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

Zeba Wunderlich* and Leonid A. Mirny†
*Biophysics Program, Harvard University, Cambridge, Massachusetts; and †Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts

ABSTRACT 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. Although 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, are more successful in predicting phenotypes of knockout strains. We reconcile these seemingly conflicting results by showing that the topology of the metabolic networks of both *Escherichia coli* and *Saccharomyces cerevisiae* are, in fact, sufficient to predict the viability of knockout strains with accuracy comparable to flux balance analysis on large, unbiased mutant data sets. 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 such as node degree, graph diameter, and node usage (betweenness) fail to predict the viability of *E. coli* 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. Our results strongly support a link between the topology and function of biological networks and, in agreement with recent genetic studies, emphasize the minimal role of flux rerouting in providing robustness of mutant strains.

INTRODUCTION

Many have suggested and debated the idea that topology determines network function. Although structures of several biological networks are available, it remains hard to separate the contributions of topology from the contributions of kinetic and equilibrium parameters. Because of their well-established structures and the wealth of related experimental data, the *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks are perfect model systems to explore the role of network topology. Is topology of a metabolic network sufficient to predict the viability of knockout mutants?

Metabolic networks have been modeled extensively using steady-state flux balance approaches (1–8). 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) (3,4,6–9), the related minimization of metabolic adjustment (MOMA) method (10), and elementary mode analysis (EMA) (11). Although these methods have been shown to be 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 as well as the viability of single-gene knockout strains. EMA can predict the viability of mutant strains with comparable accuracy. Because these methods use the network topology, the stoichiometry of metabolic chemical and transport reactions, and in some cases, the maximal rates of some of the reactions, they cannot separate the role of topology from the role played by other parameters in network function. In addition, because of the complexity of the method and the results, EMA techniques are computationally expensive (12) and provide little insight into why certain mutations are lethal, whereas 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 both *E. coli* and *S. cerevisiae* mutant strains as accurately as FBA on large, unbiased sets of mutants (9,13,14). This result supports the claim that topology plays a central role in determining network function and malfunction (15,16). We employ a novel network property, synthetic accessibility, and an intuitive and transparent way of understanding the effects of metabolic mutation (Fig. 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 that result from 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 and is similar in spirit to the ''scope'' of

Submitted December 30, 2005, and accepted for publication May 24, 2006.

Address reprint requests to Leonid Mirny, Harvard-MIT Div. of Health Sciences & Technology, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139. Tel.: 617-452-4862; Fax: 617-253-2514; E-mail: leonid@mit.edu.

0006-3495/06/09/2304/08 $2.00

doi: 10.1529/biophysj.105.080572

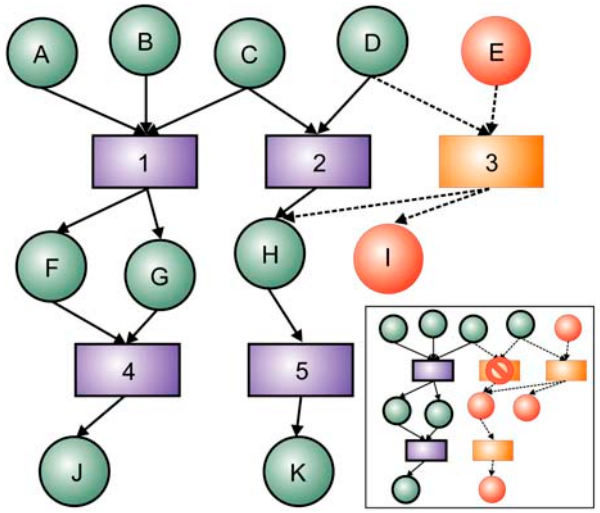

FIGURE 1 An illustration of the synthetic accessibility approach. In this representation of the metabolic network, circular nodes represent metabolites, rectangular nodes represent reactions, and directed edges indicate their relationships. Nodes with a thick outline (*green or blue*) are synthetically accessible, and nodes with a thin outline (*red or orange*) are not accessible. The algorithm begins by identifying all the reactions that neighbor the input metabolites (*nodes A–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–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 any more. 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 the branching nature of chemical reactions and the purpose of metabolic networks, to produce outputs from inputs.

metabolites (17,18). We also demonstrate that other network characteristics such as node degree or change in the graph diameter are unable to predict the viability of *E. coli* mutant strains better than random predictions, suggesting synthetic accessibility is a more appropriate characteristic for networks with directed transport, such as metabolic networks.

