Minireview

Extending knowledge of *Escherichia coli* metabolism by modeling and experiment

Eberhard O Voit* and Monica Riley†

Addresses: *Medical University of South Carolina, Charleston, SC 29425, USA. †Marine Biological Laboratory, Woods Hole, MA 02543, USA.

Correspondence: Monica Riley. E-mail: mriley@mbl.edu

Published: 28 October 2003

Genome ***Biology*** 2003, **4:**235

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/11/235

Abstract

One of the challenges for 'post-genomic' biology is the integration of data from many different sources. Two recent studies independently take steps towards this goal for *Escherichia coli*, using mathematical modeling and a combination of gene expression and protein levels to predict new gene functions and metabolic behaviors.

It has become a platitude of the post-genomic era that a deluge of data is being produced and awaits both computational/mathematical analysis and experimental verification. There is no reason to argue with this observation, yet it leads to two immediate follow-up questions. First, has the genomic era come to an end already? And second, what types of mathematical and computational models would be most beneficial for dealing with the rich streams of data? Two recent articles, by Reed *et al.* [1] and Corbin *et al.* [2], answer the first question in the negative: there is still a lot of genomic research to be done. These two articles show that, even for one of the best studied of organisms, *Escherichia coli*, there are still very many genes for which we know little beyond their sequence and location. We don't know what their functions are, exactly which genes are actively functioning at any one time, and to what degree some might even be entirely dispensable. The experimental approach proposed by Corbin *et al.* [2] sheds light on some of these issues with a combination of two methods, one for measuring gene expression and one for detecting proteins in *E. coli* cells. In contrast, Reed *et al.* [1] address the two questions with a novel and interesting application of mathematical modeling.

The combination of the two papers is intriguing, because both have the same purpose - annotating gene function and learning more about intermediary metabolism - yet the two use very different approaches to accomplish their common goal. This independence of approaches may be useful for comparisons or for mutual complementation of results, and could aid the community in answering questions about the reliability of separate approaches to interpreting genomic information.

To assign metabolic functions to unknown genes, Reed *et al.* [1] use a method that is based on a combination of mathematical modeling and data mining. The authors use the available literature and database information to construct a large stoichiometric model of intermediary metabolism that includes all known biochemical reactions in *E. coli*. A stoichiometric model describes quantitatively the flow of mass through a metabolic network. It includes one linear differential equation for each metabolite, and each of these equations consists of the sum of all fluxes leading to the production of this metabolite minus the sum of all fluxes degrading or consuming this metabolite. To determine the sizes of all internal fluxes, one measures some input and output fluxes, such as substrate uptake and lactate or carbon dioxide excretion. Under the assumption that all reactions are in a dynamic steady state, the fluxes at each metabolite should be numerically balanced. Typically there is not enough input-output information to compute all internal fluxes, but the stoichiometry severely constrains the range of possibilities, and optimization within this range leads to the desired internal flux distribution (reviewed in [3,4]).

Although the assumption of flux balance is found to be true in the majority of cases, the authors detect notable exceptions and conclude that some catalytic steps must be missing from the model structure [1]. Analysis of metabolic maps in other organisms suggests mechanisms (enzymes and catalyzed reactions) associated with the depletion of those metabolites that accumulate in the current model, or the production of metabolites that are not made available in sufficient quantities in the model. In many cases, these mechanisms have been characterized in other organisms, and often their genes and gene sequences have been determined. This information is used to search for similar sequences among unknown *E. coli* genes and thus leads to proposals for new annotations for formerly unidentified open reading frames (ORFs). Thus, through the integration of metabolic data by means of a mathematical model, inconsistencies in the model lead to new discoveries, or at least to suggestions for targeted experiments that would confirm or reject the hypothesized annotation.

The metabolites reported by the current model as accumulating without removal are called 'dead-end' metabolites [1]. The list is interesting from a biochemical point of view, because it consists of a mixture of types of compounds. Some of the metabolites are common, essential compounds whose balances must be managed by the cell, such as thymine and siroheme. The apparent accumulation of these compounds by the model may point to incomplete biochemical data. Data of this sort should be useful for finding omissions in the model and for the annotation of genes. Some of the other metabolites on the list are approaching the macromolecule category. All reactions associated with these metabolites would, therefore, not be expected to be present in any model of small-molecule metabolism. Examples of these are 'cold-adapted KDO2 lipid A' and a 'peptidoglycan subunit'. Other entries in the dead-end list may be eccentric names for normal metabolites: for example, T-trans-aconitate instead of the usual trans-aconitate, and D-D-Methionine instead of either D-methionine or L-methionine.

