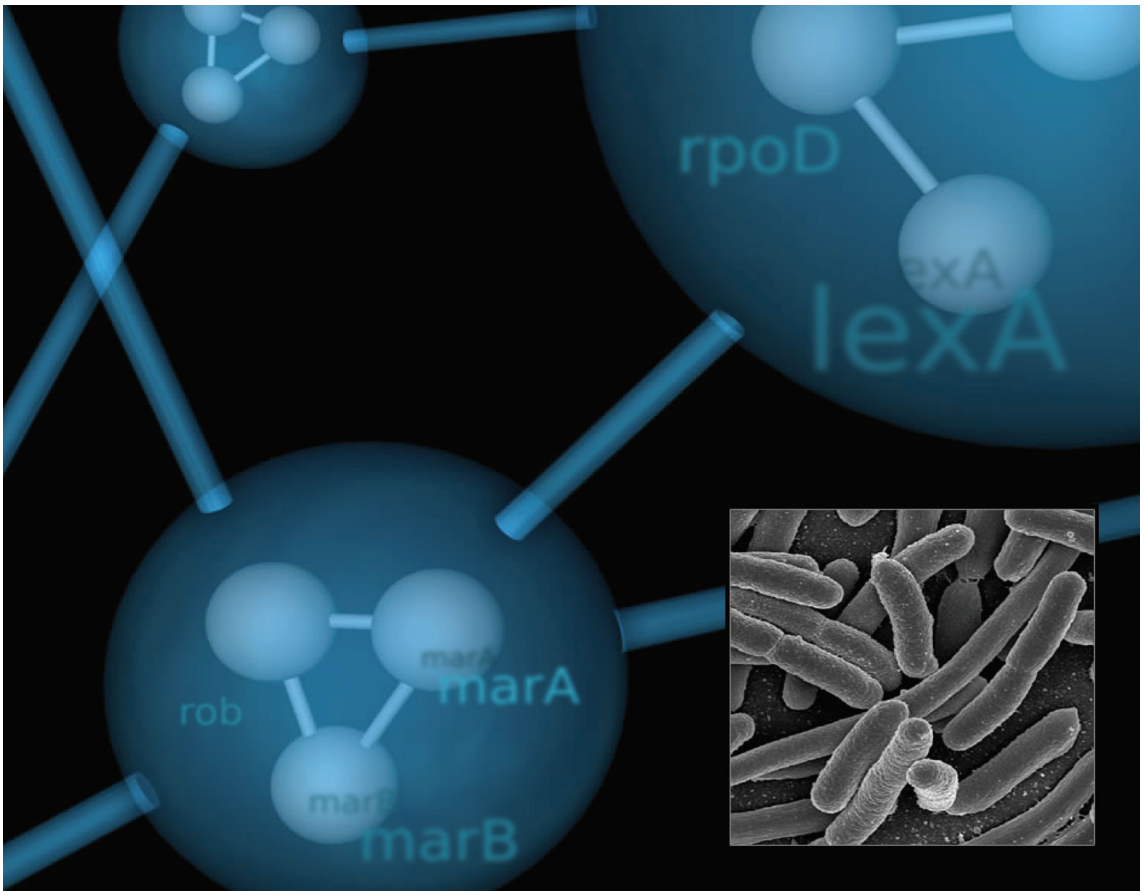
Molecular BioSystems

This article was published as part of the

Computational and Systems Biology themed issue



A network-based method for predicting gene–nutrient interactions and its application to yeast amino-acid metabolism†‡

Idit Diamant,[a] Yonina C. Eldar,[b] Oleg Rokhlenko,[c] Eytan Ruppin[ad] and Tomer Shlomi*[c]

Received 24th December 2008, Accepted 7th May 2009
First published as an Advance Article on the web 18th June 2009
DOI: 10.1039/b823287n

Cellular metabolism is highly dependent on environmental factors, such as nutrients, toxins and drugs, genetic factors, and interactions between the two. Previous experimental and computational studies of how environmental factors affect cellular metabolism were limited to the analysis of only a small set of growth media. In this study, we present a new computational method for predicting metabolic gene–nutrient interactions (GNI) that uncovers the dependence of gene essentiality on the presence or absence of nutrients in the growth medium. The method is based on constraint-based modeling, permitting the systematic exploration of a large putative growth media 'space'. Applying this method to predict GNIs in the amino-acid metabolism system of yeast reveals complex interdependencies between amino-acid biosynthesis pathways. The predicted GNIs also enable the 'reverse-prediction' of growth media composition, based on gene essentiality data. These results suggest that our approach may be applied to learn about the host environment in which a microorganism is embedded given data pertaining to gene lethality, providing a means for the identification of a species' natural habitat.

Introduction

The study of cellular systems is focused not only on the separate effects of genetic and environmental factors, but also on the interactions between the two. The investigation of gene–environment interactions has its roots in the identification of the joint effect of genetic and environmental factors on disease susceptibility[1] and many interactions of that kind are already known today.[2] Previous research in this field involved the development of high-throughput assays for measuring gene–chemical interactions in micro-organisms, which were shown to uncover molecular mechanisms that underlie drug function.[3,4] Large-scale experiments of condition-dependent gene essentiality have provided further insight into the relationship between environmental and genetic factors that affect an organism's growth.[5,6]

