Using Topology of the Metabolic Network to Predict Viability of Mutant Strains

Citation: Genome Biology. 2005 Dec 28;6(13):P15

As Published: http://dx.doi.org/10.1186/gb-2005-6-13-p15

Publisher: BioMed Central Ltd

Persistent URL: http://hdl.handle.net/1721.1/59195

Version: Original manuscript: author's manuscript prior to formal peer review

Deposited research article

Using Topology of the Metabolic Network to Predict Viability of Mutant Strains

Zeba Wunderlich and Leonid Mirny*

Addresses: Biophysics Program, Harvard University, 77 Massachusetts Avenue, 16-361, Cambridge, MA 02139, USA. *Harvard-MIT Division of Health Sciences & Technology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 16-343, Cambridge, MA 02139, USA.

Correspondence: Leonid Mirney. E-mail: leonid@mit.edu

Posted: 28 December 2005

Genome ***Biology*** 2005, **6:**P15

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/13/P15

Received: 23 December 2005

This is the first version of this article to be made available publicly.

comment
reviews
reports
deposited research
refereed research
interactions
information

Using Topology of the Metabolic Network to Predict Viability of Mutant Strains

Zeba Wunderlich and Leonid Mirny*

Biophysics Program, Harvard University
77 Massachusetts Avenue, 16-361
Cambridge, MA 02139
(617) 452-4075
wunderl@fas.harvard.edu

* Corresponding author
Harvard-MIT Division of Health Sciences & Technology, Massachusetts Institute of Technology
77 Massachusetts Avenue, 16-343
Cambridge, MA 02139
(617) 452-4862
(617) 253-2514 (fax)
leonid@mit.edu

Abstract

Background: Understanding the relationships between the structure (topology) and function of biological networks is a central question of systems biology. The idea that topology is a major determinant of systems function has become an attractive and highly-disputed hypothesis. While the structural analysis of interaction networks demonstrates a correlation between the topological properties of a node (protein, gene) in the network and its functional essentiality, the analysis of metabolic networks fails to find such correlations. In contrast, approaches utilizing both the topology and biochemical parameters of metabolic networks, e.g. flux balance analysis (FBA), are more successful in predicting phenotypes of knock-out strains.

Results: We reconcile these seemingly conflicting results by showing that the topology of *E. coli*'s metabolic network is, in fact, sufficient to predict the viability of knock-out strains with accuracy comparable to FBA on a large, unbiased dataset of mutants. This surprising result is obtained by introducing a novel topology-based measure of network transport: synthetic accessibility. We also show that other popular topology-based characteristics like node degree, graph diameter, and node usage (betweenness) fail to predict the viability of mutant strains. The success of synthetic accessibility demonstrates its ability to capture the essential properties of the metabolic network, such as the branching of chemical reactions and the directed transport of material from inputs to outputs.

Conclusions: Our results (1) strongly support a link between the topology and function of biological networks; (2) in agreement with recent genetic studies, emphasize the minimal role of flux re-routing in providing robustness of mutant strains.

Background

Many have suggested and debated the idea that topology determines network function. Although structures of several biological networks are available, it remains hard to delineate the contributions of topology from the contributions of kinetic and equilibrium parameters. Due to its well-established structure and the wealth of experimental data on cell metabolism, the *Escherichia coli* metabolic network is a perfect model system to explore the role of network topology. Is topology of a metabolic network sufficient to predict the viability of knock-out mutants?

Metabolic networks have been modeled extensively using steady state flux balance approaches [1-6]. To test the capabilities of metabolic network models, many groups have compared predicted and experimentally-measured effects of gene deletions on cell growth. Among the most effective methods are flux balance analysis (FBA) [3, 4, 6, 7], the related minimization of metabolic adjustment (MOMA) method [8], and elementary mode analysis (EMA) [9]. While these methods have been shown useful in understanding the structure and dynamics of metabolic fluxes, they deliver different experimentally testable predictions. FBA can accurately predict fluxes through individual reactions in the wild type and mutant strains [8], as well as the viability of single-gene knockout strains. EMA, in turn, was shown to predict the viability of mutant strains with comparable accuracy [9].   Since these methods use both network topology and the stoichiometry of metabolic chemical and transport reactions, they cannot separate the role of topology from the role played by other parameters in network function. In addition, due to the complexity of the method and the results, EMA techniques are computationally expensive [10] and provide little insight on why certain mutations are lethal, while others are tolerated.

Here we untangle the topology and stoichiometry of the metabolic network and show that topology alone is sufficient to predict the viability of mutant strains as accurately as FBA on a large, unbiased set of mutants [7]. This result supports the claim that topology plays a central role in determining network function and malfunction [11, 12]. We employ a novel network property, synthetic accessibility, an intuitive and transparent way of understanding the effects of metabolic mutation (Figure 1). We define synthetic accessibility, *S*, as the total number of reactions needed to transform a given set of input metabolites into a set of output metabolites, and predict that increases in *S* due to alterations in the topology of the metabolic network will adversely affect growth. The term "synthetic accessibility" is borrowed from the field of drug design where it is defined as the smallest number of chemical steps needed to synthesize a drug from common laboratory reactants [13]. We also demonstrate that other network characteristics such as node degree or change in the graph diameter are unable to predict the viability of mutant strains better than random predictions, suggesting synthetic accessibility is a more appropriate characteristic for networks with directed transport, such as metabolic networks.

