GAVIN PUBLISHERS

Advances in Biochemistry and Biotechnology

Research Article

Nazem-Bokaee H, et al. Adv Biochem Biotehcnol: ABIO-141.

SyM-GEM: A Pathway Builder for Genome-Scale Models

Hadi Nazem-Bokaee[1], Jiun Y Yen[1], Ahmad IM Athamneh[1], Advait A Apte[1], Michael J McAnulty[1], Ryan S Senger[1,2*]

[1]Department of Biological Systems Engineering, Virginia Tech, Blacksburg, VA, USA

[2]Department of Chemical Engineering, Virginia Tech, Blacksburg, VA, USA

[*]**Corresponding author:** Ryan S Senger, Department of Biological Systems Engineering and Chemical Engineering, Virginia Tech, Blacksburg, VA, USA. Tel: +15402319501; Email: senger@vt.edu

Citation: Nazem-Bokaee H, Yen JY, Athamneh AIM, Apte AA, McAnulty MJ, et al. (2017) SyM-GEM: A Pathway Builder for Genome-Scale Models. Adv Biochem Biotehcnol: ABIO-141.

Received Date: 25 September, 2017; **Accepted Date:** 11 October, 2017; **Published Date:** 18 October, 2017

Abstract

The BioSynthetic and Metabolic Pathway Builder and Genome-Scale Model Database (SyM-GEM) is a new web application capable of adding user-generated *de novo* biosynthetic pathways to genome-scale metabolic flux models relevant to biotechnology. SyM-GEM is available freely to academic users as a web application (http://www.mesb.bse.vt.edu/SyM-GEM), and all source code is available through GitHub (https://github.com/SengerLab/SyM-GEM) under an MIT license agreement. The rapid growth in genome-scale model generation from different research groups and sources has produced incompatible systems of model nomenclature. In response, SyM-GEM incorporates (i) a novel probabilistic compound grouping algorithm, based on the PubChem Compound database, to synchronize existing genome-scale models and (ii) an application to install user-generated metabolic pathway(s) to one or multiple models. Modified or unmodified models can be downloaded from SyM-GEM in Systems Biology Markup Language (SBML) format and are compatible with the popular COnstraint-Based Reconstruction and Analysis (COBRA) Toolbox.

Introduction

Genome-scale metabolic flux models enable the study of metabolism *in silico* and have clinical and metabolic engineering applications [1-3]. Previously reviewed applications of genome-scale models include (i) analyzing the properties of complicated cellular metabolic networks [4-6], (ii) predicting phenotypes and evaluating experimental data [7-9], and (iii) metabolic engineering [10-12] with industrial, medical, and environmental applications. Often, metabolic engineering applications involve the installation of a *de novo* biosynthetic pathway. We have identified challenges in synthesizing and installing these pathways in existing genome-scale metabolic flux models so their cell-wide metabolic demands can be studied. In response, we have created the BioSynthetic and Metabolic Pathway Builder and Genome-Scale Model Database (SyM-GEM). This is a web application available freely to academic users (subject to licensing agreement available on the site) at http://www.mesb.bse.vt.edu/SyM-GEM. All source code is available to academic users through GitHub at https://github.com/SengerLab/SyM-GEM (subject to an MIT licensing agreement). In the following sections, we discuss the relationship of SyM-GEM to other existing tools, why adding a *de novo* biosynthetic pathway to an existing genome-scale model is non-trivial, and what algorithms and automations are available through SyM-GEM to simplify the process.

Several repositories for biochemical data and metabolic modeling tools now exist. Many of these are described further in the Supplementary Appendix. Among the databases of metabolic network information, a wide variety of systems of nomenclature for representing biochemical reactions and compounds are used. Over the past two decades, "genome-scale" metabolic network reconstructions have been built using different representations of compounds and reactions, which are often inconsistent with other formats. This creates a challenge in synchronizing available networks and models. This is illustrated in Table 1 for the simple case of the ATP synthase catalyzed reaction. Here, the same reaction was extracted from nine different genome-scale models. Not only do these models differ in notation of reactions and compounds, they also differ in proton balancing. Moreover, inconsistencies, redundancies, and ambiguities inside individual models and databases result in difficulties developing a universal standardized approach that enables direct incorporation of metabolic information from these models. Developments such as BiGG [13], MEMOSyS [14], MetRxn [15], KBase [16], and others have attempted to address these critical issues. The integration of

genome-scale models into a unified format is a logical step for the progression of model building and systems biology in general. Ultimately, standardized formats will enable the simulation of multiple models simultaneously. This will be extremely useful and necessary for studying microbial interactions. In the current research, the focus is placed on compound synchronization among models (while maintaining original nomenclatures) and enabling the addition of *de novo* biosynthetic and metabolic pathways in multiple genome-scale models for the production of new fuels, chemicals, and pharmaceuticals.

