BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 9 2005, pages 2008–2016
doi:10.1093/bioinformatics/bti245

Systems biology

Investigating metabolite essentiality through genome-scale analysis of *Escherichia coli* production capabilities

Marcin Imieliński[1,*], Călin Belta[3], Ádám Halász[4] and Harvey Rubin[2]

[1]Genomics and Computational Biology Graduate Group and [2]Department of Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA, 19104, USA, [3]Department of Mechanical Engineering, Drexel University, Philadelphia, PA, 19104, USA and [4]General Robotics, Automation, Sensing and Perception Laboratory, University of Pennsylvania , Philadelphia, PA, 19104, USA

Received on October 22, 2004; revised on December 8, 2004; accepted on December 19, 2004
Advance Access publication January 25, 2005

ABSTRACT

Motivation: A phenotype mechanism is classically derived through the study of a set of mutants and comparison of their biochemical capabilities. One method of comparing mutant capabilities is to characterize producible and knocked out metabolites. However such an effect is difficult to manually assess, especially for a large biochemical network and a complex media. Current algorithmic approaches towards analyzing metabolic networks either do not address this specific property or are computationally infeasible on the genome-scale.

Results: We have developed a novel genome-scale computational approach that identifies the full set of biochemical species that are knocked out from the metabolome following a gene deletion. Results from this approach are combined with data from *in vivo* mutant screens to examine the essentiality of metabolite production for a phenotype. This approach can also be a useful tool for metabolic network annotation validation and refinement in newly sequenced organisms. Combining an *in silico* genome-scale model of *Escherichia coli* metabolism with *in vivo* survival data, we uncover possible essential roles for several cell membranes, cell walls, and quinone species. We also identify specific biomass components whose production appears to be non-essential for survival, contrary to the assumptions of previous models.

Availability: Programs are available upon request from the authors in the form of Matlab script files.

Contact: imielns@mail.med.upenn.edu

Supplementary information: http://www.cis.upenn.edu/biocomp/manuscripts/bioinformatics_bti245/supp-info.html

1 INTRODUCTION

Recent years have witnessed the sequencing and annotation of over 100 microbial genomes and the development of techniques that allow high throughput mapping of genotypes to phenotypes (Bernal *et al*., 2001; Bochner, 2003). Interpretation of this data is best approached through a formal framework that facilitates building of hypotheses on a systems-level scale (Karp, 2001). A genome-scale metabolic model serves such a purpose, providing a general description of the current state of knowledge regarding the genetics and biochemistry of metabolism in an organism (Covert *et al*., 2001). Analysis of such a model may yield predictions regarding *in vivo* cellular behaviors and insight into the complex relationship between cell components and systems-level cellular phenotypes.

One fundamental question we may ask regarding the capabilities of a metabolic network is whether it contains the reactions necessary to form a product from a given set of media substrates (Schuster *et al*., 2000). Though for small networks this question may be answered by manual inspection, such an approach is impractical for large networks. Alternatively, an algorithmic search for the existence of a path using graph analysis is misleading since most paths in a graph representation of a metabolic network are not in fact biochemical pathways (Arita, 2004). Another approach may be to consult descriptions of biosynthetic pathways in textbooks or databases; however, these manual annotations typically do not encompass the full set of pathways that are valid in a large network. The latter may be comprehensively revealed through algorithmic enumeration approaches such as extreme pathway or elementary flux mode analyses (Schilling *et al*., 1999b; Schuster *et al*., 2000). Unfortunately due to their significant computational complexity these methods are difficult to apply to genome-scale metabolic models (Schilling *et al*., 2000b,c; Papin *et al*., 2003, 2004).

In this paper we propose a simple mathematical criterion for determining metabolite producibility, which we define as the ability of a metabolic network to sustain the production of a chemical species given a media and a genotype. Using the stoichiometry matrix and flux constraints, we apply this criterion to each species in the metabolome to generate the producible set of metabolites for a given media and genotype combination. Comparison of producible metabolite sets between mutant and wild-type *in silico* strains reveals the set of metabolite knockouts resulting from a gene deletion. Compilation of these results for a large set of mutants yields a gene to metabolite knockout map.

Given these *in silico* predictions, we can employ *in vivo* data to investigate the essentiality of individual metabolites for a phenotype. An essential role for a metabolite in a phenotype is suggested if the knockout of the metabolite correlates consistently with the abolishment of the phenotype. Conversely, a metabolite is non-essential for a phenotype if the phenotype persists despite the knockout of that metabolite. Using our method, the existence of a metabolite knockout can be predicted from the *in silico* analysis of a genome-scale metabolic model for a mutant. Such predictions can be used as a means

[*]To whom correspondence should be addressed.

of comparing the metabolic capabilities of *in vivo* strains that differ in a phenotype. These results can then be used to suggest metabolites that are essential and non-essential for a phenotype. These essentiality characterizations can be evaluated for consistency with previous biological understanding and used to formulate novel mechanisms for cell-level phenotypes. This approach can be an especially useful application of genome-scale metabolic modeling towards the interpretation of results from a large-scale genotype to phenotype screen.

We demonstrate an application of this method toward furthering the understanding of *Escherichia coli* survival in the context of a recent genome-scale metabolic model. Analyzing the production capabilities of a large set of mutants, we are able to generate hypotheses regarding the essentiality of metabolites for survival. We also demonstrate how our producibility results can be used to reveal specific inconsistencies between the metabolic network annotation and the biomass model of survival.

2 METHODS AND IMPLEMENTATION

2.1 Mathematical formulation

2.1.1 Metabolic network representation We represent a metabolic network of n metabolites and m reactions in an $n \times m$ stoichiometry matrix S (Clarke, 1988; Heinrich *et al*., 1996). We use x to represent the vector of species concentrations and v to represent the vector of fluxes through all reactions in the system. We represent reversible reactions as two irreversible reactions in S and restrict all fluxes to be non-negative. Reactions are further constrained to be inactive or active by a vector of upper bounds u which is computed using the genotype and media composition.

In our formalism, the columns of S correspond to all known reactions of small molecule biochemistry occurring inside the cell, including transport and core metabolism. It does not include macromolecular processes such as protein synthesis, DNA replication and lipid synthesis or the dilution flux brought about by cell division.