Results

Performance of synthetic accessibility. To study the performance of synthetic accessibility in predicting viability of knock-out strains and compare it to previous studies, we tested it on two datasets, a large, unbiased dataset of insertional mutants

[7] and a smaller dataset collected for FBA analysis [3], which mainly contained knock-outs of enzymes involved in central metabolism. We used these datasets specifically because they were used in previous studies[3, 7-9] to which we compared our results. We also used the union of these datasets and refer to it below as the combined dataset. When applied to the combined dataset, our approach performed as well (62% accuracy, $p = 6 \times 10^{-8}$) as the FBA approach (62%, $p = 3 \times 10^{-8}$). (See Table 1, Figure 2 for details.) On the large dataset of 487 insertional mutants [7], the synthetic accessibility approach performed as well (60% accuracy, $p = 3 \times 10^{-5}$) as the FBA and MOMA approaches (58% and 59% accuracy, $p = 1 \times 10^{-3}$ and 1×10^{-4} respectively), with a somewhat higher statistical significance. On a smaller dataset of 79 mutants [3], FBA correctly predicted 86% of the cases, while our topology-based synthetic accessibility approach had 71% accuracy, providing correct predictions for 53/68=78% of the cases predicted correctly by FBA (Figure 3).

The difference in performance of the synthetic accessibility approach between the two datasets (Table 1) is probably due to the way the datasets were interpreted and the cases included in the two datasets. In the smaller dataset [3], the mutant strains are classified as viable or inviable, while in the insertional dataset [7], the mutants are labelled as negatively selected – the population of the mutant strain is less than one-half the wild-type population after 30 generations of competitive growth, or not negatively selected. Since the synthetic accessibility approach deems a mutant strain inviable or negatively selected based the path lengths from inputs to outputs and the accessibility of outputs, the latter classification scheme may correspond more closely to the synthetic accessibility approach – longer path lengths probably correspond to reduced growth rates rather than inviability.

The number and type of data points included in the datasets are also different. The insertional dataset is much larger (487 versus 79 data points) and includes a fairly random collection of insertions in metabolic genes, while the smaller dataset only contains data about the enzymes used in the central metabolism (glycolysis, pentose phosphate pathway, citric acid cycle, respiration processes) [3]. Because the central metabolism contains a number of alternate pathways, some of which may require fewer steps than the commonly used pathways, it is not surprising that the synthetic accessibility approach performs worse when applied to the smaller datasets.

When considering the combined dataset, synthetic accessibility had greater sensitivity, indicating it was better than FBA or MOMA at predicting strains that are viable, but it had lower specificity, indicating that it was not as good at predicting inviable strains (Figure 5). The success of synthetic accessibility on the combined dataset demonstrates reveals three important results, making transparent the difference between most of viable and non-viable strains.

1. Most non-viable mutants simply lack a pathway to synthesize some of their biomass components ($S=\infty$), i.e. one of essential metabolites cannot be produced from the network inputs (Table 4).
2. Our approach correctly predicted that most strains with longer re-routed pathways are inviable, suggesting that re-routing of metabolic fluxes plays a small role in rescuing mutant strains. This result is consistent with results of FBA analysis of yeast mutants [14].

3. Most viable mutants have either untouched primary synthetic pathways or only short re-routing (e.g. due to isozymes).

Performance of other topology-based measures. We tested the ability of other topology-based graph characteristics, such as node degree, graph diameter, and node usage (see Materials and Methods) to predict the viability of mutant strains. Several studies have suggested that nodes that have higher degree are more important for the network, and removal of such nodes in biological networks is more likely to lead to a lethal phenotype [11, 12]. To test this hypothesis, we computed the degree of each enzyme as the number of metabolites participating in reactions catalyzed by this enzyme. A strain was predicted to be inviable if the degree of the knocked-out enzyme was above a certain cutoff. Figure 2 demonstrates that for an optimized cutoff value, this procedure predicts viability worse than a random prediction.

Several theoretical studies have focused on graph diameter as a measure of network performance, defining a graph diameter as a mean of shortest paths between every pair of nodes [11, 15, 16]. To test graph diameter as a predictor of viability, we predicted a mutant to be inviable if increase in graph diameter exceeded a cutoff. Figure 2 shows that, similar to node degree, graph diameter did not perform any better than random predictions.

Similarly, we tested another topology-based measure, enzyme usage, that is analogous to node betweenness [17, 18]. Enzyme usage performed somewhat better than random predictions but worse than synthetic accessibility, which is not surprising, since it basically used a subset of the data produced by the synthetic accessibility approach.

In summary, popular topology-based measures performed more poorly than synthetic accessibility. Moreover, node degree and diameter are no more accurate than simply predicting that all the mutants are viable, which gives an accuracy of 53.8%, and while node usage performed better than node degree and diameter, it was a worse predictor than the synthetic accessibility. (See DataTable3.xls for details.)

These characteristics ignore essential properties of metabolic network: directionality and branching of reactions, and directed transport of material from cellular substrates (sugars, oxygen, etc.) to products (biomass). Synthetic accessibility, in contrast, takes into account these properties of the metabolic network. As such, synthetic accessibility can be thought of as a generalization of the concept of graph diameter for directed transport networks. While certain topological characteristics such as node degree and diameter can be predictive in information carrying networks (e.g. the internet, protein-protein interaction networks), our results suggest that other characteristics like synthetic accessibility are more appropriate for transport in directed networks, such as metabolic networks.

Robustness of synthetic accessibility. Metabolic networks are almost always incomplete and may contain some errors. To study how predictions made using synthetic accessibility depend on some errors in the network, we performed a robustness analysis. Errors were modeled by random re-assignment of certain percentage of enzymes to different reactions. Figure 4 shows how the accuracy of prediction decreased with increased fraction of introduced mistakes. The method tolerated assignment error rates of 5-10%, but the accuracy dropped to the level of random predictions when approximately 50% of enzyme-reaction assignments were shuffled.

