HHS Public Access

Author manuscript

Curr Opin Biotechnol. Author manuscript; available in PMC 2016 August 01.

Published in final edited form as:

Curr Opin Biotechnol. 2015 August ; 34: 105–109. doi:10.1016/j.copbio.2014.12.010.

Transparency in metabolic network reconstruction enables scalable biological discovery

Benjamin D. Heavner[1] and **Nathan D. Price**[1,*]

[1]Institute for Systems Biology, Seattle WA, USA

Abstract

Reconstructing metabolic pathways has long been a focus of active research. Now, draft models can be generated from genomic annotation and used to simulate metabolic fluxes of mass and energy at the whole-cell scale. This approach has led to an explosion in the number of functional metabolic network models. However, more models have not led to expanded coverage of metabolic reactions known to occur in the biosphere. Thus, there exists opportunity to reconsider the process of reconstruction and model derivation to better support the less-scalable investigative processes of biocuration and experimentation. Realizing this opportunity to improve our knowledge of metabolism requires developing new tools that make reconstructions more useful by highlighting metabolic network knowledge limitations to guide future research.

Introduction

Mapping metabolic pathways has been a focus of significant scientific efforts dating from the emergence of biochemistry as a distinct scientific field in the late 19th century [1]. This endeavor remains an important effort for at least two compelling reasons. First, cataloguing and characterizing the full range of metabolic processes across species (which because of genomics are being discovered at an incredible pace) is a fundamentally important step towards a complete understanding of our ecological environment. Second, mapping metabolic pathways in organisms – many of which can be found with specialized properties shaped by their environment – facilitates metabolic engineering to advance nascent industrial biotechnology efforts ranging from augmenting/replacing petroleum-derived chemical precursors or fuels to biopharmaceutical production [2]. However, despite laudable efforts to enable high-throughput “genomic enzymology” [3], the traditional biochemical approaches of enzyme expression, purification, and characterization remain time-, capital-, and labor-intensive, and have not expanded in scale like our ability to identify and characterize life genomically. Characterizing new metabolic function is further hampered by the challenge of cultivating environmental isolates in laboratory conditions [4]. Fortunately, recent efforts to leverage genome functional annotation and established knowledge of biochemistry have enabled the computational assembly of “draft metabolic reconstructions”

*Corresponding Author.

[5], which are parts lists of metabolic network components. In this context, a reconstruction is not just the information embodied in the stoichiometric matrix describing metabolic network structure, but also the associated metadata and annotation that entails an organism-specific knowledge base. Such a reconstruction can serve as the basis for making functional models amenable to mathematical simulation. Thus, a reconstruction is a bottom-up assembly of biochemical information, and a model can serve as a framework for integrating top-down information (for example, model constraints can be generated from statistically inferred gene regulatory networks [6]). Such computational approaches are significantly faster and less expensive than biochemical characterization [7]. They are also providing new resources facilitate cultivation of novel environmental isolates [8], and the scope of draft metabolic network coverage across the biome has increased much faster than wet lab characterization. If the distinction between reconstruction and model formulation can be strengthened and supported through software implementation, there is great opportunity for using both tasks to further advance rapid discovery of biological function.

The iterative process of manual curation of a draft metabolic network reconstruction to assemble a higher confidence compendium of organism-specific metabolism (a process termed “biocuration” [9,10]) remains time- and labor-intensive. Biocuration of metabolic reconstructions currently advances on a decadal time scale [11,12]. Thus, much research effort has focused instead on developing techniques for rapid development of models that are amenable to simulation [13,14]. Thousands of models have been derived from automatically assembled draft reconstructions [15], but most of these models consist of highly conserved portions of metabolism since they are propagated primarily via orthology. Though the number of models is large, they do not reflect the true diversity of cellular metabolic capabilities across different organisms [16]. Applying the rapid and scalable process of draft network reconstruction to support and accelerate the less-scalable processes of biocuration and *in vitro* or *in vivo* experimentation remains an unrealized opportunity. The path forward should focus on increased emphasis on transparently documenting the reconstruction process and developing tools to highlight, rather than obscure, knowledge limitations that ultimately cause limitations to model predictive accuracy.