2.1.2 Quasi-steady state assumption In the cell, the rate of change of each metabolite concentration $\dot{x}_i$ is determined by two factors: production from the metabolic network $S_i v$ (where S_i corresponds to the i-th row of S) and consumption by other cellular processes c_i. Formally, $\dot{x} = Sv - c$, where c represents the concentration rate vector corresponding to consumption of intracellular metabolites by cellular processes outside of metabolism.

Metabolic reactions are known to occur at a rapid rate with respect to slower environmental changes, cell division and transcriptional regulation. As a result, when modeling on slower time scales, one can assume intracellular metabolite concentrations to be at steady state. According to this assumption, the flux vector v will obey the following set of linear equalities and inequalities:

$$Sv - c = 0, \quad 0 \le v \le u \tag{1}$$

Since certain metabolites may be produced in net by macromolecular and other cellular processes, certain components of c may be potentially negative. (Note that in our formulation, a negative value for a component c_i represents ‘negative consumption’ or production of a metabolite by other cellular processes.) For example, during a catabolic state, there may be net production of amino acids by proteolytic processes. However, given biologically reasonable assumptions regarding the physiology of a given cellular state, we can identify the portion of our system corresponding to metabolite indices $P \subseteq \{1, \dots, n\}$ that does not contain sources outside of the metabolic network, i.e. for which $c_i \ge 0$.

In particular, during growth, intracellular metabolites such as amino acids and deoxyribonucleotides undergo net consumption by macromolecular processes which form proteins and DNA. Additionally, the diluting effect of cell division acts as a sink for all metabolites present at a non-zero intracellular concentration. Thus, it is biologically reasonable to assume that during the anabolic state of growth most components of c are constrained to values $\ge$0, i.e. they do not possess sources outside of small molecule metabolism. The few exceptions may be protein species that serve carrier roles in metabolic reactions (e.g. acyl carrier protein, thioredoxin) and are products of macromolecular processes. We can further restrict v to obey the following constraints:

$$Sv = c, \quad 0 \le v \le u, \quad c_p \ge 0, \quad p \in P. \tag{2}$$

2.1.3 Producibility and knockout criteria Given these constraints, we would like to test the ability of the metabolic network to catalyze the net production of a metabolite i. We refer to this property as the producibility of metabolite i. Formally, we test the existence of a flux configuration v satisfying:

$$0 \le v \le u, \quad S_i v > 0, \quad S_p v \ge 0, \quad p \in P, \tag{3}$$

where S_i represents the i-th row of S.

Producibility can be interpreted as a statement regarding the ability of the metabolic network to meet a consumption demand by cellular processes outside of metabolism. Since producibility only refers to the feasibility of a particular network behavior, the most informative result is when this condition fails. For example, if a metabolite i is not producible, we can positively assert that there exists no feasible flux configuration that would meet a positive consumption demand for that metabolite [i.e. the set of v satisfying $c_i > 0$ and Equation (2) would be empty]; in particular, the setting of growth institutes such a consumption demand for each intracellular metabolite due to the diluting effect of cellular volume expansion. Non-producibility in this setting asserts that a non-zero intracellular concentration of the corresponding metabolite is unsustainable at steady state, given the assumptions of the model.

One effect of a gene deletion is to disable synthetic pathways that lead to the production of a metabolite. This results in infeasibility of production in the mutant for a metabolite that was producible in the wild-type. In this manner, we can demonstrate the existence of a metabolite knockout following an *in silico* gene deletion. Applying the producibility criteria to each metabolite for the mutant and wild-type strains allows the association of a gene knockout with a set of metabolite knockouts, resulting in a metabolite knockout profile for a mutant and a gene to metabolite knockout map for a set of mutants.

2.2 Implementation

2.2.1 Genome-scale metabolic model Reaction information from the published genome-scale metabolic model of *E.coli* iJR904 was used to construct a 618 × 1176 stoichiometry matrix (Reed *et al*., 2003). Flux constraints were computed using genotype, reaction reversibility and nutrient media composition information. Given a genotype, reactions were labeled active or inactive according to the gene-protein-reaction (GPR) association file, kindly provided by B. Palsson. Inactive reactions were modeled by constraining the corresponding flux to zero. Transport reactions were additionally labeled inactive if the corresponding extracellular metabolite was not present in the media. However, when modeling rich media, we assumed the availability of all known extracellular metabolites. Abbreviations and full names of metabolites mentioned in this study are listed in Table 1.

2.2.2 Producibility algorithm The producibility of each intracellular metabolite in each *in silico* strain was tested by implementing the above flux constraints for each media and genotype combination and checking the existence of a v satisfying Equation (3) using the Matlab optimization toolbox (Mathworks). The set of metabolites corresponding to P in Equation (3) is provided as Supplementary Data. Producible metabolite sets in the wild-type and 895 single-gene mutant strains were compared to generate metabolite knockout profiles. The latter were compiled into a sparse matrix, subsets of which were chosen for visualization.

2.2.3 In vivo *data* Essential gene data for mutants were obtained from the Profiling of *E.coli* Chromosome (PEC) database (http://www.shigen.nig.ac.jp/ecoli/pec/index.jsp), which annotates *E.coli* genes as essential or non-essential according to evidence from the literature. Genes characterized