Organism Name (Model ID)	ATP Synthase Reaction	Reference
*Aspergillus nidulan*s (iHD666)	ADP[m] + PI[m] + 3.88 H+_PO -> ATP[m] + H_2O[m] + 3.88 H+_PO[m]	[24]
Bacillus subtilis (iBsu1103)	cpd00008 + cpd00009 + (4) cpd00067[e] <=> cpd00001 + cpd00002 + (3) cpd00067	[25]
Clostridium acetobutylicum (iCAC490)	ADP[c] + Orthophosphate[c] + 3 H+[e] <=> H_2O[c] + ATP[c] + 2.48 H+[c]	[8]
Corynebacterium glutamicum	ADP + Pi + 4 PROTON[e] <=> ATP + WATER + PROTON	[26]
Escherichia coli (iAF1260)	ADP[c] + (4) H[p] + pi[c] <==> ATP[c] + (3) H[c] + H_2O[c]	[27]
Geobacter sulfurreducens	ADP[c] + (4) H[e] + pi[c] --> ATP[c] + (3) H[c] + H_2O[c]	[28]
Methanosarcina barkeri (iMG746)	ADP[c] + 4 H[e] + pi[c] <=> ATP[c] + H_2O[c] + 3 H[c]	[29]
Pseudomonas putida (iJP815)	IC00008 + IC00009 + 4 EC00065[e] <==> IC00001 + IC00002 + 3 IC00065	[30]
Saccharomyces cerevisiae (iND750)	ADP[g] + (3) H[c] + pi[g] --> ATP[g] + (2) H[g] + H_2O[g]	[31]

Table 1: Representation of the ATP Synthase Reaction in Several Genome-Scale Models.

SyM-GEM was constructed with genome-scale models relevant to metabolic engineering in biotechnology and a novel algorithm that synchronizes the notation of compounds and reactions between models prepared in different formats. Overall, SyM-GEM incorporates: **(i)** deriving standardization between genome-scale models of different formats; **(ii)** storing synchronized models in an on-line database; **(iii)** adding user-derived *de novo* synthetic metabolic pathways to one, multiple, or all models in the database; and **(iv)** retrieving the original and/or modified model(s) in Systems Biology Markup Language (SBML) [17] format. To develop this application, a novel algorithm based on probabilistic associations was created to pinpoint identical compounds across models of differing formats and naming conventions. The method of indexing compound data across multiple genome-scale models used herein makes use of the well-curated data available publicly through the PubChem Compound database [18,19]. From there, the online genome-scale model database portion of SyM-GEM was created and contains the standardized models. This database serves as the platform for *de novo* biosynthetic metabolic pathway addition to genome-scale models through an interactive web application. Currently, the database contains 12 models (listed in Table 2), and all have relevance to biotechnology.

Systems and Methods

The SyM-GEM online database and web application for metabolic pathway addition was developed using MySQL and PHP and has been made freely available to academic users under a licensing agreement available on the site (http://www.mesb.bse.vt.edu/SyM-GEM). The source-code for the compound grouping algorithm is also available to academic users under a licensing agreement through GitHub (https://github.com/SengerLab/SyM-GEM). Genome-scale models were simulated using MATLAB (R2015b) (MathWorks, Natick, MA) with the Constraint-Based Reconstruction and Analysis (COBRA) Toolbox v2.0.5 [20] and the open-source GLPK (GNU Linear Programming Kit) software (http://www.gnu.org/software/glpk). To read and write models as SBML, the SBML Toolbox v4.1.0 and SBML library v5.8.0 were used [17].

