhomocysteine hydrolase deficiency (AHCY; E.C 3.3.1.1; OMIM: 180960; Figure 3). Overall, although it is indeed difficult to make such predictions *a priori* by inspecting the network topology, such inspections can be telling and informative *a posteriori* (when pointed to by the stochiometric analysis). Other interesting examples are the cases of tyrosinemia type I, type III, and Alkaptonuria (OMIM 276700, 276710, 203500, respectively), each caused by the dysfunctional behavior of one out of five tyrosine degradation pathways. A simple observation of the network topology may suggest that the existence of several alternative tyrosine

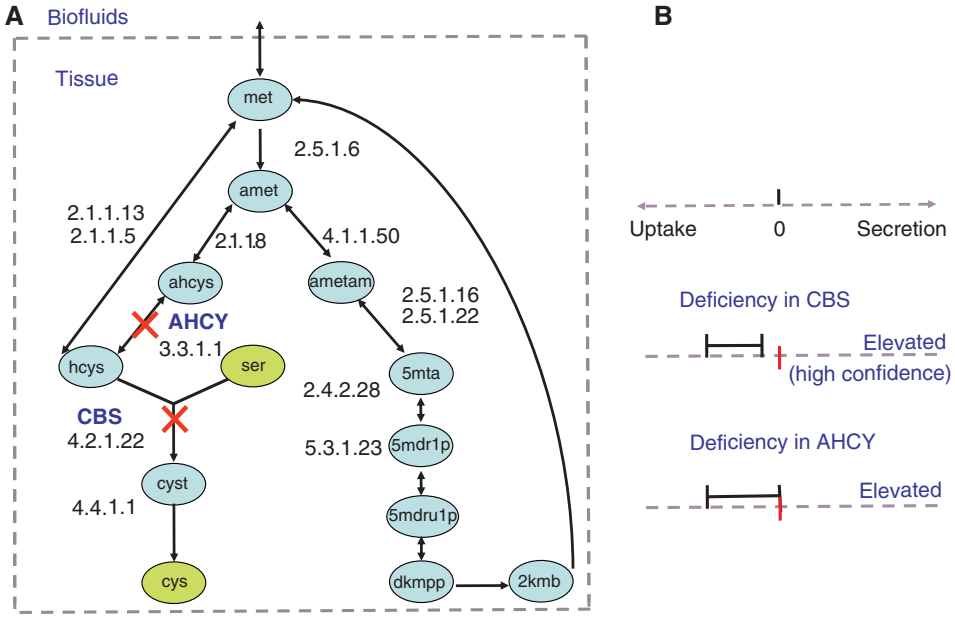

Figure 3 (**A**) A subnetwork that illustrates the effect of homocystinuria on the metabolism and transport of methionine. Circular nodes represent metabolites and edges represent biochemical reactions. For simplicity, only abbreviations of metabolite names and enzyme E.C (Enzyme Commission) numbers are specified (explicit names are given in Supplementary Table 1). Metabolites marked in green participate in other reactions that are not presented here for simplicity. Homocystinuria is caused by a dysfunctional CBS, and hypermethioninemia is caused by dysfunctional AHCY. (**B**) Prediction of concentration changes of the boundary metabolite methionine in homocystinuria and *S*-adenosylhomocysteine hydrolase deficiency based on interval comparison of its exchange reaction's flux. In both cases, inferring the elevated concentration of methionine simply based on the network topology is impossible, due to the cyclic methionine salvage pathway (involving reaction 4.1.1.50), which (though topologically plausible) cannot affect methionine concentration due to mass-balance constraints.

degradation pathways may suffice to compensate for the inactivity of a single degradation pathway, and keep the pertaining metabolites' extracellular concentrations at bay (Supplementary Figure 3). However, it turns out that due to mass-balance stoichiometric considerations the network cannot compensate for the inactivity of one degradation pathway by routing more substrates to the alternative active pathways, and, indeed, the elevated extracellular concentrations of tyrosine are correctly predicted by our method in all three diseases (Figure 2).