The central role of metabolism in cellular function and its close interaction with numerous environmental factors such as nutrients, toxins and drugs, makes studying metabolic gene–environment interactions a particularly important field of investigation. However, this is a complex task due to the highly interconnected nature of metabolic pathways. Computational models of cellular metabolism were previously used to predict chemical–chemical interactions[7] and gene–gene interactions using double[8] or higher-order knockouts.[9,10] Other computational studies have analyzed the environmental specificity of numerous factors, including gene essentiality,[11] genetic interactions,[12] and the contribution of genes to biosynthetic processes.[13] These studies relied on the constraint-based modeling approach, which enables the large-scale modeling of metabolic network function under a diverse range of genetic and environmental conditions (see ref. 14 for a review). Relying on a limited number of growth environments (ranging from 4 up to 53 media), these studies identified specific genes whose essentiality varies across different growth media. However, obtaining a comprehensive map of interactions between genes and nutrients in the growth media requires the inspection of a markedly higher number of growth media. This constitutes a major computational challenge, as the number of potential growth media to which an organism may be exposed may be very large. For example, according to the metabolic network model of ref. 15, the number of compounds that may be taken up by the yeast *Saccharomyces cerevisiae* is 116, which gives rise to an enormous set of 2^{116} possible growth media conditions consisting of subset combinations of these compounds.

In this study, we present a new constraint-based computational framework for predicting metabolic gene–nutrient interactions (GNI) that affect an organism's growth rate. Specifically, we denote the existence of a gene–nutrient interaction between gene x and nutrient y if the essentiality of gene x is modified by the presence or absence of nutrient y in the growth medium (where 'essentiality' does not necessarily entail lethality, but refers to a significant drop in growth rate). We classify GNIs according to two main criteria. The first relates to the presence/absence of nutrients in the media: a *positive* GNI is defined if the nutrient is commonly present in growth media under which the gene is essential. Conversely, a *negative* interaction is defined in cases where the nutrient is commonly

[a] *School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel*
[b] *Department of Electrical Engineering, Technion, Haifa 32000, Israel*
[c] *Department of Computer Science, Technion, Haifa 32000, Israel. E-mail: tomersh@cs.technion.ac.il*
[d] *School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel*
† This article is part of a *Molecular BioSystems* themed issue on Computational and Systems Biology.
‡ Electronic supplementary information (ESI) available: Supplementary figures. See DOI: 10.1039/b823287n

absent in media under which the gene is essential. The second, relates to the extent of the GNI: a *strong* GNI reflects a tight dependency between gene essentiality and nutrient availability. In this case, a gene is essential for growth only when a certain nutrient is always present in the growth media (strong positive interaction) or always absent from it (strong negative interaction). GNIs where the dependency is not absolute as above (*i.e.*, which are present/absent most of the time but not always) are referred to as *weak*. Positive GNIs are expected between genes that are involved in a catabolic pathway and the corresponding uptake nutrients they catabolize. For example, genes involved in galactose catabolism can be expected to have a positive interaction with galactose, as their knockout will affect the organism viability only when galactose is present in the growth media and they are normally active. Inversely, genes involved in anabolic pathways, are expected to have negative interactions with the metabolites they synthesize, as their knockout will affect growth only when the synthesized metabolite is absent from the growth media and needs to be synthesized. For example, genes involved in histidine biosynthesis are expected to have a negative interaction with histidine, as their knockout will affect growth only when histidine is absent from the media.

Considering these definitions, we conduct, for each gene, a comprehensive search through the space of possible growth media, aiming to identify media under which the gene is essential and hence account for its GNIs. We applied this computational approach to predict GNIs that affect yeast growth, focusing on amino-acid containing growth media. Amino-acid biosynthesis involves a significant fraction of the metabolic genes in the model, and we expect negative GNIs with the amino-acids that they are annotated to synthesize. The prediction of these negative GNIs serves as an initial validation of our computational approach. Focusing on GNIs for a defined set of related metabolites (*i.e.* amino-acids) has led to the identification of novel, interesting GNIs that reflect an interdependency between different biosynthetic pathways. We then show how GNIs can be used to 'reverse-predict' the growth media composition of the yeast from gene essentiality data.

Results and discussion

A constraint-based approach for predicting gene–nutrient interactions

To predict strong GNIs, we developed a method that, in contrast to common constraint-based methods, which predict gene essentiality given a growth medium of interest, actually goes the other way around. Our method searches through the space of possible growth media for a medium under which a given gene of interest is essential, thus enabling the identification of strong GNIs as shown below. A metabolic growth medium consists of a set of metabolic nutrients that can be taken up by the organism, while the space of all possible media is the collection of all such sets. A gene is considered *essential* if its knockout causes a significant drop in growth rate to a value that is below a pre-defined threshold (Methods).

The method for predicting strong GNIs is based on an optimization problem that, given a gene of interest, finds

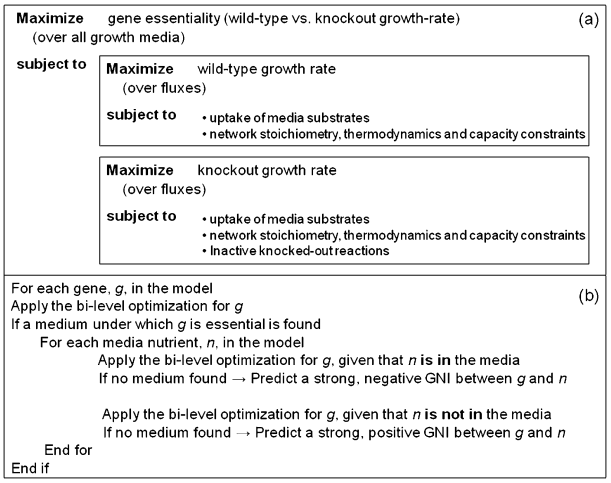


Fig. 1 An overview of the method used to predict strong gene–nutrient interactions. (a) The bi-level optimization searches for a growth medium under which a given gene is essential for growth. It is formulated as a mixed integer linear programming (MILP) problem consisting of an outer and two inner problems. The outer problem searches through the space of possible growth media for a medium under which the drop in the organism's growth rate following a gene knockout is maximized. The inner problems identify optimal flux distributions satisfying various flux constraints for the wild-type and knocked-out strains. (b) A pseudo-code of the method used to predict strong GNIs by applying constrained variants of the bi-level optimization where specific nutrients are forced to either be present or absent from the growth media.

