ARTICLES

Minimal Reaction Sets for *Escherichia coli* Metabolism under Different Growth Requirements and Uptake Environments

Anthony P. Burgard, Shankar Vaidyaraman, and Costas D. Maranas*

Department of Chemical Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802

A computational procedure for identifying the minimal set of metabolic reactions capable of supporting various growth rates on different substrates is introduced and applied to a flux balance model of the *Escherichia coli* metabolic network. This task is posed mathematically as a generalized network optimization problem. The minimal reaction sets capable of supporting specified growth rates are determined for two different uptake conditions: (i) limiting the uptake of organic material to a single organic component (e.g., glucose or acetate) and (ii) allowing the importation of any metabolite with available cellular transport reactions. We find that minimal reaction network sets are highly dependent on the uptake environment and the growth requirements imposed on the network. Specifically, we predict that the *E. coli* network, as described by the flux balance model, requires 224 metabolic reactions to support growth on a glucose-only medium and 229 for an acetate-only medium, while only 122 reactions enable growth on a specially engineered growth medium.

Introduction

The recent explosion of fully sequenced genomes has brought significant attention to the question of how many genes are necessary to sustain cellular life. A minimal genome is generally defined as the smallest set of genes that allows for replication and growth in a particular environment (*1*). Attempts to uncover this minimal gene set include both experimental and theoretical approaches. Global transposon mutagenesis was used by Hutchison et al. (*2*) to determine that 265−350 of the 480 protein-coding genes of *Mycoplasma genitalium*, the smallest known cellular genome (580 kb), are essential for survival under laboratory growth conditions. Additional experimental work (*3*, *4*) revealed that only 12% and 9%, respectively, of the yeast and *Bacillus subtilis* genomes are essential for cellular growth and replication. Theoretical methods stem from the assumption that genes conserved across large evolutionary boundaries are vital to cellular survival. On the basis of this hypothesis, a minimal set of 256 genes was compiled by Mushegian and Koonin (*5*) by assuming that genes common to *M. genitalium* and *Haemophilus influenzae* must be members of a minimal genome. Interestingly, only 6 out of 26 *Escherichia coli* open reading frames of unknown function conserved in *M. genitalium* were deemed essential to species survival (*6*). The existence of multiple, quite different, species and environment specific minimal genomes has long been speculated (*7*).

Herein we describe a computational procedure for testing this claim by estimating the minimum required growth-sustaining core of metabolic reactions under different uptake conditions. The latest stoichiometric model of *E. coli* metabolism proposed by Palsson and coworkers (*8*) is employed to identify the smallest set of enzymatic reactions capable of supporting given targets on the growth rate for either a glucose, an acetate, or a complex substrate. This flux balance analysis (FBA) model incorporates 454 metabolites and 720 reactions including the glycolysis, tricarboxylic acid (TCA) cycle, pentose phosphate pathway (PPP), and respiration pathways, along with synthesis routes for the amino acids, nucleotides, and lipids. Growth is quantified by adding an additional reaction to the model simulating a drain on the various components of *E. coli* biomass in their appropriate biological ratios (*9*). By associating a gene to each metabolic reaction in the network, gene activations and inactivations are incorporated into the FBA model using logic 0−1 binary variables. The problem of minimizing the number of active metabolic reactions required to meet specific metabolic objectives (i.e., growth rates) is shown to assume the mathematical structure of a generalized network flow problem where nodes denote metabolites and connecting arcs represent reactions. Alternatively, instead of a biomass target, minimum levels of ATP production or lowest allowable levels of key components/metabolites could readily be incorporated in the model. A mixed-integer linear programming (MILP) solver, CPLEX 6.5 (*10*) accessed via GAMS (*11*), is employed to solve the resulting large-scale combinatorial problems with CPU times ranging from minutes to days.

On the basis of the *E. coli* model, the minimal reaction network is explored for different growth requirements under two contrasting uptake environments: (i) restricting the uptake of organic material to a single organic

* Ph: (814) 863-9958. Fax: (814) 865-7846. E-mail: costas@psu.edu.

10.1021/bp0100880 CCC: $20.00 © 2001 American Chemical Society and American Institute of Chemical Engineers
Published on Web 09/14/2001

component and (ii) allowing the uptake of any organic metabolite with a corresponding transport reaction. These two extreme uptake scenarios were chosen to model maximum and minimum reliance on internal metabolism for component synthesis, respectively, and probe their effect on the minimum reaction set required. Previous attempts utilized reductionist methodologies to extract the set of essential genes through a series of gene knockouts. Here we propose an efficient computational procedure for selecting the minimal set by simultaneously considering the effect of all reactions on cell growth. A minimal gene set can then be inferred by mapping the enzyme(s) catalyzing these reactions to the corresponding coding genes. While the obtained results are, in principle, dependent on the specifics of the employed flux balance *E. coli* model (*8*), they still provide valuable insight and perspective to the questions of what is the minimal genome and how is it shaped by the environment.