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 such as amino acids, nucleotides, and other components collectively called the biomass. 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 (Fig. 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 knockout does not change S, i.e., the biomass can be produced without extra metabolic cost, the mutant is viable. If $S = \infty$, at least one essential component of the biomass cannot be produced from network inputs, and therefore we predict a lethal phenotype.

Construction of the graphic metabolism model

The reactions included in the *E. coli* metabolic network are taken from Edwards and Palsson (4), and the reactions included in the yeast metabolic network are taken from Duarte et al. (8). Although there is an updated version of the *E. coli* metabolic network available (6), we chose to use the previous version to enable the comparison of synthetic accessibility performance to previous studies (4,9–11). Each reaction and metabolite is represented as a node, and directed edges connect reactants to reactions and reactions to products, thereby accounting for the reversibility of reactions.

Selection of input and output metabolite sets

The input metabolites for *E. coli* minimal medium, *E. coli* rich medium, and the various yeast medium conditions are listed in Supplementary Material, Tables S1–S4. *E. coli* minimal medium consists of an energy source (glucose, acetate, glycerol, or succinate), the components of minimal medium, a sulfur source, carbon dioxide and oxygen, nicotinamide mononucleotide, and the regulatory protein thioredoxin (Supplementary Material, Table S1). The input metabolites are chosen to match the real composition of minimal medium as closely as possible. Nicotinamide mononucleotide and thioredoxin are included to ensure that, in the wild-type network, all components of the output biomass are accessible. They are chosen specifically because they are the most upstream metabolites of the biomass synthesis pathways. *E. coli* rich medium consists of all the metabolites in minimal medium along with biotin, riboflavin, pantoate, pyridoxine, thiamin, dihydrofolate, *p*-aminobenzoic acid, all 20 amino acids, and the three nucleotide bases included as external metabolites in the metabolic network (external thymine was not in the metabolic network). Rich medium is difficult to model accurately, but using slightly different input metabolite sets has no significant effect on the results (results not shown).

The input metabolites for yeast are all based on the descriptions in Duarte et al. (8) and include histidine, leucine, and uracil to compensate for the deletions of the His-3/Leu-2/Ura-3 in the mutant strains. Additionally, thioredoxin (oxidized), H^+ (in the endoplasmic reticulum), NADPH (in the endoplasmic reticulum), and dolichol are included as inputs, for without them, some of the components of biomass are not producible, even in the wild-type network.

The *E. coli* output metabolites are taken from the components of *E. coli* biomass (Supplementary Material, Table S5) (19). The yeast output metabolites are the components of the biomass reaction reported in (8).

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'' (Fig. 1) (20). 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 experimental results

To compare the results of our approach to the experimental data sets, we first create an 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 gene. For all

E. coli predictions, as per the previous papers, we delete all corresponding genes for reactions catalyzed by multiple isozymes. We then calculate the viability of each mutant and compare the results to the experimental data (see Supplementary Material). If $S_{\text{mutant}} = S_{\text{wild-type}}$, we predict that the mutant is viable; else we predict it is inviable. In the *E. coli* insertional mutant data set, phenotype data are given as competitive growth rates. A mutant is considered negatively selected (or inviable) if there was a twofold decrease in growth rates over 30 generations (9). For the Gerdes et al. data set (21), we create mutated adjacency matrices only for genes included in the metabolic network model, resulting in 598 mutated adjacency matrices. For the yeast experimental data, we use the preprocessed data set created in Duarte et al. (8) and do not simultaneously delete isozymes.

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 degrees 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 are used to produce biomass in the wild-type strain, according to the synthetic accessibility approach, which is essentially analogous to betweenness (22,23). 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.

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 *E. coli* network by randomly reassigning 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.