whether there is at least one medium under which this gene is essential. Specifically, we employ a bi-level optimization that searches for a medium in which: (i) there exists a feasible flux distribution for the wild-type strain, satisfying stoichiometric, thermodynamic (*i.e.* reaction directionality constraints as embedded in the network model) and flux capacity constraints, and providing substantial growth rate; and (ii) the drop in the organism's growth rate following the knockout of the gene is maximal (Fig. 1; Methods). The bi-level optimization consists of an *outer problem* that searches for a medium under which the drop in the organism's growth rate following a gene knockout is maximized, and *inner problems* that identify feasible flux distributions for the wild-type and knocked-out strains. The organism's growth rate is computed based on the maximal possible synthesis rate of its essential biomass precursors under a given growth medium, as conventionally done in the flux balance analysis (FBA) method.[16,17] The optimization problem is solved optimally *via* a mixed integer linear programming (MILP) formulation (Methods). Notably, previous constraint-based metabolic modeling studies have already used MILP and bi-level optimization for exploring metabolic flux distributions,[18] metabolic-regulatory steady-states,[19] gene knockout sets that lead to metabolite over-production,[20] possible biomass coefficients,[21] and system objectives.[22] In difference to these studies, our method is the first to systematically explore the space of possible growth media related to gene essentiality. As an alternative to employing MILP (which may require long running times), the optimization problem can be efficiently solved *via* a heuristic coordinate

descent method[23] (involving a series of linear optimization problems), which is shown to extend its scalability while maintaining accurate results (Methods).

The prediction of strong GNIs for a given gene begins with the application of the above bi-level problem to find whether it is essential for growth in at least a single growth medium. In the case that the gene is found to be essential in at least a single medium, the prediction of strong GNIs for this gene involves the application of a constrained variant of the bi-level optimization method (Fig. 1b; Methods). Specifically, to identify a strong positive GNI, the optimization method examines whether *all* growth media in which the gene is essential for growth contains a certain nutrient. This is done by restricting the search to growth media that lack the nutrient and looking for at least a single medium under which the gene is essential (*i.e.*, by re-employing the bi-level optimization). A failure to identify such a growth medium reflects a positive GNI. Inversely, to identify a strong negative GNI, the optimization method examines whether all growth media in which the gene is essential for growth lacks a certain nutrient. This technique is akin to flux variability analysis (FVA),[24] which is commonly used to explore the space of alternative flux distributions under a given growth medium, but is employed here to explore the whole space of possible growth media that give rise to gene essentiality.

To predict weak GNIs, we perform a sampling of the space of possible growth media, and apply FBA to predict gene knockout effects under each sampled medium (Methods). A positive (or negative) weak GNI is identified when a nutrient is present (or absent) in a significantly high number of media under which the gene is essential (Methods). In cases where a nutrient is found to be present (or absent) in all sampled growth media, the sampling method may also hint to the existence of strong GNIs. However, the sampling method is applicable only for genes that are essential for growth under a high number of growth media, where the sampling can obtain sufficient data to provide reliable statistics (Methods; ESI‡ Fig. S2). Notably, the optimization-based method for inferring strong GNIs, by comprehensively searching the whole of media space, does not suffer from this limitation.

An illustrative example of the method's predictions is shown in Fig. 2a–b. The growth media consist of a subset of four metabolites (M1–M4). The organism's growth depends on its ability to produce its biomass precursors M5 and M8. The method predicts strong positive interactions between gene G1 and M1 and M2, as both M1 and M2 must be present in growth media for G1 to be essential. It also predicts strong negative interactions between G1 and M3 and M4, which must be absent from the medium for G1 to be essential. In order for G2 to be essential, either M3 or M4 must be present in the growth medium but not necessarily both; hence both metabolites form a weak positive interaction with G2. G2 is predicted to have a strong negative interaction with M2, as the presence of M2 (together with either M3 or M4, which lead to the synthesis of M1) renders the gene G2 dispensable. Gene G3 is predicted to be non-essential under all possible growth media (as its activity is backed-up by an alternative pathway) and hence has no GNIs.

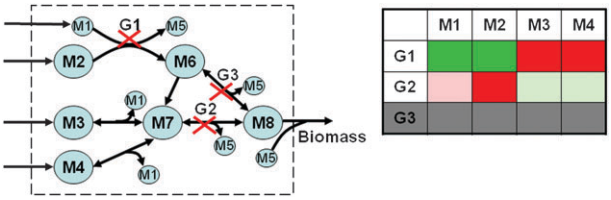


Fig. 2 An illustrative example of GNI predictions in an example network. (a) Nodes represent metabolites and edges represent biochemical reactions. The smaller nodes represent abundant (currency) metabolites that are duplicated several times. (b) Predicted GNIs, with rows representing genes and columns representing nutrients. Dark green and red table entries represent strong positive and strong negative GNIs, respectively. Light green and red entries represent weak positive and weak negative GNIs. Gene *G3* is marked with gray, denoting the fact that there is no growth medium under which it is predicted to be essential.