Results

The first case study involves identifying the minimal reaction set supporting *E. coli* growth on a glucose substrate. A detailed description of the employed modeling procedure is provided in the Appendix. A constrained amount of glucose (<10 mmol/gDW·h), along with unconstrained uptake routes for inorganic phosphate, oxygen, sulfate, and ammonia are enabled to fuel the metabolic network. Secretion routes for every metabolite capable of exiting the cell are also provided. Under these conditions, the FBA model (*8*) predicts that the *E. coli* reaction network is capable of achieving a maximum theoretical growth rate of 0.966 g biomass/gDW·h, which we will refer to as the maximum growth rate (MGR). By requiring the reaction network to match the MGR we determined that at least 234 reactions out of 720 are required for maximum growth on glucose.

The growth demands are then relaxed in subsequent studies to identify the minimal number of metabolic reactions required to meet various submaximal growth demands (% of MGR). Interestingly, the number of necessary metabolic reactions decreases only mildly with the falling growth demands imposed on the network, as indicated by Figure 1A. While a reaction set comprising 234 reactions is needed for maximum growth, the minimal reaction set corresponding to growth rates of 30% and lower involves only 224 reactions. The same minimal reaction set persists even for growth rates as low as 0.1% of the MGR. In general, the reaction set reductions are attained by successively eliminating energy-producing reactions occurring in (i) glycolysis, (ii) the TCA cycle, and (iii) the pentose phosphate pathway as the growth demands are lessened. However, certain reactions absent at higher growth rates enter the minimal sets at lower growth rates, suggesting a much more complex mechanism of flux redirection than successive reaction elimination. A detailed description of the reactions entering/leaving the minimal reaction set as the imposed growth requirements are lowered is provided in Table 1.

For comparison, a similar study enabling a constrained amount of acetate (<30 mmol/gDW·h) to enter the network instead of glucose was performed (see Figure 1A). Here the network is much less tolerant of reaction set reductions than in glucose study. While for a glucose substrate the minimal network sizes decrease from 234 to 224 reactions as the growth demands are lowered, for an acetate substrate the network sizes reduce only from 231 to 229 reactions. This implies that the minimal reaction set size is dependent not only on the imposed

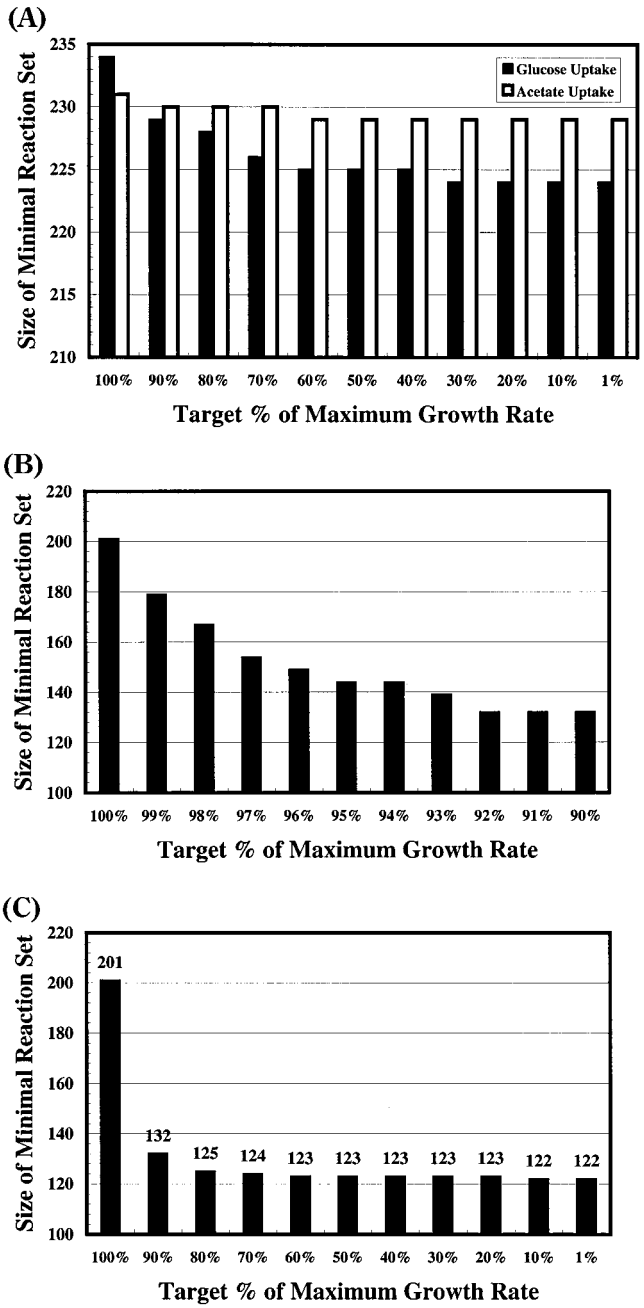

Figure 1. Number of reactions in each minimal set as a function of the imposed growth demands for (A) a glucose- or acetate-only uptake environment and (B, C) an uptake environment allowing multiple organic uptakes.

biomass production requirements but also on the specific choice for the single substrate.