Although our method successfully identifies a large set of the known biomarkers in a statistically significant manner, it failed to identify a considerable number of others. Some of the false predictions are expected to result from incompleteness of the metabolic network as well as several simplifying assumptions that underlie our computational method and enable the large-scale analysis of a network with thousands of reactions. Specifically, our method assumes a steady-state metabolic behavior (due to the lack of global enzyme kinetic data), and lacks regulatory constraints, which obviously have substantial influence on the activation of alternative pathways under different physiological conditions (but are still largely missing for the human case). In fact, the lack of cell type- and tissue-specific regulatory constraints restricts the analysis to a system that integrates the metabolic behavior of different tissues. The latter prevents the prediction of cross-tissue mechanisms that affect metabolite biofluid concentration (e.g. a metabolite secreted in one tissue and taken-up by another). However, although the method relies on several simplifying assumptions, its predicted biomarkers are significantly correlated with all tested validation datasets, with the precision and recall varying between 0.37–0.76 and 0.1–0.56, respectively, depending on the quality of the dataset. The probability that a biomarker prediction made by our method would turn out to be correct is between 6–15.8 times higher than random, across the various validation datasets (based on random generation of biomarker sets and a comparison of their prediction accuracy with our method). This suggests that using our method to prioritize metabolite measurement experiments can provide an order of a 10-fold increase in biomarker detection performance. This result is remarkably encouraging in face of the obvious model simplifications discussed above.

An in-depth investigation of the false predictions can be of significant value for further improving the metabolic network model and the analysis methods. For example, in the case of methylmalonate semialdehyde dehydrogenase deficiency (OMIM 603178), in which a reaction (E.C 1.2.1.27) in the valine and pyrimidine catabolic pathways is dysfunctional, our method falsely predicts the potential elevation of valine in biofluids instead of the elevation of its catabolic product 3-hydroxyisobutyric acid, as well as 3-hydroxypropionic acid and β-alanine. An inspection of the underlying network topology reveals that this false prediction results from missing membrane transporters for these products and their related derivatives, which prevents their exchange with surrounding biofluids and hence indirectly limits the possible uptake of their upstream substrate, valine (Supplementary Figure 4). Incorporating a putative transporter of 3-hydroxyisobutyric acid from mitochondria to cytoplasm and from there to the extracellular environment in the model, enables our method to correctly predict the increased secretion rate of hydroxyisobutyric acid in this disorder (instead of the increased concentration of its upstream substrate, valine). Similarly, the inclusion of a transporter for 3-hydroxyisobutyric acid and a mitochondrial transporter for its derivative, malonate semialdehyde, leads to a prediction of decreased uptake rate of 3-hydroxyisobutyric in these disorders (reflecting an increased biofluid concentration), and to correctly predicting the elevated concentration of β-alanine. These results suggest a future application of the biomarker prediction method for automatically identifying missing reactions in the model, in line with a previous approach for refining genome annotation (Reed *et al*, 2006).

In other cases, the lack of regulatory constraints leads to false predictions. For example, the model fails to predict the elevation in extracellular concentration of arginine in argininemia (OMIM 207800). The latter elevation is caused by a dysfunctional type I arginase (ARG1; E.C 3.5.3.1), which disrupts the conversion of arginine to urea and ornithine. This prediction failure arises from the existence of several additional alternative pathways that catabolize arginine (e.g. converting it either to *N*-hydroxyarginine, guanidinoacetate, and ornithine, or transporting it to the mitochondria) and maintain the same predicted uptake rate of arginine when ARG1 is dysfunctional. In reality, arginine catabolism through the mitochondria (through the type II arginase ARG2) is likely not to be able to fully compensate for the loss of ARG1 due to the low expression level of ARG2 compared with that of ARG1 in the liver (Levillain *et al*, 2005; Cline *et al*, 2007). Specifically, these erroneous predictions are probably caused by the current lack of tissue-specific regulatory constraints. Such prediction errors may be corrected by incorporating regulatory constraints within the model (Covert *et al*, 2004), or by incorporating tissue-specific expression data (Akesson *et al*, 2004; Duarte *et al*, 2007; Shlomi *et al*, 2008).