Discussion

In this study, we show that the topology and function of the metabolic network are intimately related. By introducing a novel topology-based measure, synthetic accessibility, we were able to correctly predict viability of about 350 of 520 mutant strains of *E. coli*. Synthetic accessibility, *S*, is essentially a network diameter specifically tailored for transport networks, and we show that an increase in *S* is correlated to an inviable phenotype. A significant increase in *S* upon mutation suggests increased metabolic costs, leading to reduction of the growth rate or death. The apparent success of synthetic accessibility can only be attributed to the contribution of network topology, since no other information has been used in these predictions.

Synthetic accessibility can be rapidly computed for a given network, has no adjustable parameters, and in contrast to FBA, MOMA and EMA, does not require the knowledge of stoichiometry or maximal uptake rates for metabolic and transport reactions. On the insertional dataset, the accuracy of synthetic accessibility approach is comparable to FBA and MOMA. The performance of synthetic accessibility as compared to FBA and EMA on the smaller dataset is worse, but this smaller dataset only has data for mutants affecting the central metabolism and therefore may be biased, while the large dataset of insertional mutants is fairly unbiased and representative.

In contrast to FBA, our model assumes that long re-routed fluxes are less efficient than native ones, predicting mutants with longer fluxes (larger synthetic accessibility) as inviable. Although this assumption fails in certain cases (see AdditionalDocumentation.pdf), the similar success rates of FBA and our approach suggest that this assumption holds true for vast majority of mutant strains. We conclude, in agreement with a recent study [14], that re-routing does not contribute significantly to robustness of knock-out mutants.

Similar accuracy achieved by techniques based on flux balance and synthetic accessibility points at the network topology as a primary determinant of the viability predictions of FBA and MOMA. Although our results suggest that network topology is sufficient to predict strain viability and use of stoichiometric coefficients and flux balances does not improve prediction accuracy, more detailed prediction of the fluxes in individual reactions by FBA/MOMA does require the knowledge of stoichiometric coefficients and maximal uptake rates.

Importantly, both flux balance and synthetic accessibility fail to predict viability of about 38% of mutants (in the combined dataset). Analysis of incorrect predictions (see AdditionalDocumentation.pdf) demonstrates well-known complexities of metabolism: the metabolic pathway used to produce a specific product is not always the shortest one; the system cannot be completely characterized by sets of input and output metabolites. Similar rates of failure of flux balance techniques suggest the importance of regulation in adaptation to mutations and the possible role of yet undiscovered metabolic and transport reactions.

We also explore other popular network characteristics like graph diameter, node degree and betweenness (usage) as predictors of mutant viability. Our results demonstrate that these characteristics fail to predict mutants' viability. We conclude, in agreement with a recent similar study [19], that node degree cannot be used to predict viability of metabolic knock-out strains.

The lack of predictive utility of node degree and graph diameter in metabolic networks is easy to understand. Both concepts have been widely applied to information exchange networks, like the internet and social networks, where every pair of nodes can potentially interact. On the contrary, the metabolic network is a transport network where products are being synthesized from a set of initial substrates. Performance of such a network is determined by its ability to synthesize products, and hence, paths from inputs to final products are of central importance, in contrast to diameter, where every pair of nodes is considered. Since chemical reactions can require more than one substrate to yield a product, the linear path used in information networks needs to be replaced by a tree of all required substrates. Considering these aspects naturally leads to the concept of synthetic accessibility to study metabolic and similar transport networks, e.g. signaling networks, which are also webs of reactions, in which the input is a chemical or physical stimulus and the output is a group of chemical responses to the stimulus. Synthetic accessibility defined this way is a generalization of graph diameter for directed, branching chemical reactions in an input-output transport network.

In summary, we show that the topology of the metabolic network is central in determining the viability of mutant strains and the success of widely-used flux balance techniques in predicting viability should be primarily attributed to topology. The addition of stoichiometric and other parameters does not significantly improve the accuracy of predictions, though they may be used by FBA to predict fluxes in individual reactions.. We introduce the concept of synthetic accessibility, which allows fast, accurate and easily interpretable analysis of metabolic networks. Our results suggest that re-routing of metabolic fluxes plays minimal role in providing viability of mutant strains. Importantly, our results strongly support the central role of network topology in determining phenotypes of biological systems.

Materials and Methods

Definition of synthetic accessibility. Consider a metabolic network that has access to certain inputs: substrates consumed from the environment (e.g. sugars, oxygen, and nitrogen), with the aim of producing certain outputs: amino acids, nucleotides and other components collectively called the biomass [20]. We define the synthetic accessibility S_j of an output j as the minimal number of metabolic reactions needed to produce j from the network inputs (Figure 1). S_j is set to infinity if j cannot be synthesized from the network inputs. Summing the synthetic accessibility over all components of the biomass, we obtain the total synthetic accessibility $S = \sum_i S_i$ of the biomass. We propose that if an enzyme knock-out does not change S, i.e. the biomass can be produced without extra metabolic cost, the mutant is viable. And if $S = \infty$, at least one essential component of the biomass cannot be produced from network inputs, causing a lethal phenotype.

Construction of the graphical metabolism model. The reactions included in the metabolic network are taken from [3]. Though there is an updated version of this metabolic network available [6], we chose to use the previous version to enable the comparison of synthetic accessibility performance to previous studies [3, 7-9]. Each reaction and metabolite is represented as a node, and directed edges connect reactants to reactions and reactions to products, therefore accounting for the reversibility of reactions.

Selection of input and output metabolite sets. The input metabolites are comprised of an energy source (glucose, acetate, glycerol or succinate), the components of minimal media, a sulfur source, carbon dioxide and oxygen, nicotinamide mononucleotide, and the regulatory protein thioredoxin (Table 2). The output metabolites are taken from the components of biomass (Table 3) [20].