Table 1. Full names of metabolites whose abbreviations are used in this study

Abbreviation	Full name
12dgr	1,2-Diacylglycerol (*E.coli*)
2dmmq8	2-Demethylmenaquinone 8
2dmmql8	2-Demethylmenaquinol 8
2me4p	2-C-methyl-D-erythritol 4-phosphate
2mecdp	2-C-methyl-D-erythritol 2,4-cyclodiphosphate
2ombzl	2-Octaprenyl-6-methoxy-1,4-benzoquinol
2omhmbl	2-Octaprenyl-3-methyl-5-hydroxy-6-methoxy-1,4-benzoquinol
2ommbl	2-Octaprenyl3-methyl-6-methoxy-1,4-benzoquinol
2omph	2-Octaprenyl-6-methoxyphenol
3hmrsACP	*R*-3-Hydroxy-myristoyl-ACP
5mthf	5-Methyltetrahydrofolate
accoa	Acetyl-CoA
acgam1p	*N*-Acetyl-D-glucosamine 1-phosphate
actACP	Acetoacetyl-ACP
adp	ADP
agpg	Acyl-glycerophosphoglycerol (*E.coli*)
amet	*S*-Adenosyl-L-methionine
atp	ATP
cdpdag1	CDPdiacylglycerol (*E.coli*)
clpn	Cardiolipin (*E.coli*)
coa	Coenzyme A
ctp	CTP
datp	dATP
db4p	3,4-Dihydroxy-2-butanone 4-phosphate
dctp	dCTP
ddcaACP	Dodecanoyl-ACP (*n*-C12:0ACP)
dgtp	dGTP
dmpp	Dimethylallyl diphosphate
dtdp	dTDP
dttp	dTTP
dxyl5p	l-Deoxy-D-xylulose 5-phosphate
fad	FAD
frdp	Farnesyl diphosphate
gam1p	D-Glucosamine l-phosphate
gdp	GDP
glu-D	D-Glutamate
glutrna	L-Glutamyl-tRNA(Glu)
glycogen	Glycogen
grdp	Geranyl diphosphate
gtp	GTP
h2mb4p	1-Hydroxy-2-methyl-2-(*E*)-butenyl 4-diphosphate
hdcea	Hexadecenoate (*n*-C16:1)
hdeACP	Hexadecenoyl-ACP (*n*-C16:1ACP)
hemeO	Heme O
ipdp	Isopentenyl diphosphate
kdo	3-Deoxy-D-manno-2-octulosonate
kdo2lipid4	KDO(2)-lipid IV(A)
kdo2lipid4L	KDO(2)-lipid IV(A)
kdo2lipid4p	KDO(2)-lipid IV(A)
kdo8p	3-Deoxy-D-manno-octulosonate 8-phosphate
kdolipid4	KDO-lipid IV(A)
lipa	KDO(2)-lipid(A)
lipa_cold	Cold adapted KDO(2)-lipid (A)
lipidA	2,3-*Bis*(3-hydroxytetradecanoyl)-D-glucosaminyl-1,6-beta-D-2,3-*bis*(3-hydroxytetradecanoyl)-beta-D-glucosaminyl 1-phosphate
lipidAds	Lipid A disaccharide
lipidX	2,3-*Bis*(3-hydroxytetradecanoyl)-beta-D-glucosaminyl 1-phosphate
lps	lipopolysaccharide (*E.coli*)
malACP	Malonyl-[acyl-carrier protein]
mql8	Menaquinol 8
mqn8	Menaquinone 8
mthgxl	Methylglyoxal
nad	Nicotinamide adenine dinucleotide
nadh	Nicotinamide adenine dinucleotide-reduced
nadp	Nicotinamide adenine dinucleotide phosphate
nadph	Nicotinamide adenine dinucleotide phosphate-reduced
ocdcea	Octadecencate (*n*-C18:1)
octeACP	Octadecenoyl-ACP (*n*-C18:1ACP)
pa	Phosphatidate (*E.coli*)
pap	Adenosine 3′,5′-bisphosphate
pe	Phosphatidylethanolamine (*E.coli*)
peptido	Peptidoglycan subunit of *E.coli*
pg	Phospatidylglycerol (*E.coli*)
pgp	Phosphatidylglycerophosphate (*E.coli*)
prpp	5-Phospho-alpha-D-ribose 1-diphosphate
ps	Phosphatidylserine (*E.coli*)
q8	Ubiquinone-8
q8h2	Ubiquinol-8
sucarg	*N*2-Succinyl-L-arginine
succoa	Succinyl-CoA
tdeACP	Tetradecenoyl-ACP (*n*-C14:1ACP)
thmpp	Thiamine diphosphate
ttdcea	Tetradecenoate (*n*-C14:1)
u23ga	UDP-2,3-*bis*(3-hydroxytetradecanoyl)glucosamine
u3aga	UDP-3-*O*-(3-hydroxytetradecanoyl)-*N*-acetylglucosamine
u3hga	UDP-3-*O*-(3-hydroxytetradecanoyl)-D-glucosamine
uaagmda	Undecaprenyl-diphospho-*N*-acetylmuramoyl-(*N*-acetylglucosamine)-L-ala-D-glu-meso-2,6-diaminopimelcyl-D-ala-D-ala
uagmda	Undecaprenyl-diphospho-*N*-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminopimeloyl-D-alanyl-D-alanine
uama	UDP-*N*-acetylmuramoyl-L-alanine
uamag	UDP-*N*-acetylmuramoyl-L-alanyl-D-glutamate
udcpdp	Undecaprenyl diphosphate
udcpp	Undecaprenyl phosphate
udpg	UDPglucose
ugmd	UDP-*N*-acetylmuramoyl-L-alanyl-D-gamma-glutamyl-meso-2,6-diaminopimelate
ugmda	UDP-*N*-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminopimeloyl-D-alanyl-D-alanine
unaga	Undecaprenyl diphospho *N*-acetyl-glucosamine
utp	UTP

Adapted from Reed *et al*. (2003).

as essential and non-essential by the PEC were represented as lethal and non-lethal, respectively, in our *in silico* rich media growth conditions. In addition to the PEC data, we incorporated essentiality results obtained from the genome-wide transposon mutagenesis study performed by Gerdes *et al*. (2003). In the case of conflicts between the two datasets, the results from the PEC database were used, since these were obtained from individual studies. Essential and non-essential genes obtained from the PEC and Gerdes *et al*. were mapped to the iJR904 *E.coli* genome-scale model via Blattner identification numbers.

2.2.4 Data mining We implemented and applied a variant of the a priori algorithm to discover associations between the *in silico* results and *in vivo* data (Agrawal *et al*., 1993). The algorithm was used to discover combinations of metabolite knockouts that associate with mutant lethality. Further details regarding the algorithm and implementation are provided in the Supplementary Methods.