Although this paper focused on predicting metabolic biomarkers for known IEM, we additionally applied our approach to predict biomarkers for the knockouts of all other genes present in the metabolic model, that is, those that are not known to cause metabolic disorders but some may potentially be discovered to do so in the future. This analysis, covering an additional set of 872 genes from the model of Duarte *et al*, resulted in a total set of 9567 biomarkers alterations, which are available for inspection and future validation as Supplementary Dataset 1. To provide means for further studying the mechanisms by which a biomarker's extracellular concentration is altered, we created network visualizations (through the Cytoscape tool (Cline *et al*, 2007)) of the metabolic alterations resulting from the knockout of each gene in the network. These are available for download from the supplemental website: www.cs.tau.ac.il/~shlomito/biomarkers.

In summary, this study presents a generic approach for the large-scale prediction of specific biomarkers that are elevated or reduced in biofluids. Future work should aim at extending the model to include additional metabolic pathways, for example, by integrating it with other large-scale networks such as the Edinburgh human metabolic network reconstruction (Ma *et al*, 2007). On the long run, an integrated human metabolic-regulatory reconstruction in the lines of that of

(Covert *et al*, 2004) will further improve the predictions in cases where the predicted activation of alternative pathways is in discordance with their real biological activity. This will additionally require the usage of computational methods for solving such integrated models (Shlomi *et al*, 2007). Building upon the basic approach presented here, these approaches may further advance the search for reliable model-driven predictions of metabolic biomarker alterations on a genomic scale.

Materials and methods

A CBM approach for predicting metabolic biomarkers

Our method is applied to the genome-scale human metabolic network model of (Duarte *et al*, 2007). The model consists of 320 *boundary metabolites* that can be taken-up or secreted from human tissues through pseudo-reactions called *exchange reactions*. A positive flux through an exchange reaction represents the secretion of the boundary metabolite, while a negative flux represents its uptake. This model defines a space of feasible flux distributions that satisfy mass-balance constraints (embedded in the stoichiometric matrix S) and flux directionality constraints (embedded in the flux bound vectors $v_{\min}$, $v_{\max}$), as shown below in equations (1) and (2). Nonlinear thermodynamic constraints that are computationally harder to consider were not accounted for (Beard *et al*, 2002). We define the *exchange interval* for a boundary metabolite by determining the minimal and maximal value of its exchange reactions. The minimal and maximal values of an exchange reaction i are computed using FVA (Mahadevan and Schilling, 2003) by solving the following two linear programming optimization problems:

$$\begin{aligned} &\text{Min } v_i \text{ or Max } v_i \\ &s.t \\ &Sv = 0 \end{aligned} \qquad (1)$$

$$v_{\min} \leqslant v \leqslant v_{\max} \qquad (2)$$

For each metabolic reaction r and every boundary metabolite m in the model, we compute the exchange interval of m when r is forced to be active (representing the healthy case), and when r is forced to be inactive (representing the disease case). To force r to be active in the healthy case, we constrain its flux to be larger than a flux activity threshold, denoted ε, and compute the forward exchange interval $H^{+}_{r,m}$. For reversible reactions, we also constrain the flux to be lower than $-\varepsilon$, and compute the backward exchange interval $H^{-}_{r,m}$. The healthy exchange interval $H_{r,m}$ can be determined by taking the union of the forward and backward exchange intervals. However, as missing thermodynamic constraints in the model (which restrict reactions' directionality) may cause the backward exchange interval to falsely account for infeasible metabolic states, we tested a second method for computing $H_{r,m}$, which considers the backward interval only in cases where the forward interval is predicted to be zero. Testing both methods for computing the healthy exchange interval showed a significant advantage to the second method while comparing our prediction with the clinical data obtained from Ramedis (yielding a precision that is 12% higher than that of the first method), and has hence been used in all further analysis. The activity threshold ε was set to a value of 1, and other thresholds in the range of 0.3–1 did not substantially change our results (less then 2% change in the predictions). To force reaction r to be inactive we simply constrain its flux to zero and compute the exchange interval $D_{r,m}$. The commercial CPLEX solver was used for solving LP problems, on a Pentium-4 machine running Linux in dozens of milliseconds per each individual problem.