Synthetic accessibility algorithm. To determine the synthetic accessibility of the outputs given the inputs, we use a type of iterative breadth first search, similar to the previously-described “forward-firing” (Figure 1) [21]. The algorithm starts by examining all the reactions that require one of the given input metabolites as a reactant. It then marks the reactions for which all the reactants are available “accessible” and marks all the metabolites produced by these reactions “accessible,” as well. The algorithm examines all the reactions that require one of the newly-marked metabolites as a starting material, determines whether each reaction is accessible or not based on the availability of its reactants and so on until no new metabolites are marked accessible. Concurrently, the number of steps needed to reach each accessible metabolite j, its synthetic accessibility S_j, is recorded; the synthetic accessibility of the network S is calculated by summing the synthetic accessibilities of all outputs.

Comparison to other predictive approaches. To compare the results of our approach to the smaller [3] and insertional mutant datasets [7], we create adjacency matrix, which represents the wild-type metabolic network topology. Then, for each mutant strain, we create a "mutated" adjacency matrix by removing all the reactions catalyzed by the mutated gene. As per the previous papers, for reactions catalyzed by multiple isozymes, we delete all corresponding genes. We then calculate the viability of each mutant and compare the results to the experimental data (DataTables1.xls, DataTable2.xls). If $S_{mutant} = S_{wild\ type}$, we predict that the mutant is viable, else we predict it is inviable. In the insertional mutant dataset, phenotype data is given as competitive growth rates. A mutant is considered negatively selected (or inviable) if there was a twofold decrease in growth rates over thirty generations [7].

Calculation of other topology-based predictions. We explore a number of other topology-based measures as predictors of *E. coli* mutant viability, including node degree, diameter, and node usage. The degree of each enzyme is calculated by summing the degree of all the reactions catalyzed by the enzyme and its isozymes. We define network diameter as the sum of all metabolites-versus-all metabolites shortest paths, and for each mutant, we calculate the change in network diameter from wild type. We define node usage for each enzyme as the number of times the reactions catalyzed by each enzyme is used to produce biomass in the wild-type strain, according to the synthetic accessibility approach, which is essentially analogous to betweenness [17, 18]. For each measure, degree, diameter, and usage, we predict an enzyme to be essential (and therefore, the corresponding mutant stain to be inviable), when the measure is greater than a given cutoff. We then vary the cutoff over the entire range of possible values to find a value that gives an optimal performance, as measured either by accuracy or significance of the χ^2 statistic (DataTable3.xls).

Quantitative analysis of performance. To assess the performance of synthetic accessibility and other methods in predicting the phenotype of mutant stains, we use four measures: accuracy, sensitivity, specificity, and the p-value of the χ^2 statistic. We define accuracy as $(TP + TN)/(TP + TN + FP + FN)$, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. We define positives and negatives in terms of the experimentally-measured phenotypes, where positives are viable strains and negatives are inviable strains, though the assignment is arbitrary and may be reversed. In a similar fashion, we define sensitivity as $TP/(TP + FP)$ and specificity as $TN/(TN + FN)$. To calculate the χ^2 statistic, we use two-by-two contingency tables that sort each mutant strain based on the *in silico* and *in vivo* phenotypes, and then calculate the appropriate p-value.

Assessment of synthetic accessibility robustness. To test the robustness of our approach, we introduce random mistakes into the network by randomly re-assigning a certain fraction of enzymes to unrelated reactions. We then measure the performance of synthetic accessibility in the erroneous network by plotting accuracy against the percentage of shuffled assignments.

Acknowledgements

We thank V. Spirin for his help in processing the raw data and comments on the manuscript, M. Kardar for his suggestions about testing for robustness, and V. Berube and M. Slutsky for comments on the manuscript. We also thank A. Trusina and K. Sneppen for useful discussions. Z.W. is a recipient of a Howard Hughes Medical Institute Predoctoral Fellowship. L.A.M. is an Alfred P. Sloan Research Fellow.

References

1. Varma A, Palsson BO: **Metabolic Flux Balancing - Basic Concepts, Scientific and Practical Use**. *Bio-Technology* 1994, **12**(10):994-998.
2. Edwards JS, Palsson BO: **Systems properties of the *Haemophilus influenzae* Rd metabolic genotype**. *J Biol Chem* 1999, **274**(25):17410-17416.
3. Edwards JS, Palsson BO: **The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities**. *Proc Natl Acad Sci USA* 2000, **97**(10):5528-5533.
4. Edwards JS, Palsson BO: **Metabolic flux balance analysis and the in silico analysis of *Escherichia coli* K-12 gene deletions**. *BMC Bioinformatics* 2000, **1**(1):1.
5. Forster J, Famili I, Fu P, Palsson BO, Nielsen J: **Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network**. *Genome Res* 2003, **13**(2):244-253.
6. Reed JL, Vo TD, Schilling CH, Palsson BO: **An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR)**. *Genome Biol* 2003, **4**(9):R54.
7. Badarinarayana V, Estep PW, 3rd, Shendure J, Edwards J, Tavazoie S, Lam F, Church GM: **Selection analyses of insertional mutants using subgenic-resolution arrays**. *Nat Biotechnol* 2001, **19**(11):1060-1065.
8. Segre D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks**. *Proc Natl Acad Sci USA* 2002, **99**(23):15112-15117.

9. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED: **Metabolic network structure determines key aspects of functionality and regulation**. *Nature* 2002, **420**(6912):190-193.

10. Klamt S, Stelling J: **Combinatorial complexity of pathway analysis in metabolic networks**. *Mol Biol Rep* 2002, **29**(1-2):233-236.

11. Albert R, Jeong H, Barabasi AL: **Error and attack tolerance of complex networks**. *Nature* 2000, **406**(6794):378-382.

12. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks**. *Nature* 2001, **411**(6833):41-42.

13. Myatt GJ: **Computer Aided Estimation of Synthetic Accessability**. *Ph.D. Thesis.* University of Leeds; 1994.

14. Papp B, Pal C, Hurst LD: **Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast**. *Nature* 2004, **429**(6992):661-664.

15. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks**. *Nature* 2000, **407**(6804):651-654.

16. Ma H, Zeng AP: **Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms**. *Bioinformatics* 2003, **19**(2):270-277.

17. Newman ME: **Scientific collaboration networks. I. Network construction and fundamental results**. *Phys Rev E* 2001, **64**(1 Pt 2):016131.

18. Newman ME: **Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality**. *Phys Rev E* 2001, **64**(1 Pt 2):016132.

19. Mahadevan R, Palsson BO: **Properties of Metabolic Networks: Structure versus Function**. *Biophys J* 2005, **88**(1):L07-09.

20. Neidhardt FC, Umbarger HE. In: *Escherichia coli and Salmonella: Cellular and Molecular Biology.* Edited by Neidhardt FC, Curtis R. Washington, D.C.: ASM Press; 1996: 13-16.

21. Romero PR, Karp P: **Nutrient-related analysis of pathway/genome databases**. *Pac Symp Biocomput* 2001:471-482.

Abbreviations

EMA, elementary mode analysis; FBA, flux balance analysis; MOMA, minimization of metabolic adjustment

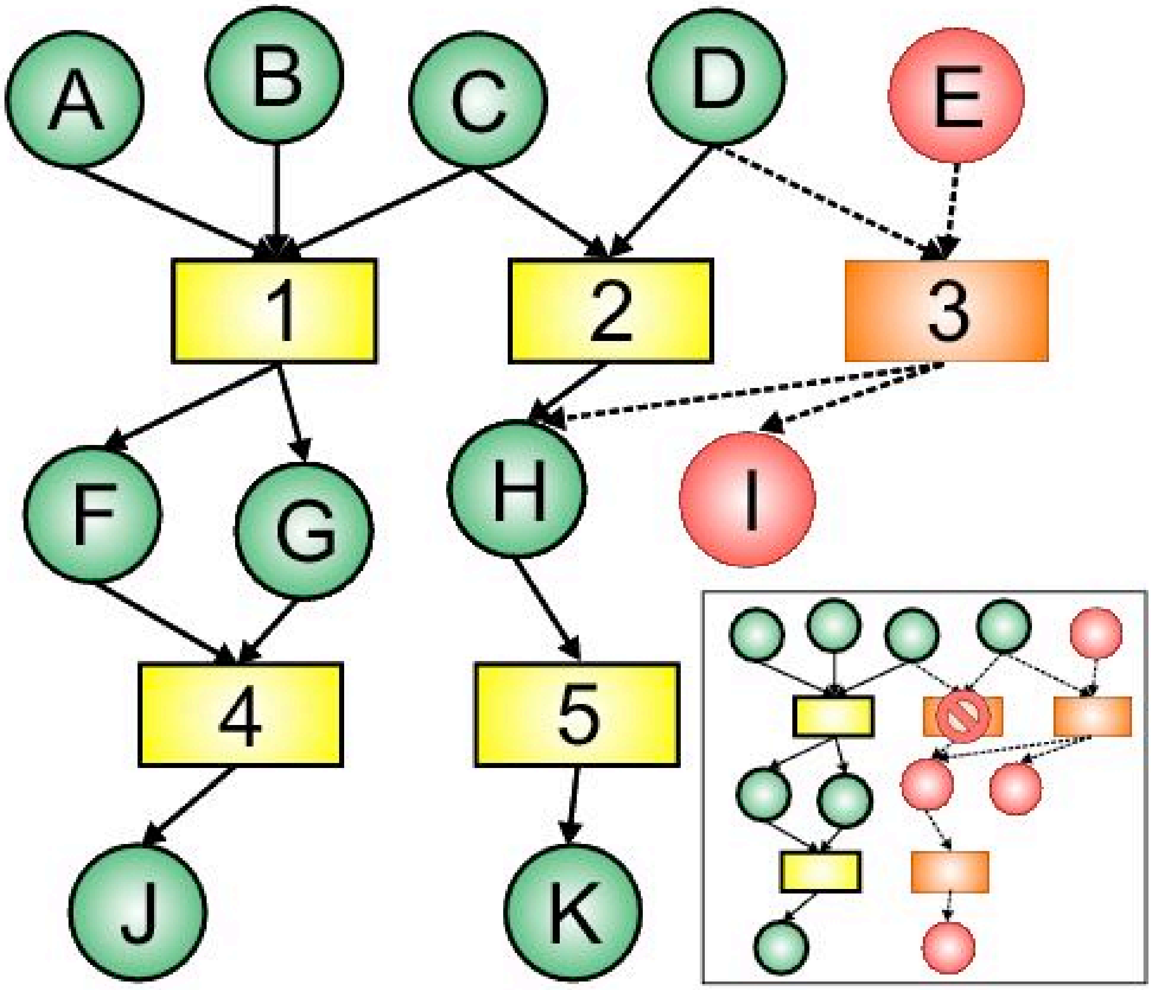

Figure 1 An illustration of the synthetic accessibility approach. In this representation of the metabolic network, circular nodes represent metabolites, rectangular nodes represent reactions, and directed edges indicate their relationships. Nodes with a thick outline are synthetically accessible, and nodes with a thin outline are not accessible. The algorithm begins by identifying all the reactions that neighbor the input metabolites (nodes A through D) and marking the reactions for which all the reactants are available as accessible (reactions 1 and 2). All the products of these reactions are marked accessible (nodes F through H). The algorithm then examines the neighboring reactions of the newly-marked metabolites as in the first step and continues until no new metabolites are marked accessible. The inset demonstrates what happens if the gene that produces the enzyme that catalyzes reaction 2 were deleted – metabolites H and K and reaction 5 would not be accessible anymore.