Predicting interactions between yeast genes and amino-acid nutrients

We applied our method to predict GNIs that affect yeast growth, focusing on amino-acid containing growth media. The analysis was performed using the genome-scale metabolic network model of the yeast *S. cerevisiae* by ref. 15, revealing 88 strong GNIs and 188 weak GNIs, involving a total number of 169 genes (Fig. 3). As expected, a high fraction of the interacting genes (83%) are annotated as involved in amino-acid biosynthesis. The predicted interactions of these genes show an expected pattern in which many of the genes have a negative interaction with the amino-acids that they are annotated to synthesize (as evident by the red diagonal mark in Fig. 3). For example, the genes *HIS1–HIS7*, which form the histidine biosynthetic pathway, have strong negative interactions with histidine, as they are essential for growth only in media that lack histidine. To validate the predicted GNIs, we extracted data on substrate auxotrophy experiments (measuring the dependency of knocked-out yeast strains on the availability of nutrients in the growth medium) from the *Saccharomyces* genome database (SGD). The substrate auxotrophy data is displayed in Fig. 3, super-imposed on-top of the GNI predictions. We found that 65% of the genes that were predicted to have a single strong interaction with a certain amino-acid are known to be auxotrophic to that amino-acid (*i.e.* they require the presence of the amino-acid in the growth media to enable growth; Fig. 3). This high overlap between the predicted GNIs and amino-acid auxotrophy is highly significant (a hyper-geometric p-value $< 10^{-300}$, which reflects the probability of achieving a similar overlap with the substrate auxotrophy data for randomly generated GNIs).

In addition to the expected negative interactions between genes and the amino-acids they are annotated to synthesize, we identify 201 GNIs that reflect the complex interdependency between amino-acid biosynthetic pathways (off-diagonal interactions in Fig. 3). An illustrative example involves *SER1* and *SER2* (3-phosphoserine aminotransferase and phosphoserine phosphatase, respectively). These genes have known, annotated interactions with serine and glycine, but the analysis reveals that they also have strong negative interactions with alanine and proline (Fig. 3). An inspection of

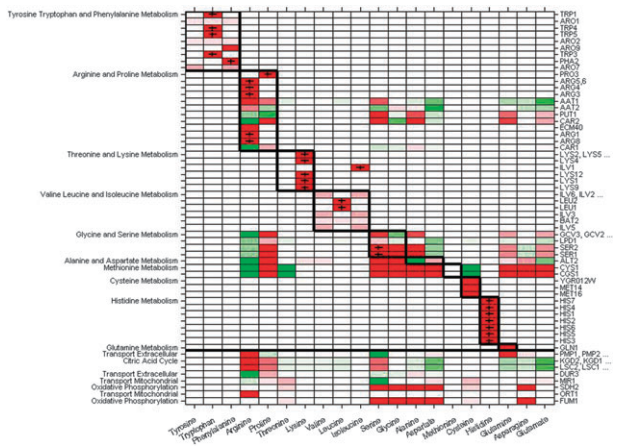


Fig. 3 Predicted gene–nutrient interactions between yeast genes and amino-acid nutrients. Rows represent genes (grouped by their known annotation, left axis) and columns represent amino-acids. Dark green and red table entries represent strong positive and strong negative GNIs, respectively. Light green and red table entries represent weak positive and weak negative interactions, respectively. For example, the top row describing gene *TRP1* shows that the gene is essential only when tryptophan is absent from the media (a strong negative GNI). The following row describing *ARO1* shows that the gene is essential when tyrosine, tryptophan, and phenylalanine are mostly absent from the media (weak GNIs). Table entries marked with a plus sign represent validated predictions based on substrate auxotrophy experiments (obtained from the SGD database). Genes that appear above the black horizontal line at the bottom of the figure are annotated (in the model of Duarte *et al.*[15]) as involved in amino-acid metabolism. Those below this line have predicted GNI interactions even though they do not have such annotations in the original model. The black block rectangles across the matrix's diagonal represent GNIs between genes and the amino-acid they are known to synthesize according to their annotation in the model.

the underlying network topology provides further insight into these predictions. The existence of strong negative interactions with both glycine and serine (which requires both of them to be missing from the growth media in order for *SER1* and *SER2* to be essential) is due to the possible conversion of glycine to serine, or *vice versa*, *via* glycine hydroxymethyl-transferase, which provides an inter-conversion backup between these two metabolites. The negative interactions between the *SER1* and *SER2* genes and alanine are due to an analogous possible backup inter-conversion between alanine and glycine *via* glyoxylate transaminase. This relationship requires alanine to be absent from the growth medium in order for *SER1* and *SER2* to be essential. However, the negative interaction between *SER1* and *SER2* and proline is not readily explained *via* a potential inter-conversion backup mechanism that can be identified by inspecting the network topology. Rather, it is the outcome of more complex stoichiometric relations. Another interesting example is the case of *CYS1* and *CGS1* genes (*O*-acetylserine (thiol) lyase and cystathionine-γ synthase, respectively), which are involved in homo-cysteine synthesis. In this case we identify several strong positive and negative interactions (Fig. 3). The negative interaction with aspartate is due to a pathway that synthesizes homo-cysteine through homo-serine that is produced from aspartate. The positive interaction with threonine is due to the essentiality of threonine in producing cystathionine (a substrate for homo-cysteine synthesis *via* *CYS1* and *CGS1*). The former must be provided in the growth medium, as it can not be produced from aspartate or glycine, which are absent from it.