Metabolic biomarkers are predicted based on a comparison of exchange intervals between the healthy case and the disease case. For exchange intervals $A=[a_1,a_2]$ and $B=[b_1,b_2]$, we define:

$$A < B \text{ if } (a_2 < b_1),$$

and

$$A \leqslant B \text{ if } (a_1 < b_1 \text{ and } a_2 \leqslant b_2) \text{ or } (a_1 \leqslant b_1 \text{ and } a_2 < b_2).$$

A metabolite m is predicted to be a biomarker of reaction r with an elevated extracellular concentration, if $H_{r,m} \leqslant D_{r,m}$, and with a reduced extracellular concentration if $D_{r,m} \leqslant H_{r,m}$. Biomarkers predicted with high confidence are determined similarly, but with the '$<$'operator. To consider only significant changes between exchange intervals, a difference in flux, denoted $a < b$, is considered only when a is at least 10% lower than b. Selection of different sensitivity thresholds in the range of (5–15%) did not substantially alter our results (causing a change of less then 2% in our total predictions).

The prediction of biomarker alterations for gene knockouts that disrupt the activity of several reactions was performed by considering all biomarkers predicted for the affected reactions (based on the gene-to-reaction mapping in the model of Duarte *et al* and the list of disease causing genes extracted from OMIM given in Online Supplementary Dataset 1). In case of inconsistency between predicted elevated or reduced extracellular concentrations of a metabolite (once different reactions associated with the same gene are inactivated), we determine the biomarker state based on a majority rule, and consider it to be unchanged in case of a tie.

Validation using the RBC kinetic model

The RBC kinetic model consists of 43 metabolites, 43 internal reactions, and 12 primary exchange reactions (Jamshidi *et al*, 2001). The set of differential equations, describing metabolite concentration dynamics, were solved through Matlab's 'ode15s' solver. Enzyme deficiencies were simulated by modifying the maximal rate constant of each enzyme in turn to 10% of its original value.

Automatic extraction of biomarker data from the OMIM database

A list of genetic metabolic diseases along with their causing genes was obtained from the OMIM database (McKusick, 2007). Biomarker data were extracted by parsing the disease description field in the OMIM database in search for metabolite names along with a mentioning of biofluids (e.g. plasma, urine). A dictionary of metabolite synonyms extracted from HMDB was used to resolve naming convention issues (Wishart *et al*, 2007).

Manual curation of biomarker data for amino acid-associated disorders

A set of in-born errors of amino-acid metabolism was obtained from the ICD-10 catalog of metabolic diseases (sections: E70–E72) as classified by the World Health Organization (World Health Organization, 2004). This set was reduced to include disorders that are cataloged in OMIM by their explicit ICD-10 name and are associated with a model gene (60 disorders). An additional 15 disorders that were predicted by our method to affect amino acids were added to this set. The known biomarkers of the amino acid-associated disorders compiled above were manually extracted from the disease description field in the OMIM database. The set of disorders was further filtered to include only the disorders that were reported to show a concentration change in at least one of the model's boundary metabolites. This resulted in a final set of 17 disorders that composed the validation set (Supplementary Dataset 1).

Clinical measurements of biomarkers

The Rare Metabolic Disease database (Ramedis) hosts an extensive set of patients' clinical data including measurements of metabolite concentration level in biofluids for 74 rare metabolic diseases (Töpel *et al*, 2006). Further data were obtained from HMDB, which records for each metabolite a list of normal and abnormal concentration levels in biofluids covering a set of 320 disorders (Wishart *et al*, 2007). Mining these databases for disorders that are associated with model genes and metabolites, as well as show a consistent view of metabolite concentration changes in both databases, resulted in a set of 29 metabolic disorders (Supplementary Dataset 1).

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

We are grateful to Markus Herrgard for very valuable discussions and comments, from the inception of this study to its final form. MC is a fellow of the Edmond J Safra Program in Tel-Aviv University. TS is supported by an Eshkol Fellowship from the Israeli Ministry of Science. This research was supported by grants from the Israeli Science Foundation (ISF) to ER.