It is important to note that neither the minimal reaction sets nor their corresponding reaction fluxes are unique. For example, for the 30% glucose uptake case we identified over 100 different minimal reaction sets containing exactly 224 enzymatic reactions without even counting the multiplicities associated with the 171 isoenzymes present in the network. Among most of these multiple minimal reaction sets, the activity and flux directions of the major pathways differ very little. Most variations are concentrated on the catabolic parts of the networks. For instance, while some minimal reaction sets secrete carbon dioxide, acetate, and fumarate as the only metabolic byproducts, other sets may also secrete varying amounts of formate, glycerol, and the amino acids phenylalanine and tyrosine. These results provide a compu-

Table 1. Evolution of Minimal Reaction Sets for Case (i) under Decreasing Growth Requirements

target % MGR	min reaction set (no. of reactions)	key features
100	234	The glycolysis, tricarboxylic acid cycle, and pentose phosphate pathways are all operating in their forward directions, optimally generating the energy cofactors ATP, NADH, and NADPH required for cell growth. All available glucose is oxidized into the cell's only secreted byproduct, carbon dioxide.
90	229	The fluxes through two TCA cycle reactions 2-ketoglutarate dehydrogenase and succinate dehydrogenase are zero while succinyl-CoA synthetase operates in its reverse direction suggesting a less demanding energetic state under the submaximal growth demands. Acetate is now secreted as a byproduct along with carbon dioxide.
80	228	Fluxes through two additional TCA cycle reactions, fumarase and malate dehydrogenase, are eliminated while a reaction secreting fumarate is added.
70	226	The pentose phosphate pathway operates solely for nucleotide biosynthesis with the reaction fluxes through ribulose phosphate 3-epimerase, transketolase I, transketolase II, and transaldolase B all operating in reverse. Fluxes through glucose-6-phosphate dehydrogenase, lactonase, and 6-phosphogluconate dehydrogenase are absent in this case, replaced by pyridine nucleotide transhydrogenase which meets the cellular NADPH needs. In addition, formate is now secreted along with acetate, fumarate, and carbon dioxide.
60, 50, 40	225	Acetate is no longer secreted as a metabolic byproduct, but is converted to acetyl-CoA by acetyl-CoA synthetase.
30, 20, 10, 1	224	Three glycolytic reactions, phosphoglycerate mutase, enolase, and pyruvate kinase are eliminated, but both serine deaminase and phosphoenolpyruvate synthase are added to supply the cell with phosphoenolpyruvate.

tational confirmation of the astounding redundancy and flux redirection versatility of the *E. coli* network. More importantly, all minimal reactions sets identified include 11 of 12 reactions whose corresponding gene deletions were determined experimentally to be lethal for growth on glucose. Earlier analyses (*8*) based on single gene deletions conducted with this model using linear optimization identified only 7 of 12 lethal gene deletions, motivating the importance of considering simultaneous gene deletions within an MILP framework.

In the second case study, the uptake or secretion of any organic metabolite is enabled. The amount of organic material entering the network is kept consistent with the first case study by allowing the uptake of a constrained amount of carbon atoms (<60 mmol/gDW·h). Unconstrained uptake routes for oxygen, inorganic phosphate, sulfate, and ammonia are also provided as in the first study. Under these "ideal" uptake conditions, we find that an MGR of 1.341 g biomass/gDW·h is attainable, requiring at least 201 metabolic reactions. The fact that only five amino acids are imported under maximum growth (i.e., MGR) conditions indicates that it is stoichiometrically more favorable to produce most amino acids internally rather than transport them into the cell from the medium.

This trend, however, is quickly reversed as the growth rate requirement is reduced. This reversal yields a corresponding sharp decrease in the total number of required reactions as a direct result of the importation of an increasing number of metabolites at submaximum target growth demands. Table 2 lists the metabolites uptaken or secreted at each target growth rate, while Figure 1B (100−90% of MGR) and Figure 1C (100−1% of MGR) illustrate the number of required metabolic reactions needed to attain various target growth demands. The rapid reduction in size of the minimal reaction sets by importing an increasing number of metabolites as the biomass demands are lessened (see Table 2) continues until the growth demands are reduced to about 90% from the MGR. Below this growth target (see Figure 1C) additional but modest reductions are achieved primarily through flux redirections. Table 3 summarizes the reactions that are being removed or added to the minimal reaction set as the growth target is successively lowered. The smallest minimal reaction network for the second case study, comprising 122 reactions, is reached when the target growth demands are lowered to 10% of the MGR. This minimal network comprises mostly cell envelope and membrane lipid biosynthetic reactions, along with a number of transport and salvage pathway reactions, as shown in Table 4. As in the glucose-only study, multiple minimal reaction sets for multi-organic uptake case are expected.