An inspection of weak GNIs reveals further high level relations between gene essentiality and nutrient availability in the growth media. For example, *ILV2*, *ILV3*, *ILV5* and *ILV6*, which were originally annotated in the model to valine, leucine and isoleucine metabolism, are predicted to have weak interactions with valine and isoleucine but not with leucine (Fig. 3). Inspecting the set of sampled growth media shows that these genes are essential when either valine or isoleucine are absent from the growth media. This is explained by the fact that both valine and isoleucine require 3-methyl-2-oxobutanoate (synthesized by ILV genes) as a precursor, while leucine does not. This prediction is supported by previous substrate auxotrophy experiments, which showed that both valine and leucine must be present in the medium to enable growth of strains with mutations in each of the above genes (SGD).

Supporting the predicted gene–nutrient interactions *via* gene–gene interactions

Similarity in patterns of GNIs between genes is one indication of functional similarity.[4] To provide a second form of large-scale validation of the predicted GNIs, we computed a GNI-based similarity measure between genes and compared it with a common measure of functional similarity, which is based on similarity between patterns of genetic interactions (GI; Methods).[25,26] In the latter measure, two genes are considered similar if they tend to have synergistic (*i.e.* synthetic sick and lethal) genetic interactions with the same set of genes. To identify genetic interactions between genes we followed previous studies, which successfully used flux balance analysis to this aim.[8,10,27] Comparing the GNI- and GI-based functional similarity scores among all pairs of metabolic genes having GNIs resulted in a highly marked Spearman correlation of 0.71 (p-value $< 10^{-300}$; Fig. 4a). The high correlation between the GNI- and GI-based functional similarity measures provides additional testimony to the veracity of the GNI predictions. To demonstrate that the GNI-based functional similarity is not directly reflected in the joint membership of genes in metabolic pathways, we computed an additional functional similarity measure between genes, based solely on pathway annotation data (Methods). The correlation between this pathway-based similarity measure and the GI-based measure was found to be markedly lower (Spearman correlation of 0.57), testifying to our method's ability to correctly capture functional similarity characteristics that are not directly reflected in the pathway annotation data. The high correlation between GNI- and GI-based gene functional similarity further strengthens previous claims regarding the evolution of genetic robustness (as reflected in the GIs) as a congruent effect of the organism's need to grow in different conditions (*i.e.*, environmental robustness; as reflected in the GNIs).[12,28] Notably, relying on experimentally determined genetic interactions could have further strengthened

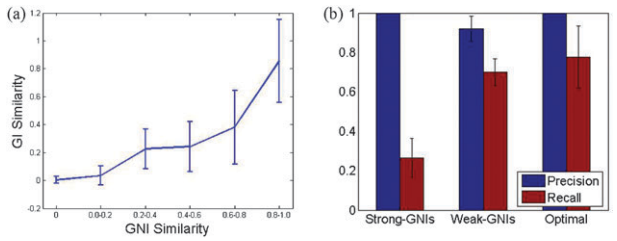


Fig. 4 GNI-*versus* GI-based gene similarity and GNI-based media prediction. (a) The correlation between the similarity in the genes' GNI patterns and their similarity in the genetic interaction (GI) patterns, supporting the biological plausibility of our method's predictions. (b) Mean accuracy of GNI-based predictions of metabolite presence/absence across a set of growth media *via* gene essentiality data. As is evident, the estimated prediction accuracy obtained with weak GNIs is close to the optimal possible accuracy (computed as described in the main text).

this analysis, but no large-scale data on genetic interactions involving yeast metabolic genes is currently publicly available.

Reverse-prediction of growth medium composition based on gene–nutrient interactions

A highly interesting potential application of this new method for predicting GNIs, is to infer the composition of a specific growth medium of interest based on phenotypic data on gene essentiality under this medium. Two previous studies with this aim made initial strides towards finding some characteristics of the natural growth environment of a pathogen within a host organism, based on gene essentiality data obtained *via in vivo* cultivation in the host.[29,30] Specifically, both studies investigated the natural growth environment of *Escherichia coli* within the mouse intestine by performing knockouts of metabolic genes in the bacteria and testing for resulting effects on its fitness, *i.e.*, its ability to grow in the intestine. Using our method and sufficiently ample knockout experimental data, a large-scale prediction of the growth medium composition can be done by summing up the constraints on nutrient availability from all GNIs identified in the environment under study (Methods).

To test this potential application of our method, we applied it to predict the composition of 2^{15} randomly sampled growth media that contain various subsets of amino-acids. In each medium, we obtained gene essentiality data (predicted *via* FBA; Methods) and used it together with predicted GNIs to reverse-predict the presence of metabolites in the medium. Relying solely on strong GNIs provided predictions regarding the presence of 27% of the metabolites on average across the growth media. As strong GNIs reflect definite and exact information about the presence of nutrients in the medium, these predictions yielded a recall of 0.27 with a precision of 1.0. When weak GNIs were added to infer the presence of metabolites in the media, the recall increased to 0.7, with a slightly lower precision of 0.92. For comparison, we computed the maximal possible recall that can be achieved (along with precision of 1.0) by calculating the average number of nutrients (across growth media) whose presence is uniquely correlated with the essentiality pattern of all genes (Methods). We found that the optimal possible recall in the example we studied was 0.77, only slightly higher than that obtained by integrating both strong and weak GNIs (Fig. 4b). Interestingly, the prediction accuracy varies significantly for different nutrients, with some (*e.g.*, tryptophan and phenylalanine) that are uniquely predicted across all sampled growth media, and others (*e.g.*, methionine) uniquely predicted in only 37% of the sampled media. The prediction accuracy may be further improved in future studies by considering higher-order GNIs, reflecting the relationship between the essentiality of a set of genes (*via* double knockouts or high-order *k*-lethality sets[10]) and the presence or absence of nutrients in the growth media.