RESULTS

Performance of synthetic accessibility on the *E. coli* metabolic network

To study the performance of synthetic accessibility in predicting viability of knockout strains and compare it to previous studies, we first applied the method to the *E. coli* metabolic network. We initially tested it on two data sets used in previous studies: a large, unbiased data set of insertional (transposon-induced) mutants (9) and a smaller data set collected for FBA analysis (4), which mainly contained knockouts of enzymes involved in central metabolism. All mutants were grown on minimal medium. We used these data sets specifically because they were used in previous studies (4,9–11), to which we compared our results. We also used the union of these data sets and refer to it below as the combined data set. When applied to the combined data set, 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, Fig. 2 for details). On the large data set of 487 insertional mutants (9), 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 data set of 79 mutants (4), FBA correctly predicted 86% of the cases, whereas our topology-based synthetic accessibility approach had 71% accuracy, providing correct predictions for $53/68 = 78\%$ of the cases predicted correctly by FBA.

The difference in the performance of the synthetic accessibility approach as compared to FBA between the first two data sets is probably related to the way the data sets were interpreted and the cases included in the data sets. In the smaller data set, the mutant strains are classified as viable or inviable, whereas in the insertional data set, the mutants are labeled 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. Because the synthetic accessibility approach deems a mutant strain inviable or negatively selected based on the path

TABLE 1 Comparison of the accuracy and statistical significance of the FBA, MOMA, EMA, and synthetic accessibility methods applied to the *E. coli* metabolic network

		Method			
Mutant data source	No. of cases	Synthetic accessibility	FBA	MOMA	EMA
Collected from literature	79	71%, 8×10^{-5}*	86%, 7×10^{-11} (4)	–	90%, 3×10^{-14} (11)
Insertional mutants	481	60%, 3×10^{-5}	58%, 1×10^{-3} (9)	59%, 1×10^{-4} (10)	–
Combined data sets	560	62%, 6×10^{-8}	62%, 3×10^{-8}	–	–
Gerdes data set	598	74%, 1×10^{-5}	–	–	–

*Accuracy, *p*-value of χ^2 statistic.

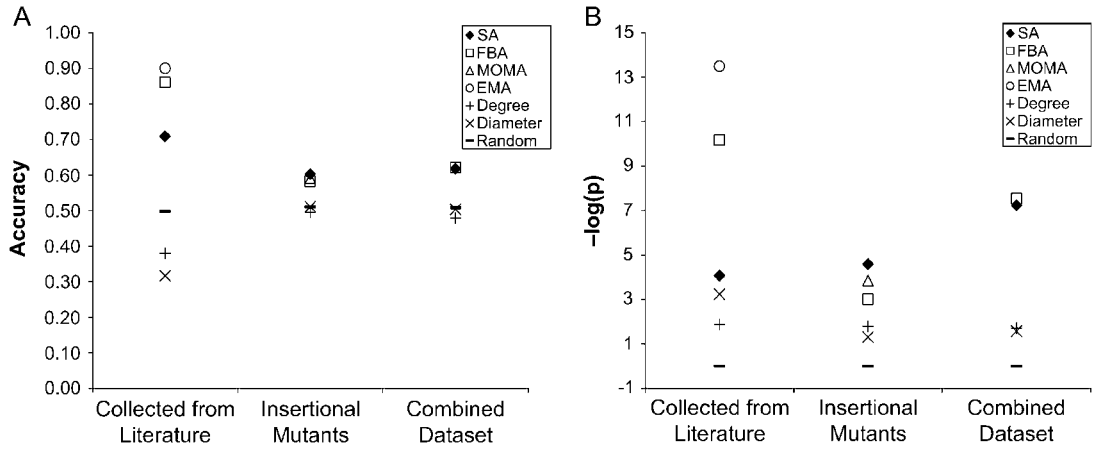

FIGURE 2 Performance of synthetic accessibility as compared to FBA, MOMA, EMA, and other topology-based measures using the *E. coli* metabolic network. The graphs illustrate the relative performance of the techniques using two measures, accuracy, $(TP + TN)/(TP + TN + FP + FN)$, and the negative log of the χ^2 statistic's *p*-value, which indicates the correlation between the in silico predictions and the in vivo observations of *E. coli* mutant strain viability. The χ^2 statistic is calculated using a contingency table like the ones in Fig. 3 for the smaller data set (79 data points, 90 data points for EMA), the insertional mutant data set (487 data points), and the combined data set (560 data points) (4,9,11). When the larger, more representative insertional mutant data set or the combined data set is used, synthetic accessibility is as accurate and statistically significant as for FBA. However, synthetic accessibility performs more poorly on the smaller data set, probably because this data set 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.