To assign functions to particular genes, possible connections between 'missing functions' (enzymes and their reactions) and particular *E. coli* gene sequences are deduced by sequence similarity searches. This data-mining and annotation step apparently used older information about *E. coli* gene products and for this reason, one finds many of the predictions cogent because they exist in the current databases. The authors list putative genes for nearly 30 functions (see the Additional data files of [1]). We compared the predictions with information in a current database [5] and found that many of the predictions verify the approach taken because they are essentially the same as the currently 'known' or 'putative' assignments. A few of the predictions seem unlikely given the functions of sequence-similar proteins. Some of the predictions, however, are indeed new connections to uncharacterized genes that could now serve to motivate experimental verification. On the whole, the approach to annotation through metabolic circuitry seems to have the capability in the future of expanding metabolic and/or genetic knowledge and directing the experimental verification of new functions.

Entirely different approaches are used by Corbin *et al.* [2] to characterize both protein and mRNA populations in *E. coli* cultures. Proteins extracted from growing cells are visualized using high-pressure liquid chromatography combined with tandem mass spectrometry (HPLC-MS/MS). Over 1,100 ORFs were detected, corresponding to a quarter of all possible gene products. It is not known what fraction of *E. coli* proteins is present in detectable amounts under the growth conditions used, but one can safely expect that not all genes are expressed at any one time; to detect more than a quarter of all gene products is therefore an impressive feat. Comparing these protein results with mRNA levels, measured separately by hybridization to an Affymetrix chip, the authors find a good correlation between the two types of measurements, provided that the intensity of the mRNA signal does not fall within the lowest 5% of the measured range. For lower intensities, the correspondence to detected proteins is no longer significant, an observation that might be attributable to the fact that reliable detection of the proteins by HPLC-MS/MS requires relatively high levels of expression [2].

The two-pronged approach of assessing proteins that are directly involved in metabolic function versus mRNAs that are only involved indirectly raises the question of whether we can actually learn anything from the mRNA results that we did not already know. The answer is that there is indeed added value. Identification of the collection of expressed genes in *E. coli* [2] allows us to ask whether the list corresponds to our *a priori* expectations of which catabolic, anabolic and macromolecular-synthesis proteins are made under the specific growth conditions. Using a relatively permissive threshold one finds that about 27% of the 955 known metabolic enzymes in the mRNA experiments are not expressed during growth on glycerol as a carbon source. This is not a surprise, because we know that many enzymes are made only in response to particular growth conditions. (Note that the computational model of Reed *et al.* [1] uses a collection of 927 entities, of which 733 are enzymes and the rest are transporters. All are presumed to be active members of the metabolic network in the computational model [1], although the experimental data of Corbin *et al.* [2] suggest that many are repressed).

The data from the mRNA experiments reveal both previously known and unknown unexpressed genes. The known interruption of the *gatR* repressor gene in *E. coli* strain MG1655 is confirmed as the galactitol genes are derepressed as expected [2,6]. The mRNA levels of the sorbitol and mannitol degradation enzymes are derepressed [2], suggesting that the GatR repressor is involved in the regulation of these other sugar alcohols as well. Another mutation, in the *pyrE* gene, was not known to the authors, but it was flagged by a derepressed mRNA signal [2]. The defect in *pyrE* was later confirmed by growth-rate studies [5]. Thus, unknown genes are readily detected experimentally.

Rich information may be extracted from the data on which isozymes are metabolically active in the selected medium. Although Corbin *et al.* [2] did not present detailed functional analysis of gene expression, data of this kind can be extracted. For instance, inspection of the data shows that shikimate kinase I is expressed more than shikimate kinase II, two of the five peptidylprolyl isomerases are most highly expressed, and three of the four FK506/rapamycin-binding protein-type peptidylprolyl isomerases are most highly expressed. Two 3-oxo(acyl carrier protein)synthases, I and III, are well expressed, but very little of isozyme II is present. Numerous other insights of this type can be deduced directly from the data.

An observation brought out by Corbin *et al.* [2] is that genes in operons are not always coordinately expressed, because in many cases only some, not all, members of an operon were detected as present at the protein level (for a similar observation see [7]). This leads to the following deduction. As enzyme activity is not a direct function of the amount of mRNA present - because mRNA half lives can differ, translation efficiencies can differ, and specific activities of enzymes range widely - the amount of active mRNA may be regulated so as to produce similar enzyme activities. If true, this conclusion from the data of Corbin *et al.* [2] opens avenues of potentially fruitful investigation.

Another direct value of the experimental results is, of course, that the identification and quantification of proteins and mRNAs in the cell under particular growth conditions provide valuable *in vivo* input for computational pathway models. Furthermore, the results can be used for validation purposes, where comparisons are made between the list of proteins found experimentally and the pathways and fluxes included in a computational model.

Returning to the questions posed at the beginning of this article, one may ask whether the computational modeling procedure [1] falls into the realm of genomic or post-genomic research. This may sound like a purely academic question, but it leads us to ask what is needed next in terms of computational and mathematical analysis. Even though the model of Reed *et al.* [1] is integrative, one would probably assign the particular use presented here to the genomic era, because the model serves as a data collection and gene-identification tool. It helps classify data in a novel fashion by using metabolite anomalies to identify possible missing reactions and enzymes, and suggests novel connections between missing enzymes and their genes through sequence analysis. This is an intriguing role for modeling, and the approach constitutes a fine example of practical model utilization.