3 RESULTS

3.1 Wild-type production capabilities

Our analysis shows that the wild-type *in silico* network is capable of sustaining the production of 558 of 618 intracellular metabolites in rich media. The producibility classifications are provided as Supplementary Results. Among the 60 metabolites not producible in the wild-type are betaine aldehyde, 3-dehydro-L-gulonate and arbutin-6 phosphate. These species have previously been annotated as 'dead ends' in the network, since they are employed as only substrate or product in each reaction in which they participate (Reed *et al*., 2003). However many species, including 8-amino-7-oxononanoate and carnitinyl-CoA, are not producible in rich media despite participating in multiple reactions as substrates and products. Though these species are technically not dead ends in the network, their non-producibility implies that they contain one or more chemical moieties that are not supplied by a single transport flux in the model. As a result, there exists no feasible flux configuration in the network that catalyzes their production from the nutrients.

Since we expect these species to actually be present in the *in vivo* wild-type metabolome, we interpret their non-producibility as an aspect of incompleteness in the metabolic network annotation. Therefore, we do not use these findings to impose additional flux constraints on the network but evaluate mutant production capabilities in reference to the wild-type model in its full catalytic capacity.

3.2 Gene deletion results

Data obtained from the PEC database and Gerdes *et al*. provide *in vivo* survival results for deletions of 895 of the 904 genes contained in the IJR904 *in silico* genome-scale metabolic model of *E.coli* (Gerdes *et al*., 2003; Reed *et al*., 2003). Of these mutants, 80 correspond to essential genes and 815 correspond to non-essential genes. We combined the latter *in vivo* data with the corresponding *in silico* metabolite knockout profiles to address the essentiality of individual metabolites for survival. The full set of gene to metabolite knockout results and *in vivo* lethality designation is provided as Supplementary Figure 1 and Supplementary Table 1.

3.3 Production of some biomass components appears non-essential for rich media survival

Flux balance analysis (FBA), a major approach towards analyzing genome-scale metabolic models, uses *in silico* biomass production to simulate *in vivo* growth and survival (Varma *et al*., 1994; Schilling *et al*., 1999a). In this approach, biomass production is modeled as flux through a reaction representing the consumption of intracellular metabolites by macromolecular processes. The feasible and optimal values of this flux are studied in the context of various nutrient conditions and genetic perturbations to generate predictions of *in vivo* behavior. This approach has yielded successful predictions regarding wild-type and mutant *in vivo* growth and survival in *E.coli* and *Saccharomyces cereviseae* in the context of various media (Edwards *et al*., 2000, 2001; Ibarra *et al*., 2002; Segre *et al*., 2002; Famili *et al*., 2003; Covert *et al*., 2004).

A necessary condition for survival in flux balance models of *E.coli* is the ability of the metabolic network to synthesize each of the biomass components. We examined the consistency of this assertion with *in vivo* survival data and our *in silico* knockout results. We found that only 40 (49%) of the 81 mutants that knock out one or more biomass metabolites are lethal *in vivo*. These results suggest that either the production of these biomass metabolites is non-essential for survival or that the metabolic network annotation is incomplete with respect to pathways facilitating their production. A gene to metabolite knockout map for these mutants and the corresponding biomass metabolites are shown in Figure 1.

Examples of viable mutants which knock out one or more biomass metabolites are *fabF, coaA, cls, psd, glgA, glgC, rfaG, lpxL* and *ribF*. Biomass metabolites knocked out by deletion of non-essential genes include glycogen, cardiolipin, succinyl- and acetyl-coA, phosphatidylglycerol, peptidoglycan subunit of *E.coli*, phosphatidylserine, phosphatidylethanolamine and LPS. Many of the latter metabolites are components of the cell membrane and cell wall, whose integrity is clearly essential for cell survival. However, our results suggest that the production of some of these metabolites is not essential for the integrity of the cell wall and cell membrane.

Deletion of the non-essential gene *ndk* (nucleotide diphosphate kinase) had widespread effects on the production capabilities of the *in silico* network. The set of 99 metabolites predicted by our algorithm to be knocked out by *ndk* included Coenzyme A, CTP, UTP and several deoxyribonucleotide triphosphates. Given the clear essentiality of these species as substrates for energy transduction and DNA synthesis, our results suggest that *ndk* deletion should be lethal. The observed inconsistency with *in vivo* data suggests that there may exist other reactions not included in the model that complement the deficiency in the *in silico* mutant.

In addition to the existence of viable mutants that knock out biomass metabolites, we find many lethal mutants that fail to knock out any biomass metabolites. Knockouts of 39 of the 80 of the essential metabolic genes fail to knock out a single biomass metabolite. Eight of these fail to knock out a single reaction (due to the presence of isozymes catalyzing the identical reaction), rendering these mutants identical to the wild-type, while 11 fail to knock out any metabolites despite knocking out one or more reactions. In addition to the above, we find 20 lethal mutants that knock out one or more metabolites despite not knocking out metabolites in the biomass set.

The lack of an observed *in silico* effect on biomass production in these mutants stems from the existence of other pathways in the model that are capable of complementing the lost metabolic function. Such pathways and isozymes may be potentially suppressed *in vivo* via transcriptional regulation and other epigenetic mechanisms. Though such effects are not currently accounted for in our model, they could be implemented by applying additional flux constraints in the manner of Covert *et al*. (2001). The implementation of these