More explicit annotation of metabolic network reconstruction and model derivation steps can help direct research efforts

The biocuration process of assembling biochemical knowledge from genomic annotation and published literature (i.e., assembling a reconstruction) involves identifying and resolving ambiguity inherent in information generated through ongoing experimental efforts to characterize biological systems. This bottom-up reconstruction process often introduces implicit hypotheses in the reconstruction. Such hypotheses, if made explicit, could be usefully exploited to prioritize experimental efforts. For example, reconstruction requires selection of a specific genome assembly, selection of a homology threshold for functional annotation, and interpretation of published literature regarding pathway information. Information about how these choices are made during the course of reconstruction can be very useful for informing subsequent research efforts, but such information is currently difficult to find because it is seldom included in published reconstructions. Furthermore, as

has been highlighted in a recent review [17], most software currently available for assembling a reconstruction does not support the detailed level of annotation that would be needed for scalable hypothesis generation.

Testing implicit hypotheses arising from reconstruction assembly provides one opportunity for guiding experimental efforts. However, the very act of identifying ambiguous information in the literature should also be exploited to contribute to experimental efforts, independent of the choices a researcher makes in assembling a reconstruction. Preliminary steps to facilitate large-scale computational identification of biological uncertainty have been made, such as the development of the Evidence Ontology [18]. However, realizing the potential for using reconstruction assembly to highlight experimental opportunities will require a broader shift to emphasize the limits of our knowledge, rather than only the predictive power of a model that can be derived from a reconstruction. Computational reconstruction of metabolic networks provides two distinct opportunities for guiding experimental efforts even before a mathematically computable model is derived from the assembled knowledge: highlighting areas of uncertainty in the current knowledge of an organism, and introducing hypotheses of metabolic function as choices are made throughout biocuration efforts.

The subsequent process of deriving a mathematically computable model from a reconstruction provides additional opportunities for scalable hypothesis generation that could be exploited to inform experimental efforts. While stoichiometrically constrained models derived from reconstructions are “parameter-light” when compared to dynamic enzyme kinetic models, they are not really “parameter free” [19]. As modelers derive a model from an assembled reconstruction, they must make choices. And, like the ambiguities and choices that are made and should be highlighted in assembling a reconstruction, highlighting the choices made in deriving a model provides further opportunity for scalable hypothesis generation. Examples of choices that often arise in deriving a functional model include adding intracellular transport reactions, filling network gaps, or trimming network dead ends to improve network connectivity [20]. Researchers seeking to conduct Flux Balance Analysis (FBA) [21] or similar approaches must formulate an objective function, can include testable parameters such as ATP maintenance requirements, and can compare model predictions to designated reference phenotype observations. Each of these model-building and tuning activities presents opportunities to rapidly develop and prioritize new hypotheses of metabolic function – for example, if a model required the addition of an intercompartmental transport reaction to function, this mathematical necessity would suggest opportunity to improve genomic annotation or to discover a previously uncharacterized enzyme or function.

Metabolic network gaps may be filled algorithmically by optimizing for shortest path connections [22] or through other penalties, such as prioritizing reactions catalyzed by enzymes that have higher homology for functional genomic annotations [23]. Either approach generates a list of reactions that permit network flux computationally, but are only predicted to have the modeled biological function. Similarly, identifying network dead ends presents opportunities to shine a light on the understudied portions of “dark metabolism”. A close look at areas where the metabolic network remains unconnected can lead to surprising

discoveries, such as the recent report of a riboneogenesis pathway in the central carbon pathway of yeast that could enable a flux bypass of the oxidative phase of the pentose phosphate pathway [24]. This function had not previously been documented despite the fact that glycolysis in yeast is perhaps the most extensively characterized pathway in the biosphere. It is thus extremely likely that many such surprises remain in even “well-studied” metabolic networks – and reconstruction biocuration and functional model development have great potential for facilitating such investigation.