We define synthetic accessibility, *S*, as the number of reactions required to transform a set of inputs into a set of outputs. Synthetic accessibility is analogous to the diameter of a directed graph, but in contrast to graph diameters, synthetic accessibility takes into account branching nature of chemical reactions and the purpose of metabolic networks, to produce outputs from inputs.

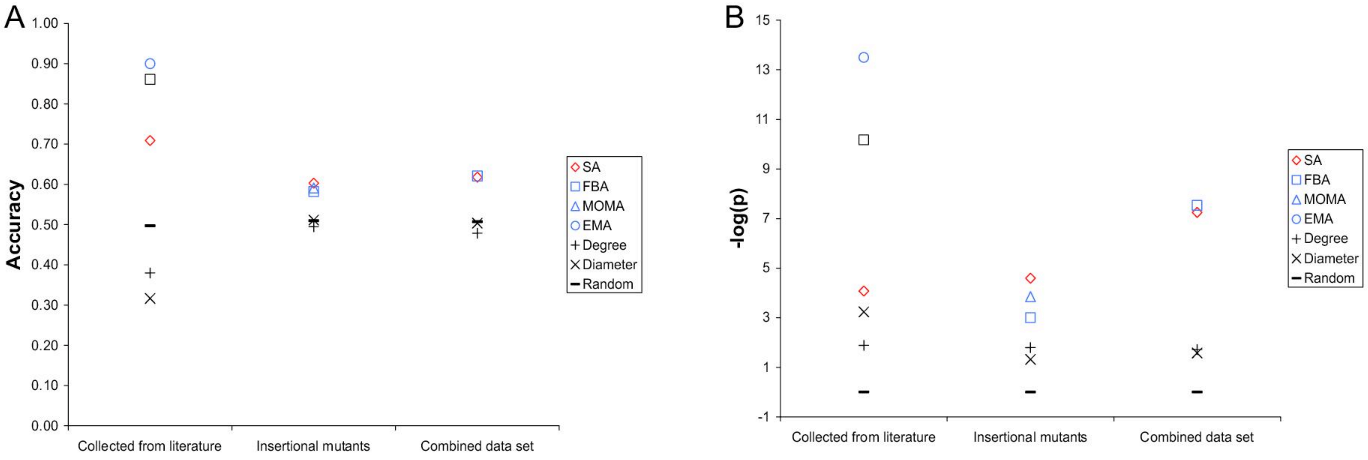

Figure 2 Performance of synthetic accessibility as compared to FBA, MOMA, EMA, and other topology-based measures. The graphs illustrate the relative performance of the techniques using two measures, accuracy, (*TP* + *TN*)/(*TP* + *TN* + *FP* + *FN*), and the negative log of the χ^2 statistic's p-value, which indicates the correlation between the *in silico* predictions and the *in vivo* observations of mutant strain viability. The χ^2 statistic is calculated using a contingency table like the one in Figure 3 for two datasets, the smaller dataset (79 data points, 90 data points for EMA), the insertional mutant dataset (487 data points), and the combined dataset (560 data points) [3, 7, 9]. When using the larger, more representative insertional mutant dataset or the combined dataset, synthetic accessibility is as accurate and statistically significant as FBA. However, synthetic accessibility performs more poorly on the smaller dataset, probably because this dataset has few data points and only covers central metabolism, a small fraction of the whole metabolic network. The other topology-based measures, degree and diameter, perform worse than FBA, MOMA, EMA and synthetic accessibility, indicating that they poorly characterize the functioning of the metabolic network. The random predictions are made using the expected values produced for the FBA χ^2 test and represent the expected performance if there were no correlation between the *in silico* and *in vivo* predictions. They vary very little if the expected values for the other χ^2 tests are used.