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 may correspond to reduced growth rates rather than inviability.

The number and type of data points included in the data sets are also different. The insertional data set is much larger (487 vs. 79 data points) and includes a fairly random collection of insertions in metabolic genes, whereas the smaller data set contains data about only the enzymes used in the central metabolism (glycolysis, pentose phosphate pathway, citric acid cycle, respiration processes) (4). 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 more poorly than FBA when applied to the smaller data sets.

In regard to the combined data set, synthetic accessibility had greater sensitivity, indicating that 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 (Figs. 3 and 4). The success of synthetic accessibility on the combined data set demonstrates three important results, making transparent the difference between most of viable and nonviable strains.

1. Most nonviable 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 2).
2. Our approach correctly predicted that most strains with longer rerouted pathways are inviable, suggesting that rerouting of metabolic fluxes plays a small role in rescuing mutant strains. This result is consistent with results of FBA analysis of yeast mutants (24).
3. Most viable mutants have either untouched primary synthetic pathways or only short rerouting (e.g., because of isozymes).

Although it has not been used in previous FBA studies, we also applied the synthetic accessibility approach to the large-scale knockout study by Gerdes et al. (21), which identified genes essential for robust growth on rich medium using a genetic footprinting technique based on transposon-based mutagenesis. The synthetic accessibility approach performed well on this data set (74% accuracy, $p = 1 \times 10^{-5}$).

Performance of synthetic accessibility on the yeast metabolic network

To ensure that the success of the synthetic accessibility method was not limited to the *E. coli* metabolic network, we tested the method on the metabolic network of *S. cerevisiae*, another metabolic network that has been reconstructed by hand (8). This reconstruction has been extensively validated by the use of FBA to predict the phenotypes of a large number of single-gene knockout yeast strains grown under a variety of conditions (13,14). The conditions include glucose minimal medium (MMD) and rich medium with a defined