By solving a related problem (see Appendix) we find that only 91 reactions are required to provide sufficient network connectivity between the available external metabolites and constituents of biomass for the multiple organic uptake study. However, because this minimal set is based strictly on network connectivity, it inherently neglects the specific stoichiometry of each reaction, thus underestimating the minimal reaction set size. Available upon request are (i) the flux distributions associated with the minimal reaction sets under different growth targets and the corresponding genes coding for the catalyzing enzymes and (ii) a partial list comprising 15 distinct alternate minimal reaction sets for the glucose-only uptake case study.

Discussion

In this study, we have identified the minimum number of *E. coli* metabolic reactions capable of supporting growth under two different uptake environments: (i) a glucose- or acetate-only uptake environment and (ii) free uptake or secretion of any organic metabolite involving a corresponding transport reaction. The obtained results quantitatively demonstrate that minimal reaction sets and thus corresponding minimal gene sets are strongly dependent on the uptake opportunities afforded by the growth medium. While an *E. coli* cell grown on a medium containing only glucose or acetate requires at least 224 or 229 metabolic reactions, respectively, to support growth, a cell cultured on a rich optimally engineered medium could theoretically support growth with as few as 122 metabolic reactions. In addition, the choice of the single substrate affects the minimal reaction set size and composition. As expected, the minimal reaction set becomes larger by increasing the required growth rate. However, the magnitude of this increase is quite different for the examined cases. While in case (i) the minimal reaction set increases only from 224 to 234 to meet the maximum growth rate on glucose and from 229 to 231 for acetate growth, in case (ii) the minimal reaction set almost doubles, going from 122 to 201. Another significant observation is the large redundancy of the *E. coli*

Table 2. Metabolites Uptaken or Secreted at Each Target Growth Rate on an Optimally Engineered Medium[a]

	percentage of 100% biomass generation required												
metabolite	100	99.5	99	98	97	96	95	90	85	80	70	60	10
acetate												**S**	**S**
acetaldehyde													U
adenine				U	U	U	U	U	U			U	
adenosine										U	U		U
alanine										U	U		
arginine	U	U	U	U	U	U	U	U	U	U	U	U	U
asparagine									U	U		U	U
aspartate									U	U	U	U	U
carbon dioxide	**S**	**S**	**S**	**S**	**S**	**S**	**S**	**S**	**S**	**S**	**S**	**S**	**S**
cysteine	U	U	U	U	U	U	U	U	U	U	U	U	U
D-alanine								U	U			U	U
thymidine		U	U	U	U	U	U	U	U	U	U	U	U
ethanol	U	U	U	U	U	U	U	U		U		U	
glycerol											U		
glycerol-3-phosphate	U	U	U	U	U	U	U	U	U	U		U	U
glutamine									U	U	U	U	U
glutamate											**S**	U	U
glycine						U	U	U	U	U	U	U	U
guanine				U	U	U	U		U	U			
guanosine								U			U	U	U
histidine		U	U	U	U	U	U	U	U	U	U	U	U
isoleucine	U	U	U	U	U	U	U	U	U	U	U	U	U
leucine							U	U	U	U	U	U	U
lysine	U	U	U	U	U	U	U	U	U	U	U	U	U
meso-diaminopimelate		U	U	U	U	U	U	U	U	U	U	U	U
methionine	U	U	U	U	U	U	U	U	U	U	U	U	U
mannitol												U	U
ammonia	U	U	U	U	U	U	U	U					
oxygen	U	U	U	U	U	U	U	U	U	U	U	U	U
phenylalanine			U	U	U	U	U	U	U	U	U	U	U
phosphate	U	U	U	U	U	U	U	U	U	U	U	U	U
proline					U	U	U	U	U	U	U	U	U
putrescine	U	U	U	U	U	U	U	U	U	U	U	U	U
pyruvate										U	U	U	U
ribose												U	U
serine								U	U	U	U	U	U
spermidine	U	U	U	U	U	U	U	U	U	U	U	U	U
threonine		U	U	U	U	U	U	U	U	U	U	U	U
tryptophan		U	U	U	U	U	U	U	U	U	U	U	U
tyrosine			U	U	U	U	U	U	U	U	U	U	U
uracil						U	U	U	U	U		U	
uridine											U	U	U
valine							U	U	U	U	U		U
no. of metabolites uptaken	**12**	**17**	**19**	**21**	**22**	**24**	**26**	**28**	**29**	**31**	**29**	**34**	**34**

[a] U denotes metabolite uptaken; S denotes metabolite secreted.

metabolic network, which is capable of supporting growth utilizing only 31% of the available metabolic reactions for growth on glucose and only 17% of the available reactions for growth on a complex medium. Even these reduced minimal reaction network sets exhibit large multiplicities. Specifically, a non-exhaustive list of 100 alternative minimal reaction sets were identified for the glucose-only uptake case.