To date, the process of selecting parameters such as gap-filling choice, biomass objective function definition, and constraint application during the model building process has been fairly implicit. Making the parameter selection process in model development more explicit provides a pathway to increase the speed of large-scale metabolic function discovery and characterization. Realizing the potential of reconstruction biocuration and model building for scalable hypothesis generation requires a traceable and reproducible method and software infrastructure for assembling a reconstruction, and for each step of deriving models from a reconstruction. Software should facilitate answering questions such as: What network gaps were filled, from what source, using what method? Why were those gaps filled? What dead ends were trimmed? How was the objective function formulated? What constraints were applied, and why? Is this set of constraints believed to be unique? How were reactions compartmentalized? What transport reactions were introduced? What changes were made for simulating different conditions?

Answering such a broad range of questions in a biocuration or subsequent model building effort is greatly facilitated by extensive annotation in a standardized data structure that can be used with a variety of software platforms. However, current standards for structured data formats that enable publishing and exchanging models (such as the Systems Biology Markup Language (SBML) [25] or the Minimal Information Required in the Annotation of Models (MIRIAM) standard [26] are designed for model exchange, rather than reconstruction exchange. It is likely that exchange and annotation of a reconstruction requires a database schema definition, rather than a markup language, which may be more suitable for exchange of functional models. This need is becoming more pressing as the scope of network reconstruction extends from metabolism to include Macromolecular Expression (ME) models [27]. One reconstruction schema that supports extensive annotation and provenance tracking has been defined as part of the Pathway Tools software [28], but this schema has not yet been adopted more widely or supported by other software implementations, nor has this schema yet been extended to support ME models or reconstructions. Similarly, there are also methods for deriving functional models from pathway databases such as those encoded by Pathway Tools [29], but these methods have not yet been standardized across the research community. Finding the necessary balance between standardization and the flexibility to enable research efforts and development of new methods remains an ongoing process of dialogue within the metabolic network research community.

Reconsidering model performance for scalable hypothesis generation

Model predictions of single gene deletion viability have been the key metric used for evaluating the performance of metabolic network models. However, over-emphasizing "improvements" to this metric carries risk of model implementation approaches that can hinder biological discovery. In contrast, careful consideration of differences between model prediction and *in vivo* observations provides another abundant opportunity for scalable hypothesis generation from metabolic network model building. It can be surprisingly difficult to generate a reference set of essential genes that can be used for evaluating model development. Even in genetic model organisms like *Saccharomyces cerevesiae*, recent work has highlighted that genes can be conditionally essential [30]; that the deletion of a single gene makes additional mutations more likely [31]; and that auxotrophic markers in laboratory strains meaningfully affect metabolic phenotype [32]. Thus, a model that has high predictive accuracy for one set of "essential" genes may in fact be over-fit, and have unexpectedly limited predictive ability of gene essentiality in non-reference environments. If recognized and highlighted, rather than obscured in an effort to improve apparent model prediction accuracy for each subsequent publication, such limitations can be usefully exploited for investigating which aspects of gene essentiality arise from stoichiometric constraints of the metabolic network itself, and which arise from other causes. Since "models based on pathway stoichiometry alone can only be used to ask questions that do not depend on more complex features" [19], model predictions that differ from observation may suggest biological functions arising from other factors.

Thus, models can be over-fit to improve performance by one metric at the expense of generality. There can be an inherent conflict between the goals of 1) assembling a reconstruction to reflect established knowledge (includes gaps, and highlights uncertainty) and 2) building a more focused model to be analyzed with a given performance metric. The same reconstruction can be used to derive models with high descriptive accuracy at the expense of predictive ability as well as models with better predictive ability but worse prediction of a reference observation. Use of condition-specific constraints makes the model more descriptive, but model developers must be conscious that applying such constraints move the scope of a model beyond stoichiometry, and can come at the expense of predictive power; such "models are not necessarily predictive but instead have a scoping nature by allowing us to assess what is metabolically feasible" [6].