This paper presents a new computational method for predicting gene–nutrient interactions that affect an organism's viability. The method enables one to computationally study an array of new questions regarding the interplay between gene essentiality and nutrient composition. Specifically, it may be used to study the constraints on the growth environments under which various genes evolved. A classical study demonstrating the effects of environmental constraints on gene evolution was previously done by Alves and Savageau,[31] and showed that an enzyme participating in the biosynthesis of a given amino-acid (*i.e.*, its gene having a negative GNI with that amino-acid, in our framework) has a lower than expected frequency of that amino-acid in its proteomic sequence. Another interesting application of our method may involve the prediction of GNIs between human metabolic disease-causing genes and various dietary nutrients, using the recent reconstruction of the first large-scale human metabolic network model.[32] Finally, as we have shown, the new method can be used to learn about the host environment in which a microorganism is embedded, providing for studies of such species in their natural habitat.

Methods

An optimization method for predicting strong gene–nutrient interactions

To identify a medium under which a given gene is essential for growth, we formulate the following bi-level optimization problem:

$$\max_{w,v,e}(w_{\text{growth}} - v_{\text{growth}})$$

$$s.t.$$

$$S \cdot w = 0 \quad (1)$$

$$w^{\min} \le w_{\text{inner}} \le w^{\max} \quad (2)$$

$$w_{\text{exchange}} \le e \quad (3)$$

$$\max_{v}(v_{\text{growth}}) \quad \text{(I)}$$

$$s.t.$$

$$\begin{cases} S \cdot v = 0, & (4) \\ v^{\min} \le v_{\text{inner}} \le v^{\max} & (5) \\ v_{\text{exchange}} \le e & (6) \\ v_g = 0 & (7) \end{cases}$$

where w denotes a feasible flux vector for the wild-type strain, and v denotes a feasible flux vector following the knockout of gene g. The vector e denotes the upper bounds on the rate of exchange reactions that enable the uptake of nutrients from the environment.

The growth rate of the wild-type and knockout strains are denoted by w_{growth} and v_{growth}, respectively, representing the maximal production rate of essential biomass precursors (as in FBA[16,17]). A feasible flux vector satisfies mass-balance, reaction directionality and flux capacity constraints. S is an $n \times m$ stoichiometric matrix, in which n is the number of metabolites and m is the number of reactions. The mass balance constraint for the wild-type and knockout flux vectors are enforced in eqn (1) and (4), respectively. Thermodynamic constraints that restrict flow direction are imposed by setting $v_{\min}$ and $v_{\max}$ as lower and upper bounds on the wild-type and knockout flux vectors in eqn (2) and (5), respectively. The constraints on the uptake of nutrients from the environment within the wild-type and knockout flux vectors are enforced in eqn (3) and (6), respectively. The knockout of gene g in the knockout flux vector v is enforced in eqn (7). The optimization is formulated as to find a growth medium e along with feasible flux vectors w and v, such that the knockout of g under growth medium e causes the maximal drop in growth rate. The growth rate of the knockout strain is computed based on the maximization of growth reaction in the inner optimization problem. The growth rate of the wild-type strain is computed based on an implicit maximization of the growth reaction in the outer optimization.

To solve the above bi-level optimization (I), we replace the inner linear programming (LP) maximization problem by its dual LP problem,[33] obtaining the following optimization problem:

$$\max_{w,z,e,x_1,x_2,x_3} (w_{\text{growth}} - (x_1^T v^{\max} - x_2^T v^{\min} + x_3^T e))$$

$$\begin{aligned} &s.t. \\ &S \cdot w = 0 \\ &w^{\min} \le w_{\text{inner}} \le w^{\max} \\ &w_{\text{exchange}} \le e \\ &-c + S^T z + x_1 - x_2 + x_3 = 0 \\ &x_1 \ge 0, x_2 \ge 0, x_3 \ge 0 \end{aligned} \qquad \text{(II)}$$

The resulting optimization problem (II) contains a bilinear term $x^T{}_3 e$. It is solved by enforcing e to be binary (representing whether a certain nutrient is available or not in the growth medium), and reformulating the problem *via* mixed integer linear programming (MILP), yielding an optimal solution. Enforcing e to binary values means that nutrients that are absent from the medium would have zero uptake rates and nutrients that are present in the medium will have uptake rates that are lower or equal to one. Allowing for higher uptake rates by multiplying e by a constant factor provided qualitatively similar results. Specifically, the bilinear term is substituted with a new variable h. If e equals one, then h is constrained to x_3 in eqn (9). If e equals zero, then h is constrained to zero in eqn (8). Note that eqn (8) and (9) require an estimation of an upper bound on the dual variable x_3, denoted x_3^{upper}. In our formulation we set $x_3^{\text{upper}} = 100$, while in practice x_3 was found to be smaller than 1 in all runs of the MILP solver. The final formulation of the MILP problem is as follows:

$$\max_{w,z,e,x_1,x_2,x_3,h} (w_{\text{growth}} - (x_1^T v^{\max} - x_2^T v^{\min} + h))$$

$$\begin{aligned} &s.t. \\ &S \cdot w = 0, \\ &w^{\min} \le w_{\text{inner}} \le w^{\max} \\ &w_{\text{exchange}} \le e \\ &0 \le h \le e \cdot x_3^{\text{upper}} \qquad (8) \\ &x_3 - x_3^{\text{upper}} \cdot (1 - e) \le h \le x_3 \qquad (9) \\ &-c + S^T z + x_1 - x_2 + x_3 = 0 \\ &x_1 \ge 0, x_2 \ge 0, x_3 \ge 0 \\ &e \in \{0, 1\} \end{aligned} \qquad \text{(III)}$$