It must be noted that our analysis provides a species-specific minimal *metabolic* reaction set, which is a subset of the complete *E. coli* minimal genome. This is a consequence of the adopted reaction-based analysis, which cannot account for genes associated with translation, replication, recombination, repair, transcription, and genes of unknown function. A comparison of our minimal metabolic reaction set with the essential gene set of Hutchison et al. (*2*) and the minimal gene set proposed by Mushegian and Koonin (*5*) in their studies with *Mycoplasma genitalium* is provided in Table 5. The obtained results agree conceptually with the finding of Hutchison and co-workers (*2*) that limited metabolic capacity can be compensated for by a proportionately greater dependence on the importation of nucleosides, amino acids, and other metabolites. Although a complete genome-based reconstruction of the *M. genitalium* metabolic network is currently unavailable, preventing a reaction-by-reaction comparison, the distributions of metabolic genes/reactions among the various functional classifications in the three studies are quite similar. Thus, perhaps the simultaneous reaction removal strategy applied to *E. coli* in this work parallels the evolutionary pressures placed on *M. genitalium* to reduce its genome size. The minimal reaction set size overestimation in our analysis may be largely due to its species-specific nature. Whereas the cellular envelope of *E. coli* contains a cell wall made up largely of peptidoglycan, the cellular envelope of mycoplasmas lacks a cell wall. Thus many of the cellular envelope reactions necessary for *E. coli* survival are not included in the genes sets of Hutchison et al. (*2*) and Mushegian and Koonin (*5*). Another contributing factor is that we assign a different reaction/gene to the uptake or secretion of each metabolite, although similar metabolites can be transported by mechanisms associated with a single gene. Furthermore, since our analysis is based on the *E. coli* model, more efficient reaction combinations, perhaps occurring in non-*E. coli* species, could further reduce the minimal gene set lowering the discrepancy.

This framework can be utilized to construct minimal reaction sets for additional species. By contrasting these minimal sets it could be inferred how minimal reaction sets (metabolic gene sets) compare along different evolutionary branches. Specifically, minimal reaction sets for *M. genitalium* and *H. influenza* could be determined

Table 3. Evolution of Minimal Reaction Sets for Case (ii) under Decreasing Growth Requirements

target % MGR	min reaction set (no. of reactions	key features
100	201	The organic material transported into the cell includes ethanol and glycerol-3-phosphate, which fuel glycolysis, the TCA cycle, and PPP. The flux directions of the glycolysis pathway are split with all reaction fluxes preceding glyceraldehyde-3-phosphate (G3P) dehydrogenase operating in reverse, and all fluxes following and including G3P dehydrogenase operate in their forward directions. Putrescine, spermidine, and five amino acids are transported into the network eliminating the need for bio-synthetic pathways for these components.
90	132	While the PPP and TCA cycle reactions are still functional, the network no longer utilizes the five glycolytic reactions from glyceraldehyde-3-phosphate dehydrogenase to pyruvate kinase. Consequently, the TCA cycle is completely fueled by imported ethanol and acetate rather than flux from the glycolysis pathway.
80	125	This network tolerates the complete elimination of the TCA cycle and glyoxylate shunt. As a result, the function of the pentose phosphate pathway reactions is no longer restricted to nucleotide biosynthesis, but now includes the formation of cellular NADPH. Most of this NADPH is subsequently converted to NADH by pyridine nucleotide transhydrogenase to replace the cellular reducing power lost from the inactivity of the TCA cycle.
70	124	A slightly less efficient set of internal metabolic reactions enables the growth demands to be met with the importation of one less metabolite (i.e. one less transport reaction) than its 80% counterpart.
60, 50, 40, 30, 20	123	Neither the TCA cycle nor PPP are utilized for reducing power. Most of the cellular reducing capabilities are now generated from the uptake of ethanol and its subsequent conversion into acetyl-CoA.
10, 1	122	This minimal network is comprised mostly of cell envelope and membrane lipid biosynthetic reactions, along with a number of transport and salvage pathway reactions. Here, the three core metabolic routes, glycolysis, the TCA cycle, and the pentose phosphate pathway are almost completely dismantled with only one glycolytic and four PPP reactions remaining.