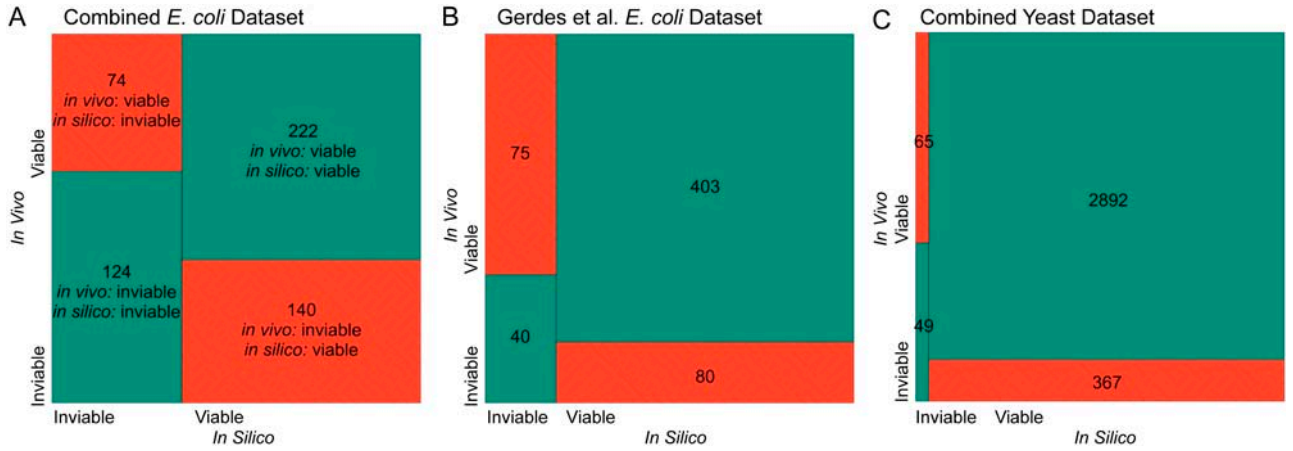

FIGURE 3 Results of the synthetic accessibility approach, divided by type of prediction. These contingency graphs allow the exploration of the types of errors that are most common. Results are reported for (*A*) the combined *E. coli* data set, (*B*) the Gerdes et al. (21) *E. coli* data set, and (*C*) the combined yeast data set (metabolic genes only). 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, hashed 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.

carbon source (YPGal, galactose; YPD, glucose; YPDGE, glucose-ethanol-glycerol, YPG, glycerol; YPE, ethanol; and YPL, lactate). Sets of essential and slow-growth genes were also identified experimentally as either genes for which mutant strains could not be constructed or genes that produced slow-growing mutant strains on rich (YPD) medium. The results (Table 3) for all the gene sets show, except the essential and slow sets, that synthetic accessibility performs comparably to FBA. When all the conditions are considered simultaneously, synthetic accessibility predicts phenotype with 83.7% accuracy, as compared to FBA with 82.6% accuracy.

We believe that the higher overall accuracy of synthetic accessibility and FBA when applied to the yeast metabolic network is probably largely a result of the way the data sets were used. For all the *E. coli* data sets, predictions were made only for knockout strains that involved genes that were included in the metabolic network model. For the yeast data sets, following the protocol of the previous FBA study (8), we made predictions for all strains, whether the gene was included in the metabolic network model or not. Because most genes are nonessential, and we predict knockouts of genes absent from the metabolic network model to be viable, this inflates the accuracy. We also report the accuracies for predictions of only metabolic gene knockouts in Table 3, and the accuracies are even higher in most cases.

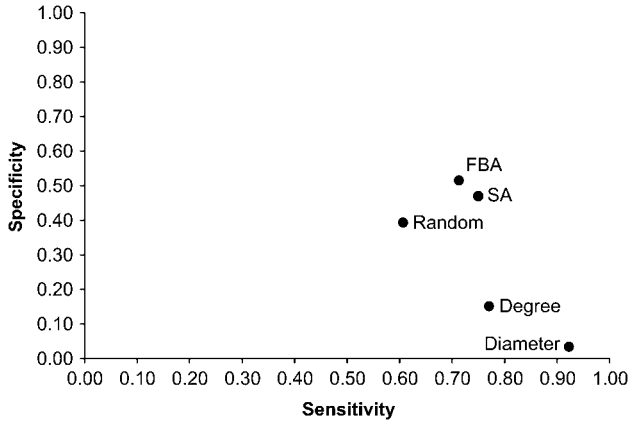

FIGURE 4 Plot of sensitivity and specificity for synthetic accessibility (SA) and other prediction methods. For the combined *E. coli* data set (560 data points), sensitivity, $TP/(TP + FP)$, and specificity, $TN/(TN + FN)$, are calculated for the predictions made using synthetic accessibility, FBA, 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 Fig. 3.

Performance of other topology-based measures on the *E. coli* metabolic network

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

TABLE 2 *E. coli* mutants predicted to be inviable by synthetic accessibility approach in the combined data set, divided by reason for predicting inviability

Reason for predicting inviability	Correct (percent)	Incorrect (percent)
No. of accessible outputs < wild-type ($S = \infty$)	89 (59%)	63 (41%)
$S >$ wild-type	10 (67%)	5 (33%)

TABLE 3 Accuracy of the synthetic accessibility and FBA methods for predicting viability of yeast deletion strains