The running time of this optimization (III) is highly dependent on the number of nutrients that may be present in the growth media and on the complexity of the network model. In the application of this method to predict strong GNIs between yeast genes and amino-acids (as described in the Results), the run time was about 12 h on a standard personal computer (using the CPLEX solver), which is 20 times faster than that required by a naïve, brute-force method. Further simulations show that the improvement factor in running time compared to brute-force grows exponentially with the number of nutrients analyzed (ESI‡ Fig. S1). The MILP relaxation of optimization problem (II) may require too much time to solve, when a higher number of media nutrients is considered. To overcome that, we tested an alternative method for solving optimization problem (II) *via* the coordinate descent (CD) method.[23] Comparing the solutions obtained with MILP and the CD-based method showed that the CD-based method correctly identifies 47% of the GNIs between yeast genes and amino-acids. The run time of the CD-based method (which involves a series of linear programming problems) is on the order of minutes regardless of the number of nutrients considered. All results presented in this paper were obtained with the MILP optimization.

To identify strong GNIs for gene x, the optimization method is first used to check whether there exists at least a single medium under which x is essential for growth. A gene is considered essential in a certain growth medium if the drop in the growth rate after its knockout is higher than 20% of the wild-type strain, following previous studies.[10,13] Different thresholds on growth rate drop provided qualitatively similar results (data not shown). Specifically, to identify a strong interaction between gene x and nutrient y, the above optimization method is used to check whether y is present or absent in *all* growth media in which x is essential. This is done by solving a constrained version of the optimization problem where y is forced to be present or absent, by constraining the

corresponding variable e_i to either one or zero, respectively. A failure of the optimization method to identify a growth medium in which nutrient y is forced to be absent reflects a strong positive GNI. Conversely, a failure to identify a growth medium in which the nutrient is forced to be present reflects a strong negative GNI.

A sampling-based method for predicting weak gene–nutrient interactions

To predict weak GNIs for a gene of interest, we perform a random sampling of the space of possible growth media. For each growth medium, FBA is applied to predict the growth rate of the wild-type strain and the growth rate following the knockout of the gene. The set of growth media under which the gene is found to be essential for growth is referred to as *the sampled essential media space*. A weak positive (negative) GNI is defined for nutrients that are significantly over-(under-) represented in the sampled essential media space. The significance is estimated based on a geometric distribution with p-values corrected for multiple-testing *via* the false-discovery rate (FDR) procedure. Notably, the number of sampled growth media should be high to provide a sampled essential media space that is large enough to obtain significant statistics. This may be problematic for genes that are essential for growth in a small fraction of the media space. For example, the dependency between the sampling size required to identify a strong GNI (*via* sampling) and the sizes of the media space and the essential media space is shown in Fig. S2 in the ESI.‡ To predict weak GNIs for the amino-acid containing growth media, we performed a sampling of 2^{15} random media, out of a total space of 2^{20} possible media. To reduce the running time, this method was applied only for genes that are found to be essential in at least a single growth medium (by the above optimization method) for which weak GNIs may exist. The total running time was 13 h on a standard personal computer (instead of the 58 h that would have been required to apply the method for all genes).

Computing functional-similarity measures between genes

The GNI-based functional similarity measure between two genes is calculated based on the Jaccard similarity coefficient between the corresponding two sets of nutrients that interact with each gene. The GI-based similarity measure is analogously based on the Jaccard coefficient between the two sets of genes that interact with each gene in the gene pair examined. GI are predicted using FBA under a rich medium, which consists of all amino-acid nutrients. A GI is predicted in cases where the knockout of a gene pair causes a significant reduction in growth, yielding a growth rate that is at least 20% lower than that obtained under the single knockout of each gene separately.[10,13] The annotation-based similarity measure is computed based on the Jaccard similarity between the two sets of pathway-annotations (as included in the model of ref. 15) associated with each gene pair.

GNIs-based prediction of growth media composition

The prediction of the composition of a growth medium based on strong GNIs and gene essentiality data is straightforward, with strong positive and negative GNIs involving essential genes representing the presence and absence of nutrients in the growth medium, respectively. A similar method is applied for weak GNIs, considering the GNI with the lowest p-value (see sampling procedure above) for each nutrient. The optimal possible prediction accuracy is computed based on nutrients whose presence/absence is uniquely correlated with the essentiality pattern of all genes. Specifically, for each medium, we computed the number of nutrients whose presence/absence is consistent in all media that give rise to the same pattern of gene essentiality.