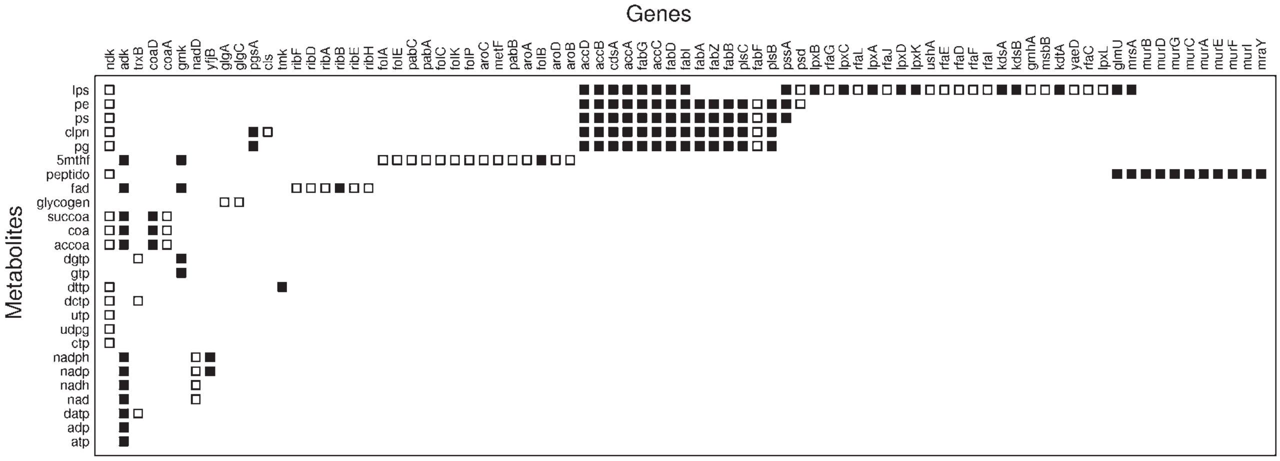

Fig. 1. Gene to metabolite knockout map representing the results of the *in silico* analysis for 81 single gene deletion mutants that knock out one or more biomass metabolites in rich media. The presence of a square in row i and column j represents the knockout of metabolite i by the gene deletion j. Only biomass metabolites knocked out by one or more gene deletions are included. Squares are empty or filled depending on whether the *in vivo* mutant is viable or lethal according to published experimental data. According to experimental results in the literature, 40 of the 81 mutants shown are viable *in vivo*. This is contrary to the assumption that these metabolites are essential for survival. Metabolite notation is taken from Reed *et al*. (2003).

constraints may cause additional metabolite knockouts to emerge for some of the above mutants.

The observed inconsistencies may also arise from incorrect *in vivo* lethality designations for these mutants. This may occur if phenotyping experiments do not observe the mutant through a course of prolonged adaptive evolution, during which an initially slow-growing strain may greatly increase its growth rate and emerge as viable (Fong *et al*., 2004).

A final alternative explanation stems from the possibility that essential metabolites may exist outside of the biomass set. We explore this possibility in the following section through the application of a data mining approach.

3.4 Data mining suggests essential metabolite sets

Using our *in silico* metabolite knockout results, we seek sets of metabolites suggested to be essential by *in vivo* survival data. For a metabolite to be essential, its knockout should consistently correlate with *in vivo* lethality. Figure 2 shows a gene to metabolite knockout map for metabolites whose knockout associates with lethality in >80% of mutants.

This figure shows several species whose knockout associates exclusively with lethality, including 2-*C*-methyl-D-erythritol 4-phosphate (2me4p), KDO (kdo), thiamine diphosphate (thmpp) and ACGAM 1-phosphate (acgam1p). In addition to the latter, there are several species whose knockout associate with lethality in all but one case. For example, the metabolite undecaprenyl-diphospho-*N*-acetylmuramoyl-(*N*-acetylglucosamine)-L-ala-D-glu-meso-2, 6-diaminopimeloyl-D-ala-D-ala (uaagmda) is knocked out in 21 *in silico* mutants, of which 20 are lethal *in vivo*. uaagmda is a component of the cell wall, a structure that is clearly essential for *E.coli* survival. However, its specific essentiality with respect to cell wall integrity has not been fully investigated. The sole exception to this and other associations is *ndk*, whose deletion is non-lethal in rich media. However, as discussed above, the metabolite knockout profile of the *ndk* mutant suggests that some of the *in vivo* biochemical function of this gene may be complemented by reactions not included in the model. If so, the effect of *ndk* deletion would be overestimated in the *in silico* genome-scale metabolic model and *ndk* may not serve as a true exception to these associations.

Though individual metabolites may have the property of being essential, more complex associations may exist between metabolite production and survival. For example, if two metabolites share an essential moiety it may be necessary to produce only one of the pair to allow survival. As a result, lethality will result only if both of these species are knocked out. To reveal such complex associations between metabolite knockout and lethality we applied a standard machine learning approach called association rules of data mining (Agrawal *et al*., 1993). The analysis generated several significant rules linking the *in silico* knock out of a Boolean combination of metabolites to *in vivo* lethality. The metabolite knockout combinations and the corresponding gene deletions are depicted in Supplementary Figure 2.

One such rule associates the knockout of tetradecenoate (ttdcea) or hexadecenoate (hdcea) and 1,2-diacylglycerol (12dgr) or phosphatidate (pa) with *in vivo* lethality. This rule is supported by 12 mutants that correspond to genes in the *fab*, *acc* and *pls* gene clusters, all of which are essential for survival. However, individual knockout of ttdcea, hdcea, 12dgr or pa does not associate exclusively with lethality (Fig. 3). Unlike a simple biomass requirement, such a rule represents the potential flexibility of metabolite requirements for survival or any other phenotype, which may not hinge on the production of a single metabolite but the production of metabolite combinations.

Another complex association links the *in silico* knockout of Heme O (hemeO) and any one of several quinone derivatives (i.e. menaquinol) to *in vivo* lethality in nine out of ten cases. As above, the sole exception to this association was *ndk*. Interestingly, gene deletions which knock out quinone species but not hemeO were viable in 11 of the 12 cases. Similarly, knockout of hemeO without the knockout of quinone species was viable in 9 of the 12 cases. However the knockout of both of these species strongly associates with lethality (Fig. 3). The quinones and heme compounds, though involved in the respiratory apparatus and cellular red-ox balance, are