Concluding remarks

The effort to computationally reconstruct biochemical knowledge to compile organism-specific reconstructions, and to derive computable models from these reconstructions, is a relatively young field of research with abundant opportunity for facilitating biological discovery of metabolic function. Judgment is required in assembling a reconstruction, and there should be careful consideration of the fact that judgment calls represent an implicit hypothesis. Making these hypotheses more explicit would help guide subsequent investigation. Bernhard Palsson and colleagues call for "an open discussion to define the minimal quality criteria for a genome scale reconstruction" [16] – an effort we fully support. We believe that such a beneficial "minimal quality criteria" should be guided by the goals of

reproducibility and transparency, including those aspects that can help to guide discovery of novel gene functions. A structured format and software tools that facilitate transparent reconstruction, specific model derivation, and a variety of model test metrics would facilitate biological knowledge generation, enable new research approaches, and improve model utility. To assess the quality of a reconstruction, it is critical to have detailed and accurate information on how it was assembled, such as knowing which choices were made in generating a simulatable model, and what methods were used to inform such choices.

Acknowledgments

The authors gratefully acknowledge Allison Kudla's assistance preparing the summary graphic and funding provided by the National Institute of Health Center for Systems Biology/2P50GM075647, National Science Foundation grant IOS-1256705, Department of Energy grant DE-AR0000426, and the Camille Dreyfus Teacher-Scholar Award.

References

1. Cornish-Bowden, A. New beer in an old bottle: Eduard Buchner and the growth of biochemical knowledge. Universitat de València; 1998.
2. Milne CB, Kim P-J, Eddy JA, Price ND. Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. Biotechnol J. 2009; 4:1653–1670. [PubMed: 19946878]
3*. Lukk T, Sakai A, Kalyanaraman C, Brown SD, Imker HJ, Song L, Fedorov AA, Fedorov EV, Toro R, Hillerich B, et al. Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. Proc Natl Acad Sci. 2012; 109:4122–4127. Lukk et al. attack the problem of functional characterization of proteins from genomic annotation. They find computational methods based upon structural homology are useful for guiding experimental characterization of enzymes in an automated, large-scale fashion. [PubMed: 22392983]
4. Pham VHT, Kim J. Cultivation of unculturable soil bacteria. Trends Biotechnol. 2012; 30:475–484. [PubMed: 22770837]
5. Swainston N, Smallbone K, Mendes P, Kell DB, Paton N. The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. J Integr Bioinforma. 2011; 8
6. Saha R, Chowdhury A, Maranas CD. Recent advances in the reconstruction of metabolic models and integration of omics data. Curr Opin Biotechnol. 2014; 29:39–45. [PubMed: 24632194]
7. Kell DB, Goodacre R. Metabolomics and systems pharmacology: why and how to model the human metabolic network for drug discovery. Drug Discov Today. 2014; 19:171–182. [PubMed: 23892182]
8. Richards MA, Cassen V, Heavner BD, Ajami NE, Herrmann A, Simeonidis E, Price ND. MediaDB: A Database of Microbial Growth Conditions in Defined Media. PLoS ONE. 2014; 9:e103548. [PubMed: 25098325]
9. Bateman A. Curators of the world unite: the International Society of Biocuration. Bioinformatics. 2010; 26:991–991. [PubMed: 20305270]
10. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, et al. Big data: The future of biocuration. Nature. 2008; 455:47–50. [PubMed: 18769432]
11. Österlund T, Nookaew I, Nielsen J. Fifteen years of large scale metabolic modeling of yeast: Developments and impacts. Biotechnol Adv. 201110.1016/j.biotechadv.2011.07.021
12. McCloskey D, Palsson BØ, Feist AM. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol Syst Biol. 2013; 9
13. Büchel F, Rodriguez N, Swainston N, Wrzodek C, Czauderna T, Keller R, Mittag F, Schubert M, Glont M, Golebiewski M, et al. Path2Models: large-scale generation of computational models from biochemical pathway maps. BMC Syst Biol. 2013; 7:116. [PubMed: 24180668]

14. Mueller TJ, Berla BM, Pakrasi HB, Maranas CD. Rapid construction of metabolic models for a family of Cyanobacteria using a multiple source annotation workflow. BMC Syst Biol. 2013; 7:142. [PubMed: 24369854]

15. Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, Li L, He E, Henry A, Stefan MI, et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst Biol. 2010; 4:92. [PubMed: 20587024]

16*. Monk J, Nogales J, Palsson BO. Optimizing genome-scale network reconstructions. Nat Biotechnol. 2014; 32:447–452. Monk et al. examined 117 reconstructions and found that many new reconstructions are based on existing reconstructions. Thus, the full extent of characterized enzyme function is underrepresented in reconstructions, and only a few reconstructions have added a substantial number of new reactions. [PubMed: 24811519]

17. Hamilton JJ, Reed JL. Software platforms to facilitate reconstructing genome-scale metabolic networks: Software platforms for network reconstruction. Environ Microbiol. 2014; 16:49–59. [PubMed: 24148076]

18. Karp PD, Paley S, Krieger CJ, Zhang P. An evidence ontology for use in pathway/genome databases. Pac Symp Biocomput Pac Symp Biocomput. 2004 [no volume].