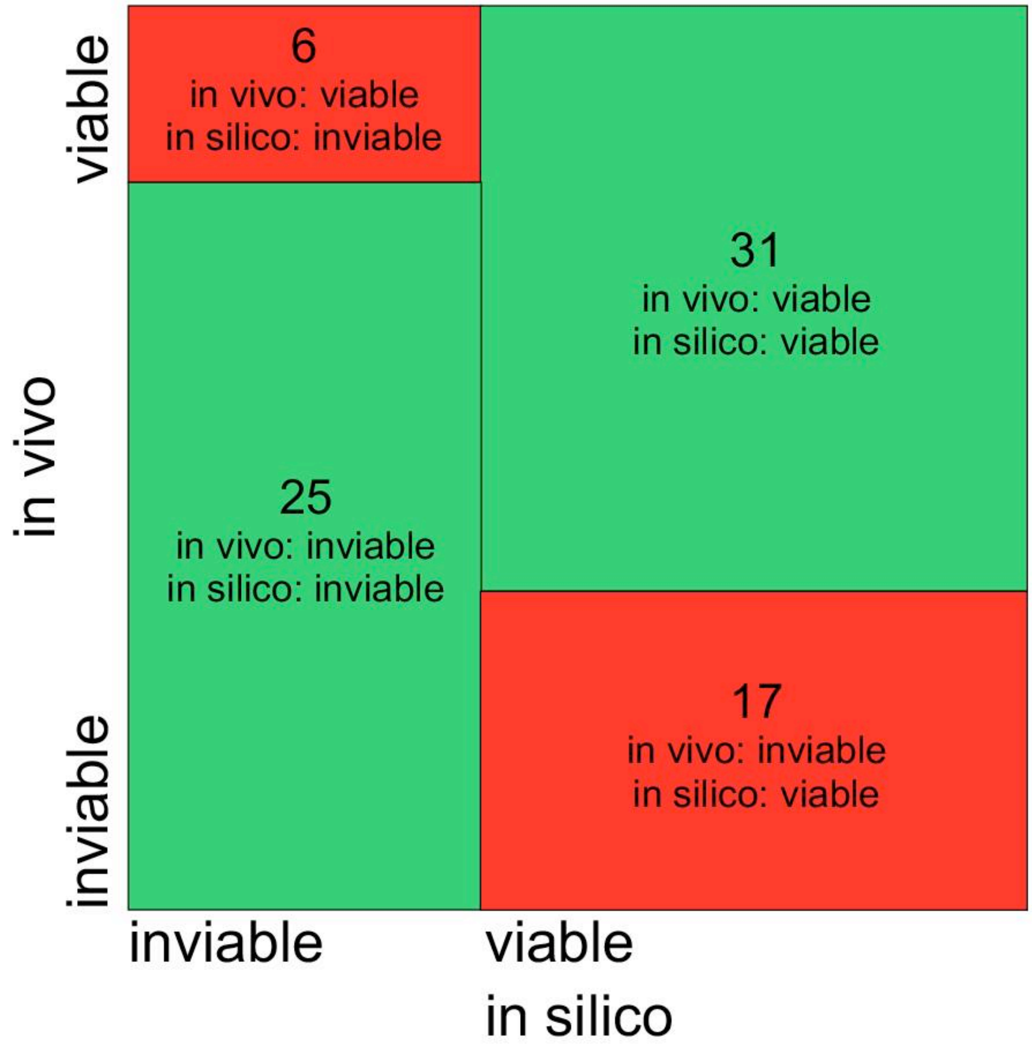

Figure 3 Results of the synthetic accessibility approach applied to the smaller mutant dataset [3]. This contingency graph allows the exploration of the types of errors that are most common. The x-axis represents the phenotypes predicted by the synthetic accessibility method, and the y-axis represents the experimental phenotypes. The green blocks correspond to cases where prediction matches experiment, and the red blocks correspond to errors. The area of each box is proportional to the number of cases in each category. From this diagram, we can see that the most common type of error is when the synthetic accessibility approach predicts the mutant viable when it is actually inviable.

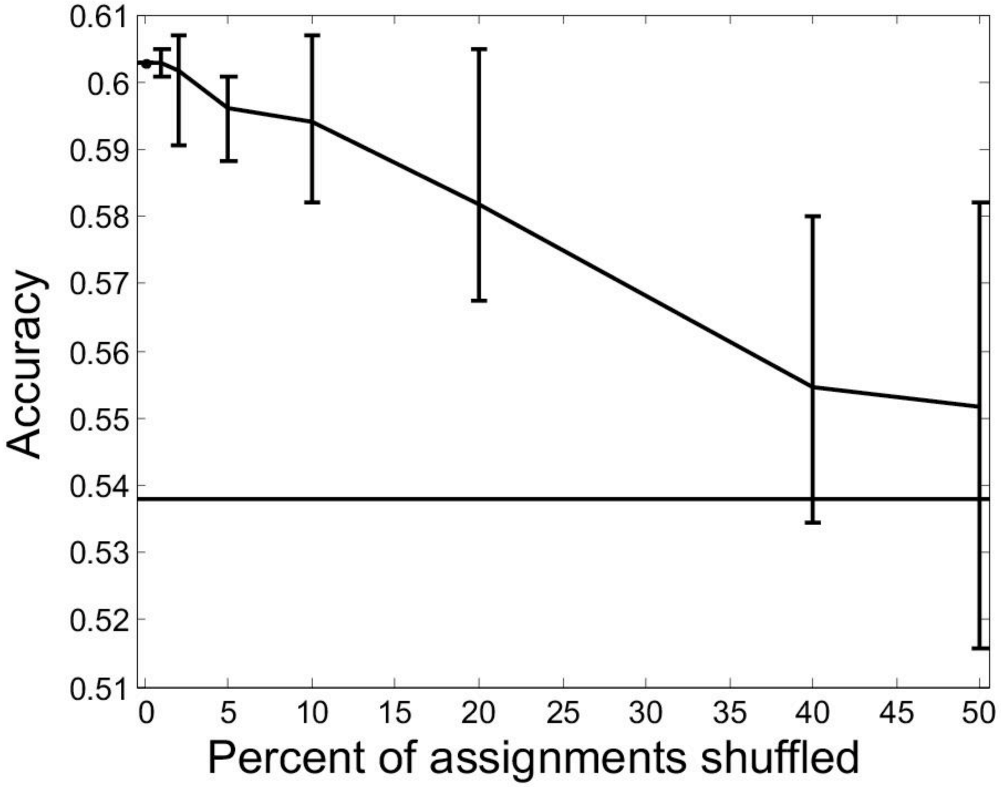

Figure 4 Accuracy of the synthetic accessibility approach with a percentage of enzyme-reaction assignments shuffled. To assess the robustness of the synthetic accessibility method to errors in the topology of the metabolic network, we randomly shuffle a given percentage of the assignments between enzymes and reactions and calculate the accuracy of the synthetic accessibility method for 10 trials. We plot average accuracy against the percent of assignments shuffled, with the error bars noting the minimum and maximum observed accuracy. The horizontal line denotes the accuracy of predicting all mutants to be viable – the best expected result in a random network. The approach is relatively robust to random errors in the enzyme-reaction assignments, although there is a clear and expected trend towards lower accuracy and great variability in accuracy as the number of shuffled assignments increases.

Figure 5. Plot of sensitivity and specificity for synthetic accessibility and other prediction methods.