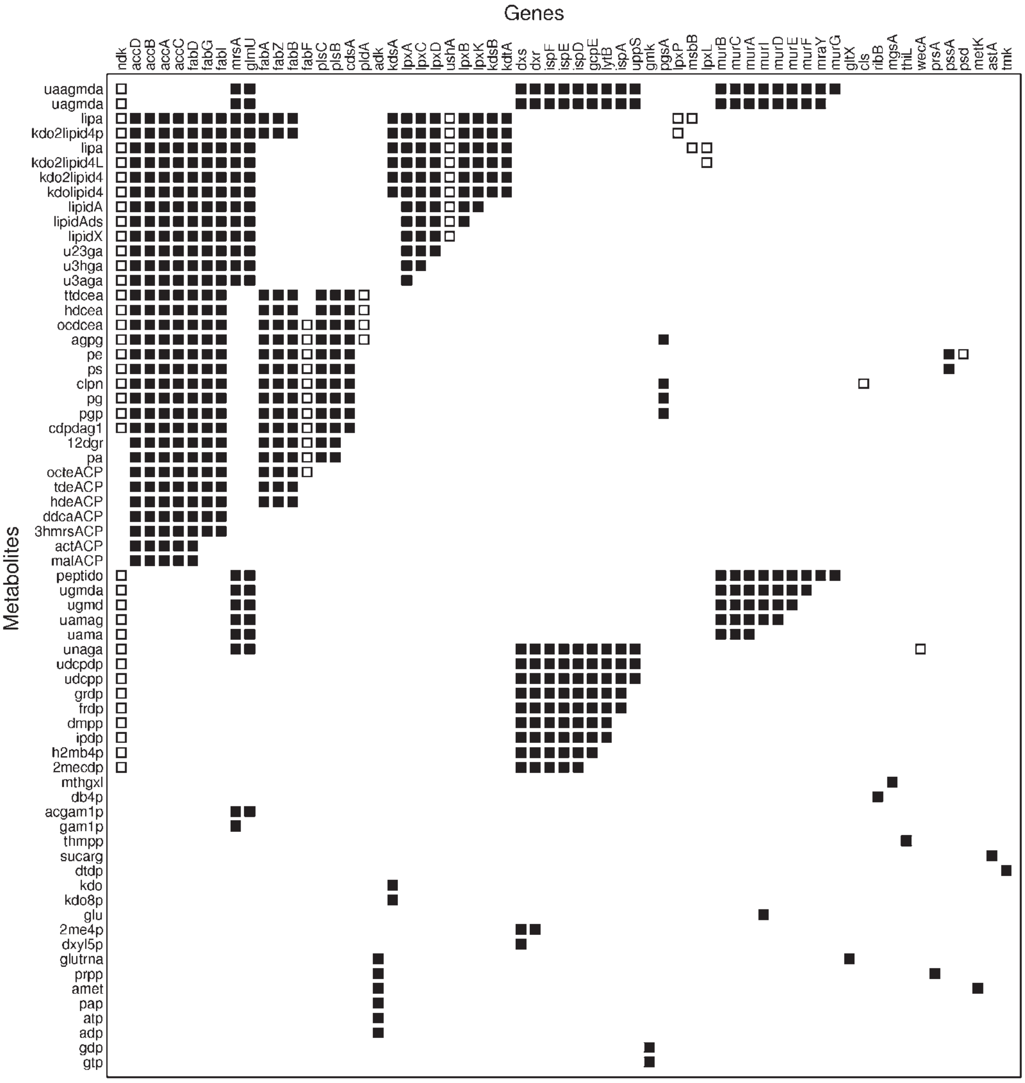

Fig. 2. Gene to metabolite knockout map representing the results of the analysis for 67 metabolites whose *in silico* knockout associates with *in vivo* lethality in >80% of mutants. The presence of a square in row i and column j represents the knockout of metabolite i by the gene deletion j. Each row contains all single-gene deletion mutants that result in the knock out of the corresponding metabolite and for which *in vivo* survival data exists. Squares are empty or filled depending on whether the *in vivo* mutant is viable or lethal according to published experimental data. These results suggest potential novel essential roles for metabolites such as uaagmda and 2-*C*-methyl-D-erythritol 4-phosphate, which are knocked out almost exclusively by lethal genes. Production of these species is not considered essential by previous FBA-based models of survival. Metabolite notation is taken from Reed *et al*. (2003).

not commonly thought to be essential for cellular survival. However, analysis of their *in silico* effect on the network suggests that their knockout is a potential mechanism for the lethality of a large set of essential genes.

There are 31 lethal gene deletion mutants that do not form the support of any association generated by our analysis. Of these, 19 fail to knock out any metabolites *in silico* and are discussed above. The remaining 12 gene mutants knock out one or more metabolites but nevertheless do not form the support of any association. This arises because metabolites knocked out by these gene deletions are also knocked out by the deletion of some non-essential gene. As a result, it is impossible to associate the lethality of these mutants with the knockout of a potentially essential metabolite set. The lack of such an essential metabolomic effect may stem from the overestimation of the set of feasible metabolic states, suggesting the inclusion of reactions *in silico* that are inactive *in vivo*. Causes for the latter include incorrect annotation in the gene to reaction map and the exclusion of gene regulation. Another reason for the lack of mechanism in this set of mutants may arise from uncharacterized non-biosynthetic roles for the corresponding genes. Such roles may include participation in the breakdown of a lethal metabolite or a regulatory effect outside of metabolism, both of which would not be captured by our approach. Finally, the mechanism of lethality in some of these 31 mutants may be mediated by subtle quantitative aspects of metabolite production, such as failure to achieve a critical rate of production or a specific steady-state concentration range. Since our analysis does not include kinetic parameters, such a mechanism would not be resolved by our approach.