Table 4. Functional Classification of Minimal Network Reactions for Growth on an Optimally Engineered Medium

functional classification	no. of reactions
ALA isomerization	1
alternative carbon source	7
anaplerotic reactions	1
cell envelope biosynthesis	29
EMP pathway	5
membrane lipid biosynthesis	16
pentose phosphate pathway	4
pyrimidine biosynthesis	1
respiration	5
salvage pathways	17
transport	36
	122

Table 5. Comparison of Minimal Metabolic Gene/Reaction Sets Based on Functional Classification[a]

metabolic function	essential gene set[b] ref 2; no. of genes	minimal gene set ref 5; no. of genes	minimal reaction set no. of reactions
amino acid biosynthesis	0	0	1
biosynthesis of cofactors, prosthetic groups, and carriers	4	3	0
cell envelope	2	11	29
central intermediary metabolism	7	7	1
energy metabolism	31	32	21
fatty acid and phospholipid metabolism	5	7	16
purines, pyrimidines, nucleosides, and nucleotides	17	14	18
transport and binding proteins	17	25	36
	83	**99**	**122**

[a] Gene functions based on the current TIGR (20) version of the *M. genitalium* annotation. [b] Obtained from disrupting the expression of individual genes and determining which are essential for cellular survival.

and benchmarked with earlier studies (*5*). Additionally, a species-independent minimal metabolic reaction set can be pursued by lumping reactions occurring in many different species (*12*−*14*) within a Universal stoichiometric matrix (*15, 16*). As more elaborate models are developed describing elementary functions of minimal cells, such as the work of Browning and Shuler (*17*) for the initiation of DNA replication, more detail can be added to the model. Apart from utilizing this MILP framework for rationally identifying "minimal" metabolic networks, it can also be used to predict in silico lethal gene deletions for different organisms and uptake environments. By identifying lethal gene deletions for pathogenic microbes (e.g., *H. pylori*), a ranked list of promising targets for therapeutic intervention (i.e., interruption of gene expression) can be compiled. Even though the proposed computational procedure is dependent upon the assumptions of the adopted FBA model (*8*), it affords the versatility to study different uptake/secretion environments as well as encompass reaction sets from multiple species in the search for the minimal genome.

Appendix: Modeling and Computational Protocol

Flux balance analysis relies on the stoichiometry of biochemical pathways and cellular composition information to identify the flux distributions potentially available to the cell. For a metabolic network comprising N metabolites and M metabolic reactions we have

$$\frac{dX_i}{dt} = \sum_{j=1}^{M} S_{ij} v_j, \quad i = 1, ..., N \tag{1}$$

where X_i is the concentration of metabolite i, S_{ij} is the stoichiometric coefficient of metabolite i in reaction j, and v_j represents the flux of reaction j. Typically, the resulting system of equations is underdetermined (the number of reactions exceeds the number of metabolites). The maximization of growth rate is sometimes employed (*18*) as a surrogate for cell fitness. The key assumption is that the cell is capable of spanning all flux combinations allowable by the stoichiometric constraints and thus achieving any flux distributions that maximize a given metabolic objective. This may overestimate the region of accessible fluxes

by neglecting kinetic and/or regulatory constraints. The optimization model (linear programming) for maximizing biomass production or, equivalently, growth rate (assuming a 1 gDW·h basis) is

$$\text{Maximize } Z = v_{\text{biomass}} \tag{2}$$

$$\text{subject to } \sum_{j=1}^{M} S_{ij} v_j = b_i, \quad i = 1, ..., N$$

$$v_j \in \mathscr{R}^+, \quad j = 1, ..., M$$

$$b_i \in \mathscr{R}, \quad i = 1, ..., N$$

where v_{biomass} is the corresponding reaction flux comprising all necessary components of biomass in their respective ratios (*9*). One gram of biomass is produced per unit flux of v_{biomass}. Variable b_i quantifies the uptake (negative sign) or secretion (positive sign) of metabolite i. In case (i), only ammonia, glucose, oxygen, phosphate, and sulfate are allowed to have a negative value for b_i and any metabolite with a transport reaction out of the cell can be secreted, whereas in case (ii) all organic metabolites can be imported. In this study we explore what is the minimum number of metabolic reactions capable of maintaining maximum and submaximal levels of biomass production. By mapping reactions to their corresponding genes, a connection between biomass production and gene expression is established. The presence/absence of reactions, and therefore genes, is described mathematically by incorporating logic 0−1 variables into the flux balance analysis framework. These binary variables

$$y_j = \begin{cases} 1 & \text{if reaction flux } v_j \text{ is active} \\ 0 & \text{if reaction flux } v_j \text{ is not active} \end{cases}, \quad j = 1, ..., M \tag{3}$$

assume a value of 1 if reaction j is active and a value of 0 if it is inactive. The following constraint

$$\nu_j^{\min} \cdot y_j \le v_j \le \nu_j^{\max} \cdot y_j, \quad j = 1, ..., M \tag{4}$$