References

Akesson M, Forster J, Nielsen J (2004) Integration of gene expression data into genome-scale metabolic models. *Metab Eng* **6:** 285–293

Beard DA, Liang SD, Qian H (2002) Energy balance for analysis of complex metabolic networks. *Biophys J* **83:** 79–86

Cellarier E, Durando X, Vasson MP, Farges MC, Demiden A, Maurizis JC, Madelmont JC, Chollet P (2003) Methionine dependency and cancer treatment. *Cancer Treat Rev* **29:** 489–499

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2:** 2366–2382

Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429:** 92–96

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci USA* **104:** 1777–1782

Durmus Tekir S, Cakir T, Ulgen KO (2006) Analysis of enzymopathies in the human red blood cells by constraint-based stoichiometric modeling approaches. *Comput Biol Chem* **30:** 327–338

Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y (2000) Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci USA* **97:** 1143–1147

Jamshidi N, Edwards JS, Fahland T, Church GM, Palsson BO (2001) Dynamic simulation of the human red blood cell metabolic network. *Bioinformatics* **17:** 286–287

Kaddurah-Daouk R, Kristal BS, Weinshilboum RM (2008) Metabolomics: a global biochemical approach to drug response and disease. *Annu Rev Pharmacol Toxicol* **48:** 653–683

Kenny LC, Dunn WB, Ellis DI, Myers J, Baker PN, Kell DB, GOPEC Consortium (2005) Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. *Metabolomics* **1:** 227–234

Lanpher B, Brunetti-Pierri N, Lee B (2006) Inborn errors of metabolism: the flux from Mendelian to complex diseases. *Nat Rev Genet* **7:** 449–460

Lee PJ, Briddon A (2007) A rationale for cystine supplementation in severe homocystinuria. *J Inherit Metab Dis* **30:** 35–38

Levillain O, Balvay S, Peyrol S (2005) Mitochondrial expression of Arginase II in male and female rat inner medullary collecting ducts. *J Histochem Cytochem* **53:** 533–541

Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, Goryanin I (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* **3:**135

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

McKusick VA (2007) Mendelian inheritance in man and its online version, OMIM. *Am J Hum Genet* **80:** 588–604

Nicholson JK, Connelly J, Lindon JC, Holmes E (2002) Metabonomics: a platform for studying drug toxicity and gene function. *Nat Rev Drug Discov* **1:** 153–161

Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* **2:** 886–897

Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO (2006) Systems approach to refining genome annotation. *Proc Natl Acad Sci USA* **103:** 17480–17484

Seegmiller JE (1968) Detection of human inborn errors of metabolism by examination of urinary metabolites. *Clin Chem* **14:** 412–417

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

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

Shlomi T, Cabili MN, Herrgård MJ, Palsson BØ, Ruppin E (2008) Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol* **26:** 1003–1010

Shlomi T, Eisenberg Y, Sharan R, Ruppin E (2007) A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol Syst Biol* **3:** 101–107

Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles E (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420:** 190–193

Töpel T, Hofestädt R, Scheible D, Trefz F (2006) RAMEDIS: the rare metabolic diseases database. *Appl Bioinformatics* **5:** 115–118

Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L, Fernie AR (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep* **4:** 989–993

Vangala S, Tonelli A (2007) Biomarkers, metabonomics, and drug development: can inborn errors of metabolism help in understanding drug toxicity? *AAPS J* **9:** E284–E297

Wagner M, Naik DN, Pothen A, Kasukurti S, Devineni RR, Adam B-L, Semmes OJ, Wright GL (2004) Computational protein biomarker prediction: a case study for prostate cancer. *BMC Bioinformatics* **5:** 26

Wiback SJ, Palsson BO (2002) Extreme pathway analysis of human red blood cell metabolism. *Biophys J* **83:** 808–818

Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G *et al.* (2007) HMDB: the Human Metabolome Database. *Nucleic Acids Res* **35** (Database issue)**:** D521–D526

WHO (2004) *International statistical classification of diseases and related health problems*, Vol. 1, 2nd edn, tenth revision. Geneva: World Health Organization

Yang J, Xu G, Hong Q, Liebich HM, Lutz K, Schmulling RM, Wahl HG (2004) Discrimination of type 2 diabetic patients from healthy controls by using metabonomics method based on their serum fatty acid profiles. *J Chromatogr* **813:** 53–58