Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Recent advances in the reconstruction of metabolic models and integration of omics data

Rajib Saha[1], Anupam Chowdhury[1] and Costas D Maranas

With the ever-accelerating pace of genome sequencing and annotation information generation, the development of computational pipelines for the rapid reconstruction of high-quality metabolic networks has received significant attention. Herein, we review the available biological databases and automated/semi-automated reconstruction tools. In addition, we describe available methodologies for the integration of high-throughput omics data to increase metabolic phenotype prediction accuracy. Data heterogeneity and lack of better integration of metabolic reconstruction pipelines with omics data generation protocols have hampered rapid progress thus far.

Addresses
Department of Chemical Engineering, The Pennsylvania State University, University Park, PA, USA

Corresponding author: Maranas, Costas D (costas@engr.psu.edu, costas@psu.edu)
[1] Joint first authors.

Current Opinion in Biotechnology 2014, **29**:39–45

This review comes from a themed issue on **Cell and pathway engineering**

Edited by **Tina Lütke-Eversloh** and **Keith EJ Tyo**

Available online XXX

0958-1669/$ – see front matter,

http://dx.doi.org/10.1016/j.copbio.2014.02.011

Introduction

A metabolic network captures the inter-conversion of metabolites through chemical transformations catalyzed by enzymes. To this end, a metabolic model describes reaction stoichiometry and directionality, gene to protein to reaction associations (GPRs), organelle-specific reaction localization, transporter/exchange reaction information, transcriptional/translational regulation and biomass composition [1]. By defining the metabolic space, a metabolic model can assess allowable cellular phenotypes under specific environmental and/or genetic conditions [2,3]. The number of metabolic models developed in the past several years is a testament to their increasing usefulness and penetration in many areas of biotechnology and biomedicine [4,5,6••]. Initially, metabolic models have been used to characterize biological systems and develop non-intuitive strategies to reengineer them for enhanced production of valuable bioproducts [7]. More recently, models have been developed and applied for a variety of goals ranging from metabolic disease drug, target identification, study of microbial pathogenicity and parasitism (as highlighted in [5]).

The validation of high-quality [8] models is critical for not only recapitulating known physiological properties but also improving their prediction accuracy. Towards this end, strategies have been developed to incorporate other cellular processes such as gene/protein expression to better understand the emergence of complex cellular phenotypes [9,10]. For example, genome-scale metabolic models of pathogens have been reconstructed to develop novel drugs for combating infections and also minimize side effects in the host [11]. An integrated model [12] of *E. coli* has been developed by combining **M**etabolism with gene **E**xpression (i.e. ME model) to increase the scope and accuracy of model-computable phenotypes corresponding to the optimal growth condition. In addition, by combining all of the molecular components as well as their interactions, a whole-cell model [13••] has been developed for *Mycoplasma genitalium*, a human pathogen, to study previously unexplored cellular behaviors including protein-DNA association and correlation between DNA replication initiation and replication itself. Tissue specific models have also been developed for eukaryotic organisms, such as *Homo sapiens* [14] and *Zea mays* [2], to scope out novel therapeutic targets and characterize metabolic capabilities, respectively. Moving beyond the single cell/tissue level, multi-cell/multi-tissue type metabolic models have been reconstructed for higher organisms. For example, *Homo sapiens* [14,15] models have been employed for biomedicine applications and a *Hordeum vulgare* [16] model has been deployed for studying crop improvement and yield stability.

With rapid improvements in sequencing (and annotating) tools and techniques, the number of complete genomes (and annotations) is increasing at an exponential pace [17]. Metabolic models can greatly facilitate the assessment of the potential metabolic phenotypes attainable by these organisms. Therefore, rapid development of high-quality metabolic models and algorithms for analyzing their content are of critical importance. The recent genome-scale metabolic models, their automated generation, improvements and applications have been reviewed elsewhere [4,18,19,20••] and will not be covered in detail in this review. Rather, in this mini-review we will critically evaluate the available repositories, model-building and data integration techniques and existing challenges