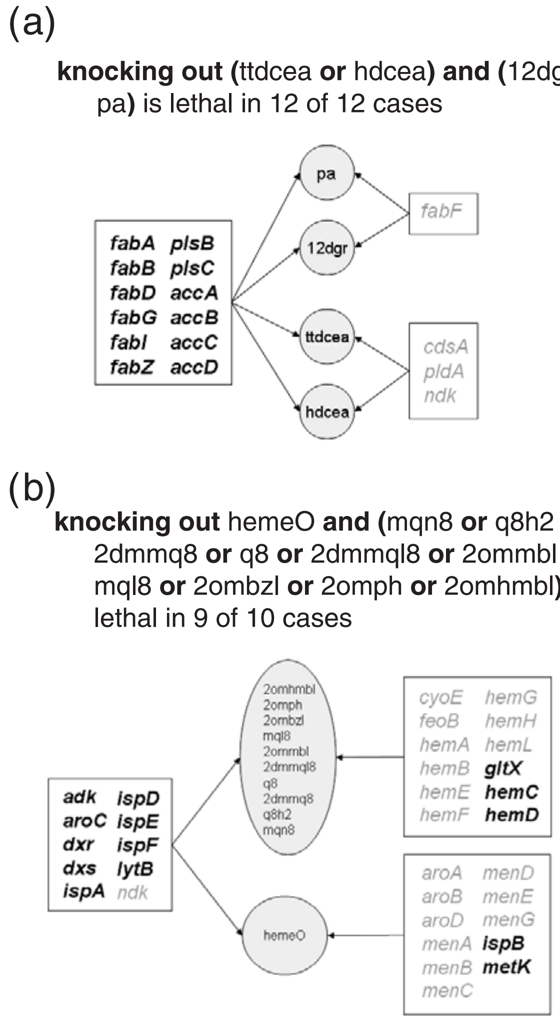

Fig. 3. Diagram of the gene deletion mutants underlying the support of two complex lethality associations. Arrows link genes to metabolites knocked out as a result of their deletion in rich media. The knockout of any single gene in a rectangular box results in the knockout of all metabolites in the rounded boxes pointed to by the arrows. Genes marked in bold are essential while genes marked in light gray are non-essential *in vivo.* (**a**) Evidence behind the association linking the knockout of phosphatiditic acid (pa) or 1,2-diacylglycerol (12dgr) and tetradecenoate (ttdcea) or hexadecenoate (hdcea) to lethality. All the 100% of mutants in the dataset that knock out this metabolite combination are lethal, while mutants that only knock out one of these pairs are viable. (**b**) Illustrates support for the association linking the knockout of quinone species and hemeO to lethality. A total of 90% of mutants knocking out this combination of metabolites are lethal, with ndk as the sole exception. In contrast, there are numerous examples of viable mutants that only knock out either quinone components or hemeO, suggesting that these two sets of metabolites may serve complementary essential functions in rich media survival.

4 DISCUSSION

4.1 Revisiting metabolite roles

An important goal of systems level metabolic analysis is to expand the understanding of the roles of metabolites in cellular phenotypes. Traditionally, small-molecule metabolism is portrayed in terms of its role in generating energy currency and biomass. Though the objectives are undoubtedly important in cellular function, it is likely they do not comprise the totality of the metabolic network's biosynthetic role. In addition to forming biomass and energy substrates, the metabolic network is responsible for the synthesis of all molecules that constitute the steady-state metabolome. Though dilute in comparison to biomass, the set of small-molecule species is remarkably diverse numbering in the hundreds or even thousands. Though some of these species may serve only as intermediates in biochemical processes, others may play essential catalytic and/or homeostatic roles in cellular functions. In addition, some species may have previously uncharacterized roles as substrates for macromolecular cellular processes. The discovery of such roles can be approached through comparison of results from *in vivo* experiments with *in silico* predictions of metabolic network behavior. In particular, our method of enumerating producible and knocked out metabolites in a given media and genotype allows such hypotheses to be generated and tested.

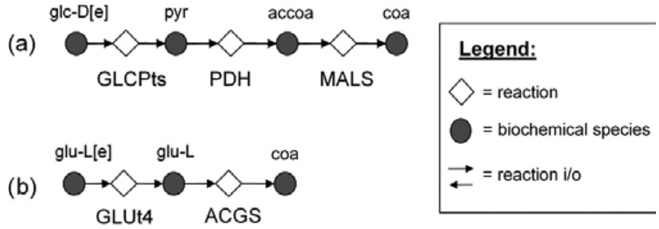

Fig. 4. Path-based graph theoretic notions of connectivity are not well suited to capture the subtle biochemical property of producibility. Shown in (**a**) and (**b**) are two paths in the *coaD* (pantetheine-phosphate adenylyltransferase) mutant metabolic network that link nutrients to the metabolic species CoA (Coenzyme A). According to our algorithm, CoA is knocked out following the deletion of *coaD* in rich media. However, despite the lack of a biosynthetic pathway leading to the production of CoA in this mutant, there exist many paths that join nutrients to this species in a graph representation of the metabolic network. Graphs were visualized using Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

4.2 Relationship to previous approaches

4.2.1 Graph analysis of paths Graph analysis of paths in a metabolic network overestimates real biochemical connectivity by considering all pairs of species on the opposite sides of a reaction that is to be connected (Arita, 2004). However, true biochemical connectedness of two species implies the sharing of moieties and a substrate–product relationship in the context of a biochemical pathway (Arita, 2004). Producibility asserts the existence of precisely such a connection between the nutrient media and a metabolite in the form of a feasible flux configuration. As a result, our approach can reveal a loss of connectivity in mutant networks that is undetectable by graph analytic techniques.

For example, CoA (Coenzyme A) is not producible in rich media in the *in silico coaD* (pantetheine-phosphate adenylyltransferase) deletion strain; however, one can connect nutrient media to CoA in the graph representation of this mutant metabolic network via several paths, examples of which are shown in Figure 4. This effect would not be captured using a simple path based criteria, such as the method used by Lemke *et al*. (2004) to assess the metabolomic damage of enzyme removal. Our results suggest that the notion of producibility may serve as a better indicator of metabolic network robustness than path-based parameters such as graph diameter, giant component size or damage.

4.2.2 Network-based pathway analysis methods Our method of identifying producible metabolites is related to previous approaches for enumerating metabolic capabilities, namely elementary flux mode or extreme pathway analysis (Schilling *et al*., 2000a; Schuster *et al*., 2000). Extreme pathways (EP) and elementary fluxes modes (EFM) correspond to feasible steady-state flux vectors that obey a non-decomposability property. In a system equipped with an output channel for each metabolite, producibility of a metabolite corresponds to the existence of an EP/EFM that has non-zero flux through the respective output channels. Thus, by generating the full set of EFM/EP for such a network, one could theoretically determine the existence or non-existence of such a vector.

Though possible, application of EFM/EP towards determining producibility is not practical for genome-scale application. Since the computational complexity of the EP/EFM search algorithm increases exponentially with respect to the number of reactions in the system, direct application to a genome-scale metabolic model is currently intractable (Schilling *et al*., 2000c). This difficulty has been addressed through manual division of the network into subsystems and computation of extreme pathways for each partition (Schilling *et al*., 2000c, 2002). However the producibility of a metabolite would not immediately follow from these results, since the set of EP/EFM generated is a subset of all possibilities and varies with alternate subsystem decompositions (Schuster *et al*., 2002).