ensures that reaction flux v_j is set to 0 when no gene coding for the enzyme catalyzing reaction j is present and functional. Alternatively, when such a gene is active, v_j is free to take values between a lower bound $\nu_j^{\min}$ and an upper bound $\nu_j^{\max}$. The mixed-integer linear programming problem of minimizing the total number of functional reactions in the network capable of meeting a target for biomass production $v_{\text{biomass}}^{\text{target}}$ is as follows:

$$\text{Minimize } Z = \sum_{j=1}^{M} y_j \tag{5}$$

$$\text{subject to } \sum_{j=1}^{M} S_{ij} v_j = b_i, \quad i = 1, ..., N$$

$$v_{\text{biomass}} \ge v_{\text{biomass}}^{\text{target}}$$

$$\nu_j^{\min} \cdot y_j \le v_j \le \nu_j^{\max} \cdot y_j, \quad j = 1, ..., M$$

$$y_j \in \{0,1\}, \quad j = 1, ..., M$$

$$v_j \in \mathscr{R}^+, \quad j = 1, ..., M$$

$$b_i \in \mathscr{R}, \quad i = 1, ..., N$$

The above MILP belongs to the class of generalized network problems (*19*). Here each metabolite constitutes a node and each reaction represents an arc in the network.

The presence of over 1000 binary variables causes the problem to become computationally intractable for some instances. In particular, the computational burden increases for lower biomass targets, and it is much greater for case (ii) than case (i) as a result of the added complexity associated with multiple uptakes. To alleviate the computational burden, four preprocessing techniques are employed: (i) isoenzyme grouping, (ii) futile cycle exclusion, (iii) flux bounds generation, and (iv) connectivity constraint addition. Isoenzyme grouping refers to the aggregation of the 171 reactions catalyzed by isoenzymes. Reactions differing only in the catalyzing enzyme (i.e., isoenzymes) are grouped together, treating all isoenzymes as a single reaction. This reduces complexity by pruning the total number of binary variables. Futile cycle exclusion addresses the removal of sets of reactions (two or more) that collectively recycle fluxes in a loop without any net effect on metabolism. A special case is reversible reactions with non-zero fluxes for both directions. In general, a set $\mathscr{K}$ composed of K reactions forms a futile cycle if

$$\sum_{j \in \kappa} S_{ij} = 0, \quad i = 1, ..., N \tag{6}$$

The following constraint ensures that at least one of them will be inactive, breaking the cycle:

$$\sum_{j \in \kappa} y_j \le K - 1 \tag{7}$$

Overall, 346 futile cycles were identified and eliminated from the model. Most of the futile cycles involved simply reversible reactions.

The solution time of the resulting MILP problems is highly dependent on the tightness of the imposed lower $\nu_j^{\min}$ and upper $\nu_j^{\max}$ bounds on the fluxes v_j. Tight bounds $\nu_j^{\min}$ and $\nu_j^{\max}$ are obtained by minimizing and maximizing, respectively, every single reaction flux v_j subject to the flux balance constraints and the biomass target specification.

$$\text{Maximize/Minimize } v_{j^*} \tag{8}$$

$$\text{subject to } \sum_{j=1}^{M} S_{ij} v_j = b_i, \quad i = 1, ..., N$$

$$v_{\text{biomass}} \ge v_{\text{biomass}}^{\text{target}}$$

$$v_j \in \mathscr{R}^+, \quad j = 1, ..., M$$

$$b_i \in \mathscr{R}, \quad i = 1, ..., N$$

This is a linear programming (LP) problem (no binary variables) and is quickly solved (i.e., less than a few seconds) for all cases. Note that different bounds are generated for different biomass targets, and the higher the biomass target is, the tighter the obtained bounds are.

Connectivity constraints are also added to ensure that if a reaction producing an intracellular metabolite is active, then at least one reaction consuming this metabolite must be active and vice versa. In addition, if a reaction transporting an extracellular metabolite into the cell is active, then at least one intracellular reaction

consuming this metabolite must be active and vice versa. These relations are incorporated in the model as follows after partitioning the reaction set $\mathcal{J}$ into two subsets: $\mathcal{J}_{\text{int}}$ representing intracellular reactions and $\mathcal{J}_{\text{trans}}$ representing reactions transporting metabolites to and from the cell. The metabolite set I is also partitioned into two subsets with I_{int} and I_{ext} representing intracellular and extracellular metabolites, respectively.

$$y_{j'} \le \sum_{\substack{S_{ij}<0 \\ j\in J}} y_j \;\forall i \in I_{\text{int}}, \quad \forall\, j' \in \{j|S_{ij} > 0\} \tag{9}$$

$$y_{j''} \le \sum_{\substack{S_{ij}>0 \\ j\in J}} y_j \;\forall i \in I_{\text{int}}, \quad \forall\, j'' \in \{j|S_{ij} < 0\} \tag{10}$$

$$y_{j}' \le \sum_{\substack{S_{ij}<0 \\ j\in \mathcal{J}_{\text{trans}}}} y_j \;\forall i \in I_{\text{ext}}, \quad \forall\, j' \in \{j|S_{ij} > 0\} \tag{11}$$

$$y_{j''} \le \sum_{\substack{S_{ij}>0 \\ j\in \mathcal{J}_{\text{trans}}}} y_j \;\forall i \in I_{\text{ext}}, \quad \forall\, j'' \in \{j|S_{ij} < 0\} \tag{12}$$