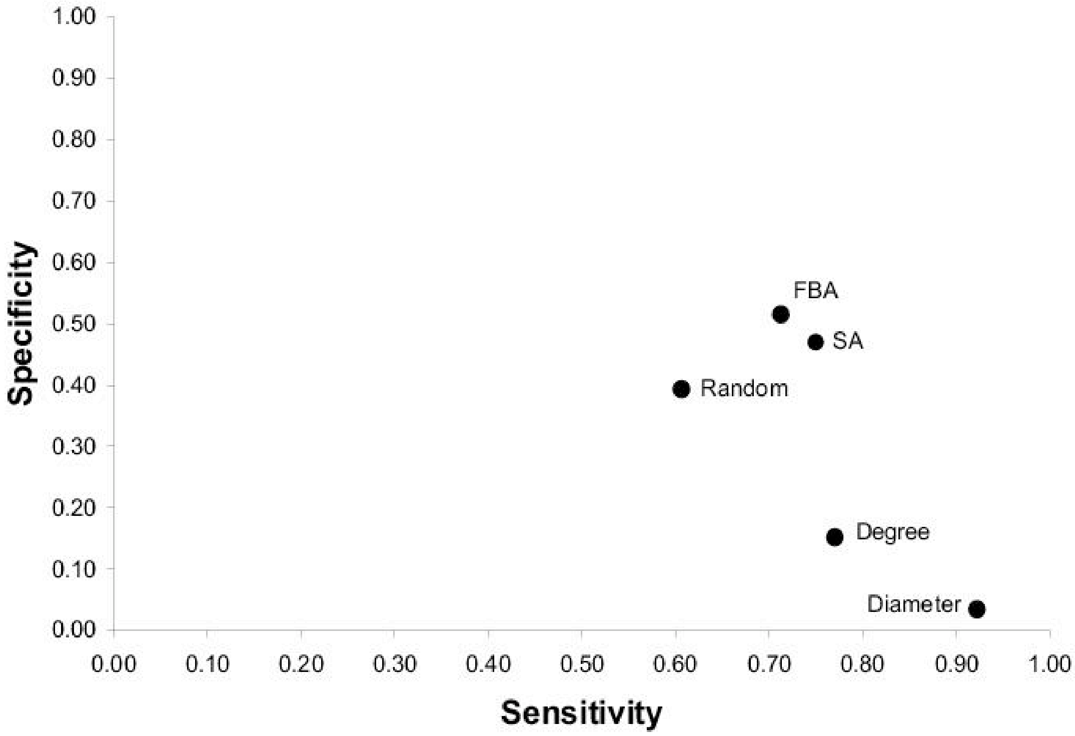

Figure 5 Plot of sensitivity and specificity for synthetic accessibility and other prediction methods. For the combined dataset (560 data points), sensitivity, *TP*/(*TP* + *FP*), and specificity, *TN*/(*TN* + *FN*), are calculated for the predictions made using synthetic accessibility, FBA, degree and diameter. The cutoff values for degree and diameter are selected to minimize the χ^2 test p-value. The random values are calculated using the expected values calculated for the χ^2 test for FBA and are essentially the same if the values for synthetic accessibility are used instead. Though both degree and diameter give good sensitivity, their specificity is quite low. Both synthetic accessibility and FBA have more moderate values for sensitivity and specificity. In all cases, the sensitivity is always greater, implying the viable predictions are more reliable than the inviable predictions, as can also be seen in Figure 3.

Tables

Table 1 Comparison of the accuracy and statistical significance of the FBA, MOMA, EMA and synthetic accessibility methods

		Method			
Mutant Data Source	Number of Data Points	Synthetic Accessibility	FBA	MOMA	EMA
Collected from literature	79	71%, $8x10^{-5}$ *	86%, $7x10^{-11}$ [3]		90%, $3x10^{-14}$ [9]
Insertional mutants	481	60%, $3x10^{-5}$	58%, $1x10^{-3}$ [7]	59%, $1x10^{-4}$ [8]	
Combined datasets	560	62%, $6x10^{-8}$	62%, $3x10^{-8}$		

* Accuracy, χ^2 statistic p-value

Table 2 Input metabolites

Ammonia, external (NH_3)

Carbon dioxide (CO_2)

Coenzyme A (CoA)

Hydrogen, external (H^+)

Inorganic phosphate, external (PO_4^-)

Nicotinamide mononucleotide, external (NMN)

Oxygen, external (O_2)

Phosphoribosyl pyrophosphate (PRPP)

Potassium, external (K^+)

Sodium, external (Na^+)

Sulfate, external (SO_4^-)

Thioredoxin, oxidized

Energy source (glucose, glycerol, succinate or acetate)

Table 3 Components of E. coli biomass*

Category	Number
Amino acids	22
Nucleotides	9
Cofactors	9
Cell membrane constituents	5
Carbohydrates	2
Total	47

* Adapted from [20].

Table 4 Mutants predicted to be inviable by synthetic accessibility approach, divided by reason for predicting inviability

Reason for predicting inviability	Correct (percent)	Incorrect (percent)
Number of accessible outputs < wild-type	89 (59%)	63 (41%)
S > wild-type	10 (67%)	5 (33%)

Additional Files

File Name	File Format	Title
AdditionalDocumentation.pdf	PDF	Analysis of incorrect predictions
DataTable1.xls	MS Excel	Synthetic Accessibility Applied to Edwards, Palsson 2000 Dataset
DataTable2.xls	MS Excel	Synthetic Accessibility Applied to Badarinarayana, *et al.* 2001 Dataset
DataTable3.xls	MS Excel	Enzyme degree, enzyme usage, and network diameter of Badarinarayana, *et al.* 2001 Dataset