References

1 G. M. Lower Jr., T. Nilsson, C. E. Nelson, H. Wolf, T. E. Gamsky and G. T. Bryan, *Environ. Health Perspect*., 1979, **29**, 71–79.
2 D. J. Hunter, *Nat. Rev. Genet*., 2005, **6**, 287–298.
3 G. Giaever, P. Flaherty, J. Kumm, M. Proctor, C. Nislow, D. F. Jaramillo, A. M. Chu, M. I. Jordan, A. P. Arkin and R. W. Davis, *Proc. Natl. Acad. Sci. U. S. A*., 2004, **101**, 793–798.
4 A. B. Parsons, R. L. Brost, H. Ding, Z. Li, C. Zhang, B. Sheikh, G. W. Brown, P. M. Kane, T. R. Hughes and C. Boone, *Nat. Biotechnol*., 2004, **22**, 62–69.
5 G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K. D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Guldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kotter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C. Y. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis and M. Johnston, *Nature*, 2002, **418**, 387–391.
6 J. A. Brown, G. Sherlock, C. L. Myers, N. M. Burrows, C. Deng, H. I. Wu, K. E. McCann, O. G. Troyanskaya and J. M. Brown, *Mol. Syst. Biol*., 2006, **2**, 0001.
7 J. Lehar, G. R. Zimmermann, A. S. Krueger, R. A. Molnar, J. T. Ledell, A. M. Heilbut, G. F. Short 3rd, L. C. Giusti, G. P. Nolan, O. A. Magid, M. S. Lee, A. A. Borisy, B. R. Stockwell and C. T. Keith, *Mol. Syst. Biol*., 2007, **3**, 80.
8 D. Segre, A. Deluna, G. M. Church and R. Kishony, *Nat. Genet*., 2005, **37**, 77–83.
9 J. Behre, T. Wilhelm, A. von Kamp, E. Ruppin and S. Schuster, *J. Theor. Biol*., 2008, **252**, 433–441.
10 D. Deutscher, I. Meilijson, M. Kupiec and E. Ruppin, *Nat. Genet*., 2006, **38**, 993–998.
11 B. Papp, C. Pal and L. D. Hurst, *Nature*, 2004, **429**, 661–664.
12 R. Harrison, B. Papp, C. Pal, S. G. Oliver and D. Delneri, *Proc. Natl. Acad. Sci. U. S. A*., 2007, **104**, 2307–2312.
13 T. Shlomi, M. Herrgard, V. Portnoy, E. Naim, B. O. Palsson, R. Sharan and E. Ruppin, *Genome Res*., 2007, **17**, 1626–1633.
14 N. D. Price, J. L. Reed and B. O. Palsson, *Nat. Rev. Microbiol*., 2004, **2**, 886–897.
15 N. C. Duarte, M. J. Herrgard and B. O. Palsson, *Genome Res*., 2004, **14**, 1298–1309.
16 D. A. Fell and J. R. Small, *Biochem. J*., 1986, **238**, 781–786.
17 K. J. Kauffman, P. Prakash and J. S. Edwards, *Curr. Opin. Biotechnol*., 2003, **14**, 491–496.
18 T. Shlomi, O. Berkman and E. Ruppin, *Proc. Natl. Acad. Sci. U. S. A*., 2005, **102**, 7695–7700.
19 T. Shlomi, Y. Eisenberg, R. Sharan and E. Ruppin, *Mol. Syst. Biol*., 2007, **3**, 101.
20 A. P. Burgard, P. Pharkya and C. D. Maranas, *Biotechnol. Bioeng*., 2003, **84**, 647–657.

21 A. P. Burgard and C. D. Maranas, *Biotechnol. Bioeng.*, 2003, **82**, 670–677.
22 E. P. Gianchandani, M. A. Oberhardt, A. P. Burgard, C. D. Maranas and J. A. Papin, *BMC Bioinf.*, 2008, **9**, 43.
23 D. P. Bertsekas, *Nonlinear programming*, Athena Scientific, Cambridge MA, 2nd edn, 1999.
24 R. Mahadevan and C. H. Schilling, *Metab. Eng.*, 2003, **5**, 264–276.
25 A. Beyer, S. Bandyopadhyay and T. Ideker, *Nat. Rev. Genet.*, 2007, **8**, 699–710.
26 S. R. Collins, M. Schuldiner, N. J. Krogan and J. S. Weissman, *GenomeBiology*, 2006, **7**, R63.
27 J. L. t. Hartman, B. Garvik and L. Hartwell, *Science*, 2001, **291**, 1001–1004.
28 J. A. de Visser, J. Hermisson, G. P. Wagner, L. Ancel Meyers, H. Bagheri-Chaichian, J. L. Blanchard, L. Chao, J. M. Cheverud, S. F. Elena, W. Fontana, G. Gibson, T. F. Hansen, D. Krakauer, R. C. Lewontin, C. Ofria, S. H. Rice, G. von Dassow, A. Wagner and M. C. Whitlock, *Evol. Int. J. Org. Evol.*, 2003, **57**, 1959–1972.
29 A. J. Fabich, S. A. Jones, F. Z. Chowdhury, A. Cernosek, A. Anderson, D. Smalley, J. W. McHargue, G. A. Hightower, J. T. Smith, S. M. Autieri, M. P. Leatham, J. J. Lins, R. L. Allen, D. C. Laux, P. S. Cohen and T. Conway, *Infect. Immun.*, 2008, **76**, 1143–1152.
30 D. E. Chang, D. J. Smalley, D. L. Tucker, M. P. Leatham, W. E. Norris, S. J. Stevenson, A. B. Anderson, J. E. Grissom, D. C. Laux, P. S. Cohen and T. Conway, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 7427–7432.
31 R. Alves and M. A. Savageau, *Mol. Microbiol.*, 2005, **56**, 1017–1034.
32 N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas and B. O. Palsson, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 1777–1782.
33 D. Bertsimas and J. N. Tsitsiklis, *Introduction to linear optimization*, Athena Scientific, Belmont, Mass., 1997.