The computational advantage of our approach lies in that it directly seeks to determine the existence of a single feasible flux configuration rather than attempting to enumerate all possibilities. As a result, it offers a practical and automated approach for determining producibility for a large number of potential metabolic outputs.

4.2.3 Optimization approaches Our method differs from previous genome-scale modeling techniques such as FBA and minimization of metabolic adjustment (MOMA) in its discovery-based approach toward the study of the metabolome. The latter methods employ *in vivo* data for the purpose of testing an existing *in silico* model of a phenotype; in contrast, our method employs *in vivo* data to infer novel essential roles for metabolites and build a biochemical model of a phenotype. In addition, our approach is capable of identifying specific inconsistencies between the genome-scale model, *in vivo* data and previous biological knowledge. An example of this is our analysis of the effects of *ndk* deletion, which is non-lethal *in vivo* but according to the metabolic network annotation results in the *in silico* knockout of many metabolites thought to be necessary for survival. This inconsistency can be addressed by either reexamining the biological assertion that these metabolites are essential or revisiting the metabolic network annotation to discover what genes may potentially complement the observed *in silico* deficiency.

4.3 Future directions

Though we have chosen the well outlined phenomenon of *E.coli* survival as the subject of this proof-of-concept study, promising future applications of this method may lie in the study of emerging pathogens. Given data from high-throughput mutant screens, this method could yield predictions regarding the essentiality of metabolites for clinically relevant phenotypes such as survival during infection and drug resistance. This may facilitate a rational approach towards target discovery and pharmaceutical design. In addition, our approach shows promise for metabolic network annotation validation and refinement in newly sequenced organisms whose metabolism has not been as well characterized as that of *E.coli*. Predictions arising from this method can also potentially be tested against *in vivo* metabolomic measurements and the results used to drive discovery of new reactions and re-annotation.

ACKNOWLEDGEMENTS

We thank Seth Berger, Mark Goulian, Howard Goldfine and Vijay Kumar for their valuable inputs. We thank Bernhard Palsson for kindly providing a text form of the GPR associations underlying the *E.coli* genome-scale model. Research was supported by DARPA BioCOMP and NIH.

REFERENCES

Agrawal,R., Imielinski,T. and Swami,A.N. (1993) Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM/SIGMOD International Conference on Management of Data*, Washington DC, pp. 207–216.

Arita,M. (2004) The metabolic world of *Escherichia coli* is not small. *Proc. Natl Acad. Sci. USA*, **101**, 1543–1547.

Bernal,A., Ear,U. and Kyrpides,N. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.*, **29**, 126–127.

Bochner,B.R. (2003) New technologies to assess genotype–phenotype relationships. *Nat. Rev. Genet.*, **4**, 309–314.

Clarke,B.L. (1988) Stoichiometric network analysis. *Cell Biophys.*, **12**, 237–253.

Covert,M.W., Schilling,C.H., Famili,I., Edwards,J.S., Goryanin,II, Selkov,E. and Palsson,B.O. (2001) Metabolic modeling of microbial strains *in silico*. *Trends Biochem. Sci.*, **26**, 179–186.

Covert,M.W., Knight,E.M., Reed,J.L., Herrgard,M.J. and Palsson,B.O. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, **429**, 92–96.

Edwards,J.S. and Palsson,B.O. (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl Acad. Sci. USA*, **97**, 5528–5533.

Edwards,J.S., Ibarra,R.U. and Palsson,B.O. (2001) *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.*, **19**, 125–130.

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

Fong,S.S. and Palsson,B.O. (2004) Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat. Genet.*, **36**, 1056–1058.

Gerdes,S.Y., Scholle,M.D., Campbell,J.W., Balazsi,G., Ravasz,E., Daugherty,M.D., Somera,A.L., Kyrpides,N.C., Anderson,I., Gelfand,M.S. *et al*. (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.*, **185**, 5673–5684.

Heinrich,R. and Schuster,S. (1996) *The Regulation of Cellular Systems.* Chapman & Hall, New York.

Ibarra,R.U., Edwards,J.S. and Palsson,B.O. (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature*, **420**, 186–189.

Karp,P.D. (2001) Pathway databases: a case study in computational symbolic theories. *Science*, **293**, 2040–2044.

Lemke,N., Heredia,F., Barcellos,C.K., Dos Reis,A.N. and Mombach,J.C. (2004) Essentiality and damage in metabolic networks. *Bioinformatics*, **20**, 115–119.

Papin,J.A., Price,N.D., Wiback,S.J., Fell,D.A. and Palsson,B.O. (2003) Metabolic pathways in the post-genome era. *Trends Biochem. Sci.*, **28**, 250–258.

Papin,J.A., Stelling,J., Price,N.D., Klamt,S., Schuster,S. and Palsson,B.O. (2004) Comparison of network-based pathway analysis methods. *Trends Biotechnol.*, **22**, 400–405.

Reed,J.L., Vo,T.D., Schilling,C.H. and Palsson,B.O. (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.*, **4**, R54.

Schilling,C.H., Edwards,J.S. and Palsson,B.O. (1999a) Toward metabolic phenomics: analysis of genomic data using flux balances. *Biotechnol. Prog.*, **15**, 288–295.

Schilling,C.H., Schuster,S., Palsson,B.O. and Heinrich,R. (1999b) Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.*, **15**, 296–303.

Schilling,C.H., Edwards,J.S., Letscher,D. and Palsson,B.O. (2000a) Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol. Bioeng.*, **71**, 286–306.

Schilling,C.H., Letscher,D. and Palsson,B.O. (2000b) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, **203**, 229–248.

Schilling,C.H. and Palsson,B.O. (2000c) Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J. Theor. Biol.*, **203**, 249–283.

Schilling,C.H., Covert,M.W., Famili,I., Church,G.M., Edwards,J.S. and Palsson,B.O. (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.*, **184**, 4582–4593.

Schuster,S., Fell,D.A. and Dandekar,T. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, **18**, 326–332.

Schuster,S., Pfeiffer,T., Moldenhauer,F., Koch,I. and Dandekar,T. (2002) Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae. Bioinformatics*, **18**, 351–361.

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

Varma,A. and Palsson,B.O. (1994) Metabolic flux balancing—basic concepts, scientific and practical use. *Biotechnology*, **12**, 994–998.