These connectivity constraints are also employed to identify the smallest set of reactions capable of ensuring adequate connectivity between the external metabolites and the components of biomass. This problem involves minimizing $_j\Sigma\, y_j$ subject to constraints (9−12) with an active biomass reaction, $y_{\text{biomass}} = 1$. An algorithm for solving two similar network connectivity problems is presented by Romero and Karp (*21*) and applied to the EcoCyc (*13*) database.

The iterative generation of the multiple minimal reaction sets is achieved by accumulating integer cuts and resolving the MILP formulation. Each integer cut excludes one previously found solution. For example, solution y_{j*} is excluded from consideration by adding the following integer cut:

$$\sum_{j:y_{j*}=1} y_j + \sum_{j:y_{j*}=0} (1 - y_j) \le M - 1 \tag{13}$$

All optimization problems are solved using CPLEX 6.5 (*10*) accessed through the modeling environment GAMS (*11*) on an IBM RS6000-270 workstation. The total cumulative CPU expended for this study was in the order of 400 h.

Acknowledgment

Financial support by NSF Awards CTS9701771 and BES0120277 is gratefully acknowledged.

References and Notes

(1) Cho, M. K.; Magnus, D.; Caplan, A. L.; McGee, D. et al. Ethical considerations in synthesizing a minimal genome. *Science* **1999**, *286*, 2087−2090.

(2) Hutchison, C. A., III et al. Global transposon mutagenesis and a minimal mycoplasma genome. *Science* **1999**, *286*, 2165−2169.

(3) Goebl, M. G.; Petes, T. D. Most of the yeast genomic sequences are not essential for cell growth and division. *Cell* **1986**, *46*, 983.

(4) Itaya, M. An estimation of minimal genome size required for life. *FEBS Lett.* **1995**, *362*, 257.

(5) Mushegian, A. R.; Koonin, E. V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 1026.

(6) Arigoi, F. et al. A genome-based approach for the identification of essential bacterial genes. *Nat. Biotechnol.* **1998**, *16*, 851−856.

(7) Huynen, M. Constructing a minimal genome. *Trends Genet.* **2000**, *16*, 116.

(8) Edwards J. S.; Palsson, B. O. The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5528−5533.

(9) Neidhardt, F. C. *Escherichia coli and Salmonella: Cellular and Molecular Biology*; ASM Press: Washington, DC, 1996.

(10) Brooke, A.; Kendrick, D.; Meeraus, A.; Raman, R. *GAMS: The Solver Manuals*; GAMS Development Corporation Washington, D. C., 1998.

(11) Brooke, A.; Kendrick, D.; Meeraus, A.; Raman, R. *GAMS: A User's Guide*; GAMS Development Corporation: Washington, DC, 1998.

(12) Kanehisa, M.; Goto, S. KEGG: The Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 29−34.

(13) Karp, P. D.; Riley, M.; Saier, M.; Paulsen, I. T.; Paley, S. M.; Pellegrini-Toole, A. The EcoCyc and MetaCyc databases. *Nucleic Acids Res.* **2000**, *28*, 56−59.

(14) Maranas, C. D.; Burgard, A. P. Review of EcoCyc and MetaCyc databases. *Metab. Eng.* **2001**, *3*, 98−99.

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

(16) Burgard, A. P.; Maranas, C. D. Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. *Biotechnol. Bioeng.* **2001**, *74*, 364−375.

(17) Browning, S. T.; Shuler, M. L. The Initiation of DNA Replication in a Mathematical Model of a Minimal Cell. Presented at AIChE Annual Meeting 2000, Session 69, Los Angeles, CA.

(18) Varma, A.; Palsson, B. O. Metabolic flux balancing: Basic concepts, scientific and practical use. *Biotechnol. Bioeng.* **1998**, *12*, 994−998.

(19) Ahuja, R. K.; Magnanti, T. L.; Orlin, J. B. *Network Flows, Theory, Algorithms, and Applications*; Prentice Hall: Englewood Cliffs, NJ, 1993.

(20) TIGR Web Site. TIGR microbial database. http://www.tigr.org (accessed 2001).

(21) Romero, P.; Karp, P. D. Nutrient-related analysis of pathway/genome databases. *Proceedings of the Pacific Symposium on Biocomputing*; Altman, R., Klein, T., Eds.; World Scientific: Singapore, 2001; pp 471−482.

Accepted for publication July 16, 2001.

BP0100880