Is the model useful beyond this role? In the work of Reed *et al.* [1] and the related literature [3,4] it is claimed that a large stoichiometric model describes metabolism with sufficient reliability to make predictions of organismal responses under untested conditions and to serve as a basis for optimizing *E. coli* strains for particular tasks of biotechnological interest. Indeed, examples have been presented where such predictions were successful [8]. Nevertheless, it must be recognized that the mathematical structure of any purely stoichiometric model precludes a true inclusion of kinetic and regulatory features. Under novel conditions, the cell is likely to respond by calling up its regulatory-control mechanisms, but this cannot be modeled with stoichiometry alone, except that once all regulation is done, the metabolic network should again reside at a steady state, in which all metabolites are balanced.

The question then becomes whether a constrained linear optimization of a stoichiometric model would actually reach the same balanced state that the real cell would assume through its regulatory mechanisms. At this point, this question cannot be answered with any generality, except that there will almost certainly be cases where the linear prediction is correct but there will also be cases where that is not so. For instance, the cell may ‘decide’ to export unwanted metabolites, or it may resort to pathways that are minimized under normal conditions and used only under specific conditions. As an example, it would seem difficult to predict with a stoichiometric model alone that a yeast cell would respond to heat shock with an enormous production of trehalose, which exists only in traces under cooler conditions. Thus, there need to be additional phases of model development and analysis on the path towards understanding organismal function.

The most obvious extension beyond stoichiometric models is the construction of nonlinear models, which can account for regulatory features (see, for example, [9]). These clearly require much more input in terms of pathway information and kinetic and regulation data but will have an improved chance of adequately representing tested and untested organismal behaviors. Like stoichiometric models, however, nonlinear models will eventually also encounter the ‘curse of combinatorial explosion’. Once these models reach a certain size, it becomes an overwhelming task to implement them numerically, to test the reliability of explicit and implicit assumptions associated with the model set-up, and to interpret the results. For instance, if a model contains ten parameters with ten possible values each, and if each model analysis takes one second, an exhaustive evaluation would require 317 years of computation time. Obviously, clever coding, parallelization and other advancements will reduce this time, but it is nevertheless quite obvious that such an approach is bound to break down eventually.

What is needed in addition to these direct extensions of modeling and simulation is the discovery of general principles that govern the behavior of organisms and their responses to stimuli [10]. Such principles provide an objective rationale for a particular design and operation of a gene-regulatory, metabolic or physiological system and will ultimately allow us to dissect large systems into interacting functional modules. They will also give us confidence in predicting responses under novel conditions, optimizing strains, or ultimately designing new strains from scratch. Both extensions, toward nonlinearities and toward the exploration of design and operating principles, will require solid and detailed information on the components of biological systems. The two papers discussed here [1,2] provide some such information and are therefore important in that they help us, in independent ways, to make the current 'parts catalog' of *E. coli* more complete, precise and reflective of the contents of the cell in specified conditions.

References

1. Reed JL, Vo TD, Schilling CH, Palsson BO: **An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR).** *Genome Biol* 2003, **4:**R54.
2. Corbin RW, Paliy O, Yang F, Shabanowitz J, Platt M, Lyons CE Jr, Root K, McAuliffe J, Jordan MI, Kustu S, *et al.*: **Toward a protein profile of *Escherichia coli*: comparison to its transcription profile.** *Proc Natl Acad Sci USA* 2003, **100:**9232-9237.
3. Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BO: **Metabolic pathways in the post-genome era.** *Trends Biochem Sci* 2003, **28:**250-258.
4. Reed JL, Palsson BO: **Thirteen years of building constraint-based *in silico* models of *Escherichia coli*.** *J Bacteriol* 2003, **185:**2692-2699.
5. ***E. coli* genome and proteome database** [http://genprotec.mbl.edu]
6. Soupene E, van Heeswijk WC, Plumbridge J, Stewart V, Bertenthal D, Lee H, Prasad G, Paliy O, Charernnoppakul P, Kustu S: **Physiological studies of *Escherichia coli* strain MG1655: growth defects and apparent cross-regulation of gene expression.** *J Bacteriol* 2003, **185:**5611-5626.
7. Voit EO, Radivoyevitch T: **Biochemical systems analysis of genome-wide expression data.** *Bioinformatics* 2000, **16:**1023-1037.
8. Edwards JS, Ibarra RU, Palsson BO: ***In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data.** *Nat Biotechnol* 2001, **19:**125-130.
9. Voit EO: *Computational Analysis of Biochemical Systems. A Practical Guide for Biochemists and Molecular Biologists*. Cambridge, UK: Cambridge University Press, 2000.
10. Savageau MA: **Reconstructionist molecular biology.** *New Biol* 1991, **3:**190-197.