	Data set									
	Essential	Slow	MMD	YPGal	YPD	YPDGE	YPG	YPE	YPL	All
No. of cases	118	83	564	564	565	565	565	565	565	4154
FBA (8)	31.4%	19.3%	84.0%	85.6%	84.4%	85.3%	86.5%	85.7%	86.4%	82.6%
Synthetic accessibility	11.9%	1.20%	94.0%	97.2%	85.3%	84.4%	83.7%	84.1%	84.1%	83.7%
No. of cases (only metabolic enzymes)	100	45	459	459	462	462	462	462	462	3373
FBA	33.0%	4.44%	95.0%	97.6%	87.9%	87.9%	89.2%	87.4%	88.3%	87.6%
Synthetic accessibility	14.0%	2.22%	94.3%	96.9%	88.5%	88.3%	89.0%	88.7%	88.7%	87.1%

E. coli mutant strains. Several studies have suggested that nodes that have higher degrees are more important for the network, and removal of such nodes in biological networks is more likely to lead to a lethal phenotype (15,16). 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. Fig. 2 shows 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 (15,25,26). To test graph diameter as a predictor of viability, we predicted a mutant to be inviable if increase in graph diameter exceeded a cutoff. Fig. 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, which is defined as the number of times the reactions catalyzed by each enzyme are used to produce biomass in the wild-type strain according to the synthetic accessibility approach. Enzyme usage is analogous to node betweenness, which is the number of shortest paths between all pairs of nodes that go through the node (22,23). Enzyme usage performed somewhat better than random predictions but worse than synthetic accessibility, which is not surprising because 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 although node usage performed better than node degree and diameter, it was a worse predictor than the synthetic accessibility (see Supplementary Material).

These characteristics ignore essential properties of a 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. Although 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 such as 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 using the *E. coli* metabolic network. Errors were modeled by random reassignment of certain percentages of enzymes to different reactions. Fig. 5 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

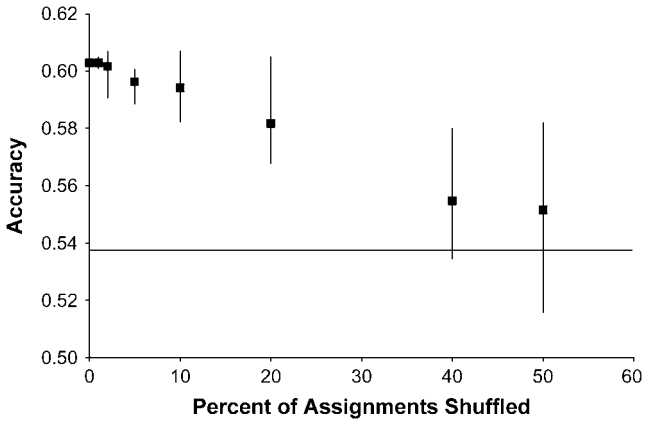

FIGURE 5 Accuracy of the synthetic accessibility approach with a percentage of enzyme-reaction assignments shuffled. To assess the robustness of the synthetic accessibility method to errors in the topology of the *E. coli* metabolic network, we randomly shuffle a given percentage of the assignments between enzymes and reactions and calculate the accuracy of the synthetic accessibility method for 10 trials. We plot average accuracy against the percentage 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 toward lower accuracy and great variability in accuracy as the number of shuffled assignments increases.

dropped to the level of random predictions when ~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 443 of 598 mutant *E. coli* strains from a comprehensive, reliable data set (21) and 3477 of 4154 mutant yeast strains grown under several conditions (13,14). 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 on 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 because 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 *E. coli* insertional data set, the accuracy of the synthetic accessibility approach is comparable to those of FBA and MOMA. The performance of synthetic accessibility as compared to FBA and EMA on the smaller *E. coli* data set is worse, but this smaller data set only has data for mutants affecting the central metabolism and therefore may be biased, whereas the large data set of insertional mutants is fairly unbiased and representative. Synthetic accessibility also performs comparably to FBA on the yeast data sets.

Unlike FBA, synthetic accessibility also does not assume optimality with regard to biomass production. But our model assumes that long rerouted fluxes are less efficient than native ones, predicting mutants with longer fluxes (larger synthetic accessibility) as inviable. Although this assumption fails in certain cases, 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 recent studies (24,27), that rerouting does not contribute significantly to robustness of knockout 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 that the 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 a significant number of mutants. Analysis of incorrect predictions in *E. coli* (see Supplementary Material) 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 such as 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 (28), that node degree cannot be used to predict viability of metabolic knockout 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, such as 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. Because 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 rerouting of metabolic fluxes plays a 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.