Results

The overall methodology developed for SyM-GEM is shown in Figure 1 and is described in detail in the Supplementary Appendix. The major components of SyM-GEM are described in the following 6 steps. Step 1: Read and convert genome-scale models for use with COBRA and SyM-GEM. Genome-scale models encoded in SBML format can be read directly by the COBRA Toolbox in MATLAB. However, to our knowledge, no universal script exists to read models prepared in other formats. In response, a MATLAB script was created (and made available through GitHub) that uses regular expressions to parse genome-scale models encoded in virtually any format including those used in the BiGG [13], KEGG [21], MetaCyc [22], Model SEED [23], and MetRxn [15] databases and makes them compatible with the COBRA Toolbox in MATLAB. Steps 2-5: Synchronization of compounds using a novel Compounds Grouping Algorithm (CGA). Here, compound names are interfaced with the PubChem

Compound database [18] to identify identical compounds with different names and identifiers used by different models. The CGA is separated into distinct steps for clarity, and the CGA MATLAB script is also available through GitHub. Step 2: Find identical compounds across all models by querying against PubChem. This step uses direct search function to identify compound names in multiple models and in PubChem. Step 3: Find additional identical compounds by calculating the probability of a compound being identical or similar to others across all models and the PubChem Compound database. This probabilistic model for grouping identical and similar compounds across all models is novel to SyM-GEM and is described in further detail and demonstrated in the Supplementary Appendix. Step 4: Generate a repository of identical compounds. Step 5: Evaluate the groups in the repository, detect anomalies, and remove redundancies. Step 6: Build the SyM-GEM genome-scale model database and web application.

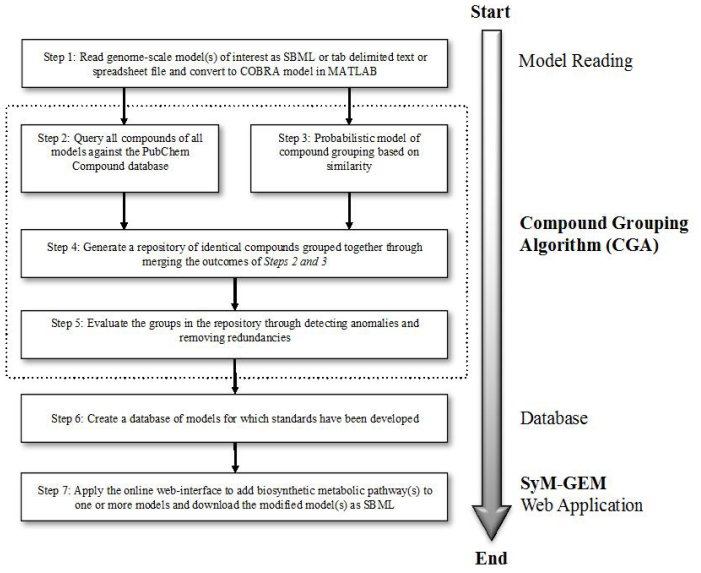

Figure 1: Flowchart of SyM-GEM development: **(i)** custom reading of differently formatted genome-scale models, **(ii)** synchronization of models through a novel Compound Grouping Algorithm (CGA), **(iii)** development of an on-line database of genome-scale models, and **(iv)** a web application to add synthetic metabolic pathway(s) to one or multiple models.

The initial version of SyM-GEM was constructed using 11 available models and contains 7,098 unique compounds. SyM-GEM has since been updated to contain 12 models (Table 2). The initial CGA resulted in construction of 2,510 non-redundant compound groups. A comprehensive step-by-step tutorial of how to use SyM-GEM to add a *de novo* biosynthetic pathway to an existing genome-scale model and download a COBRA-compatible SBML version of the new model is also presented in the Supplementary Appendix.

Organism Name (Model ID)	Reference
Bacillus subtilis (iBsu1103)	[25]
Clostridium acetobutylicum ATCC 824 (iCAC498)	[8]*
Clostridium beijerinkii NCIMB 8052 (iCM925)	[32]
Corynebacterium glutamicum ATCC 13032 (Seed196627.4)	[33]
Escherichia coli K-12 MG1655 (iAF1260)	[27]
Geobacter sulfurreducens	[28]
Lactobacillus plantarum WCFS1	[34]
Methanosarcina barkeri (iMG746)	[29]
Pseudomonas putida KT2440 (Seed160488.1)	[33]
Saccharomyces cerevisiae (iND750)	[31]
Synechocystis sp. strain PCC6803 (iJN678)	[35]
Yarrowia lipolytica (iYL619_PCP)	[36]
*The iCAC498 model has replaced the originally published iCAC490 model.	

Table 2: Genome-Scale Models Currently Available in SyM-GEM.