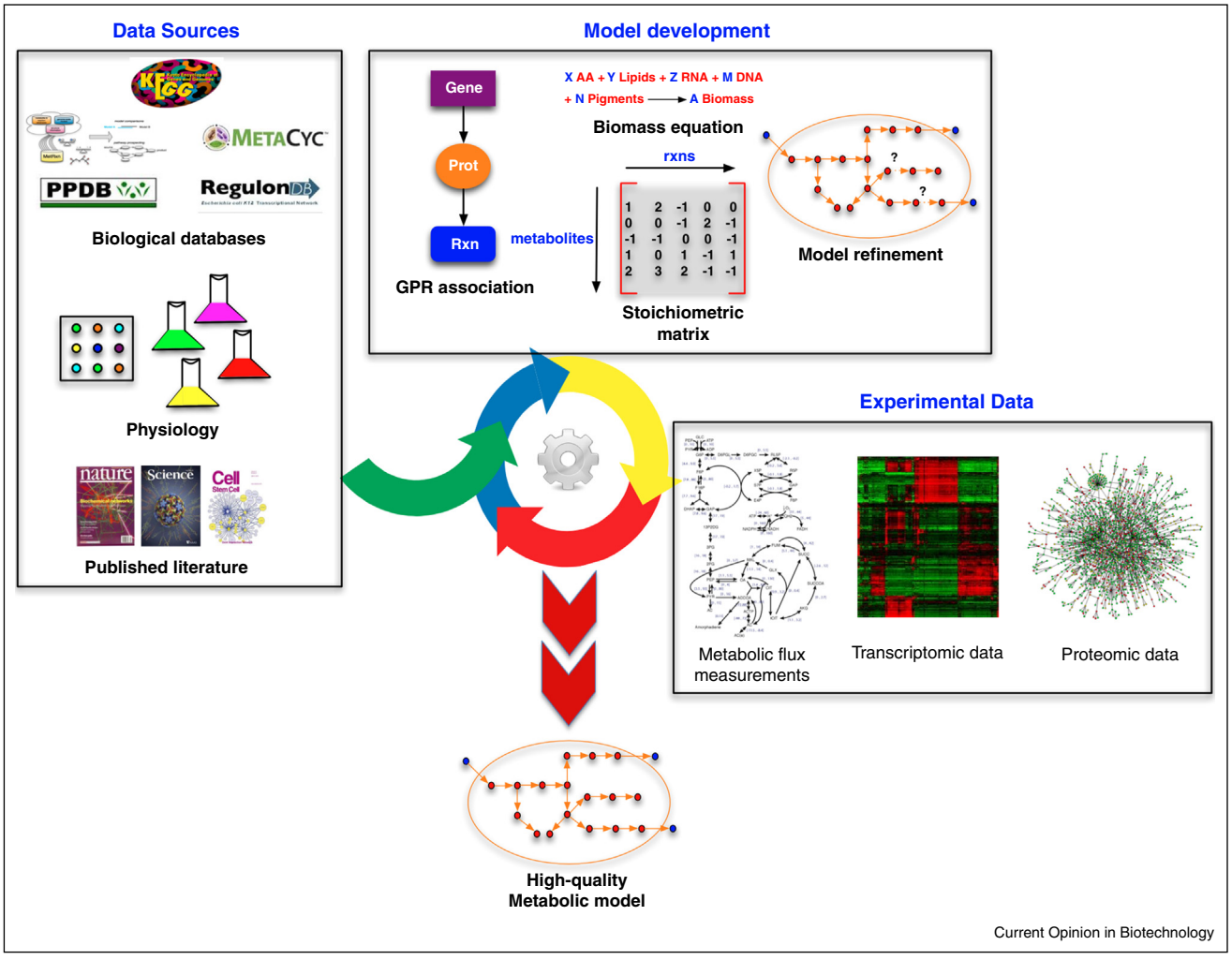
related to rapid reconstruction of high-quality metabolic models.

Metabolic model reconstruction approaches

Metabolic network/model reconstruction process follows three major steps (as highlighted in Figure 1). Initially, upon sequencing and annotating a genome of interest, literature sources and/or homology searches are used to assign function to all the **O**pen **R**eading **F**rames (i.e. ORFs). For every function with a metabolic fingerprint a specific chemical transformation is assigned. Therefore, by iteratively marching along the entire genome, a compilation of reactions encompassing the entire chemistry repertoire of the organism can be achieved. It must be noted that these models are not necessarily predictive but instead have a scoping nature by allowing us to assess what is metabolically feasible. Regulatory constraints on reaction fluxes are incorporated based on the thermodynamic (i.e. reaction reversibility) and omics (i.e. transcriptomic/proteomic) data that can further sharpen predictions.

One of the most critical steps of metabolic model building is to establish GPR information of a specific organism from biological databases and/or literature sources. To this end, biological databases (as highlighted in [1]) such as KEGG, SEED, Metacyc, BKM-react, Brenda, Uniprot, Expasy, PubChem, ChEBI and ChemSpider provide information about reactions/metabolites and associated enzymes and genes. However, as illustrated by Kumar

Figure 1



Outline for the development of a high-quality metabolic model: the first step involves retrieving data from different biological databases, physiology and biochemistry of the organism as well as published literature. In the next step, GPR associations are established, the biomass equation is described based on experimental measurements and the model is represented in the form of a stoichiometric matrix. Furthermore, gaps in the model are identified and reconciled based on established gap filling techniques. Finally, in the third step high-throughput experimental measurements, such as transcriptomic, proteomic and fluxomic, data, are utilized to improve the model accuracy.