SUPPLEMENTARY MATERIAL

An online supplement to this article can be found by visiting BJ Online at http://www.biophysj.org.

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., and B. O. Palsson. 1994. Metabolic flux balancing—basic concepts, scientific and practical use. *Biotechnology*. 12:994–998.
2. Edwards, J. S., and B. O. Palsson. 1999. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.* 274:17410–17416.
3. Edwards, J. S., and B. O. Palsson. 2000. Metabolic flux balance analysis and the in silico analysis of *Escherichia coli* K-12 gene deletions. *BMC Bioinformatics*. 1:1.
4. Edwards, J. S., and B. O. Palsson. 2000. The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA*. 97:5528–5533.
5. Forster, J., I. Famili, P. Fu, B. O. Palsson, and J. Nielsen. 2003. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* 13:244–253.
6. Reed, J. L., T. D. Vo, C. H. Schilling, and B. O. Palsson. 2003. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* 4:R54.
7. Burgard, A. P., and C. D. Maranas. 2001. Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. *Biotechnol. Bioeng.* 74:364–375.
8. Duarte, N. C., M. J. Herrgard, and B. O. Palsson. 2004. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* 14:1298–1309.
9. Badarinarayana, V., P. W. Estep 3rd, J. Shendure, J. Edwards, S. Tavazoie, F. Lam, and G. M. Church. 2001. Selection analyses of insertional mutants using subgenic-resolution arrays. *Nat. Biotechnol.* 19:1060–1065.
10. Segre, D., D. Vitkup, and G. M. Church. 2002. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA*. 99:15112–15117.
11. Stelling, J., S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. 2002. Metabolic network structure determines key aspects of functionality and regulation. *Nature*. 420:190–193.
12. Klamt, S., and J. Stelling. 2002. Combinatorial complexity of pathway analysis in metabolic networks. *Mol. Biol. Rep.* 29:233–236.
13. Giaever, G., 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, et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*. 418:387–391.
14. Steinmetz, L. M., C. Scharfe, A. M. Deutschbauer, D. Mokranjac, Z. S. Herman, T. Jones, A. M. Chu, G. Giaever, H. Prokisch, P. J. Oefner, and R. W. Davis. 2002. Systematic screen for human disease genes in yeast. *Nat. Genet.* 31:400–404.
15. Albert, R., H. Jeong, and A. L. Barabasi. 2000. Error and attack tolerance of complex networks. *Nature*. 406:378–382.
16. Jeong, H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks. *Nature*. 411:41–42.
17. Myatt, G. J. 1994. Computer aided estimation of synthetic accessibility. PhD Thesis. University of Leeds, Leeds, UK.
18. Handorf, T., O. Ebenhoh, and R. Heinrich. 2005. Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J. Mol. Evol.* 61:498–512.
19. Neidhardt, F. C., and H. E. Umbarger. 1996. Chemical composition of *Escherichia coli*. In *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. F. C. Neidhardt and R. Curtis, editors. ASM Press, Washington, DC. 13–16.
20. Romero, P. R., and P. Karp. 2001. Nutrient-related analysis of pathway/genome databases. *Pac. Symp. Biocomput.* 471–482.
21. Gerdes, S. Y., M. D. Scholle, J. W. Campbell, G. Balazsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, et al. 2003. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* 185:5673–5684.
22. Newman, M. E. 2001. Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E*. 64:016131.
23. Newman, M. E. 2001. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*. 64:016132.
24. Papp, B., C. Pal, and L. D. Hurst. 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*. 429:661–664.
25. Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. 2000. The large-scale organization of metabolic networks. *Nature*. 407:651–654.
26. Ma, H., and A. P. Zeng. 2003. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*. 19:270–277.
27. Blank, L. M., L. Kuepfer, and U. Sauer. 2005. Large-scale ^{13}C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol.* 6:R49.
28. Mahadevan, R., and B. O. Palsson. 2005. Properties of metabolic networks: structure versus function. *Biophys. J.* 88:L07–L09.