19. Fernie AR, Stitt M. On the Discordance of Metabolomics with Proteomics and Transcriptomics: Coping with Increasing Complexity in Logic, Chemistry, and Network Interactions Scientific Correspondence. PLANT Physiol. 2012; 158:1139–1145. [PubMed: 22253257]

20. Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010; 5:93–121. [PubMed: 20057383]

21. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010; 28:245–248. [PubMed: 20212490]

22. Satish Kumar V, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics. 2007; 8:212. [PubMed: 17584497]

23*. Benedict MN, Mundy MB, Henry CS, Chia N, Price ND. Likelihood-Based Gene Annotations for Gap Filling and Quality Assessment in Genome-Scale Metabolic Models. PLoS Comput Biol. 2014; 10:e1003882. Benedict et al. present a novel gap-filling approach that seeks to fill reconstruction gaps with reactions that are likely present based upon genetic homology, rather than an approach that seeks the most parsimonious set of reactions to fill a gap. [PubMed: 25329157]

24. Clasquin MF, Melamud E, Singer A, Gooding JR, Xu X, Dong A, Cui H, Campagna SR, Savchenko A, Yakunin AF, et al. Riboneogenesis in Yeast. Cell. 2011; 145:969–980. [PubMed: 21663798]

25. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 2003; 19:524–531. [PubMed: 12611808]

26. Novère NL, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, et al. Minimum information requested in the annotation of biochemical models (MIRIAM). Nat Biotechnol. 2005; 23:1509–1515. [PubMed: 16333295]

27. Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Zengler K, et al. In silico method for modelling metabolism and gene product expression at genome scale. Nat Commun. 2012; 3:929. [PubMed: 22760628]

28. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, et al. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform. 2010; 11:40–79. [PubMed: 19955237]

29. Latendresse M, Krummenacker M, Trupp M, Karp PD. Construction and completion of flux balance models from pathway databases. Bioinformatics. 2012; 28:388–396. [PubMed: 22262672]

30. Villa-García MJ, Choi MS, Hinz FI, Gaspar ML, Jesch SA, Henry SA. Genome-wide screen for inositol auxotrophy in Saccharomyces cerevisiae implicates lipid metabolism in stress response signaling. Mol Genet Genomics MGG. 2011; 285:125–149. [PubMed: 21136082]

31. Teng X, Dayhoff-Brannigan M, Cheng W-C, Gilbert CE, Sing CN, Diny NL, Wheelan SJ, Dunham MJ, Boeke JD, Pineda FJ, et al. Genome-wide Consequences of Deleting Any Single Gene. Mol Cell. 201310.1016/j.molcel.2013.09.026

32**. VanderSluis B, Hess DC, Pesyna C, Krumholz EW, Syed T, Szappanos B, Nislow C, Papp B, Troyanskaya OG, Myers CL, et al. Broad metabolic sensitivity profiling of a prototrophic yeast deletion collection. Genome Biol. 2014; 15:R64. VanderSluis et al. have made a prototrophic *Saccharomyces cerevisiae* deletion library that can grow in a defined minimal medium, unlike previous deletion libraries which included auxotrophic markers to facilitate genetic studies. They profile growth phenotypes of their new library in 28 different environments, and discover previously uncharacterized fitness phenotypes. They found that yeast metabolic network models could predict fitness phenotypes with above random, but modest predictive ability. [PubMed: 24721214]

Highlights

- Assembling a network reconstruction can reveal knowledge gaps
- Building a functional metabolic model enables testable prediction
- Recent work has found that most models contain the same reactions
- Reconstruction and functional model building should be explicitly separated