Discussion

SyM-GEM is a novel database and web application that allows systems biologists and metabolic engineers to access synchronized genome-scale metabolic flux models through a user-friendly on-line platform. Users can further modify available models by constructing and adding customized synthetic metabolic pathways. The original and/or modified models are downloadable as SBML files that can be readily used by common software packages such as MATLAB with the popular COBRA Toolbox. The underlying approach that achieved synchronization among differently formatted genome-scale models, the CGA, uses a series of data-mining steps and probabilistic modeling that collectively provide a fast and reliable method for finding and grouping identical compounds. This algorithm does not require large storage space for pooling all available biochemical data. In Step 2 of compound synchronization between models, the CGA directly queries information from the on-line well-curated PubChem Compound database. Thus, the problem of redundancies and inconsistencies can be bypassed due to its immense size and inclusion of all known names for each compound. Moreover, the PubChem Compound database has cross-linked information on compounds from over more than 200 sources and databases. The probabilistic association methodology (Step 3) along with the final compound treatment and evaluation procedure of the CGA resolve compound associations for those compounds that could not be found in the PubChem Compound database or lack specificity in their names. These inconsistencies arise mainly because (i) there are compounds in models with poor

annotations (e.g., "salcn," "Unknown product," and "Acceptor"), too specific annotations (e.g., "Cell wall of B. subtilis" and "(O16 antigen)x4 core oligosaccharide lipid A"), or wrong annotations (e.g., "3',5'-Cyclic AMP, Adenosine 3[-,5[-cyclic monophosphate, Cyclic adenylic acid"), and (ii) the PubChem Compound database does not contain all macromolecules (e.g., tRNAs and lipids). The CGA allowed for finding identical compounds from a pool of 7,098 compounds collected from the original 11 genome-scale models and combining them into 2,510 unique groups. For example, the CGA was capable of grouping "L-Glutamate 5-Phosphate," "L-Glutamyl 5-phosphate," "L-Glutamyl_5-phosphate," "glu5p," and "cpd02097" into the same compound group. All of these, in fact, represent the same compound. The algorithm has the potential of improving its performance as more information is fed by adding additional genome-scale models to the database. For example, in the initial version of SyM-GEM, compounds "Triphosphate" and "Inorganic triphosphate" could not be linked together under the same group because none of the models in the database support this relationship. In addition, the CGA has the potential, in future implementations, to incorporate user knowledge to improve and/or assign associations; thus, further curation and improvements through "Crowd sourcing" are possible in future versions of SyM-GEM.

The SyM-GEM web application offers an effective method for incorporating new biosynthetic metabolic pathways into one or multiple genome-scale models. These modified models enable comprehensive analysis and comparison among various metabolic engineering strategies for optimal production of a specific target. SyM-GEM allows construction of a *de novo* biosynthetic pathway and its simultaneous addition to multiple models without user regard for the different model formats or specific nomenclatures used by individual model. In the current version of the SyM-GEM, there are both highly-curated models and draft reconstructions. The approach designed in this research has the potential to be complemented with automated genome-scale model reconstruction pipelines or even other databases of models to enable: (i) synchronization among newly published models and those residing in SyM-GEM, (ii) addition of biosynthetic metabolic pathways to an expanded number of models, and (iii) producing modified SBML files of models for widespread distribution. Ultimately, the goal of SyM-GEM is to support systems biology and metabolic engineering where one or more biosynthetic pathways are added and evaluated in multiple hosts simultaneously. This will aid in choosing a proper host for expression of a desired pathway by allowing exploration of factors such as cofactor availability, biosynthesis of competing products, substrate utilization, and bioprocess considerations.

References

1. Feist AM, Palsson BO (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nature biotechnology 26: 659-667.
2. Milne CB, Kim PJ, Eddy JA, Price ND (2009) Accomplishments in genome-scale *in silico* modeling for industrial and medical biotechnology. Biotechnology journal 4: 1653-1670.
3. Kim TY, Sohn SB, Kim YB, Kim WJ, Lee SY (2012) Recent advances in reconstruction and applications of genome-scale metabolic models. Current opinion in biotechnology 23: 617-623.
4. Schilling CH, Edwards JS, Letscher D, Palsson BO (2000) Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. Biotechnology and bioengineering 71: 286-306.
5. Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2: 886-897.
6. Senger RS (2010) Biofuel production improvement with genome-scale models: The role of cell composition. Biotechnology journal 5: 671-685.
7. Edwards JS, Ibarra RU, Palsson BO (2001) *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. Nature biotechnology 19: 125-130.
8. McAnulty MJ, Yen JY, Freedman BG, Senger RS (2012) Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism *in silico*. BMC Syst Biol 6: 42.
9. Senger RS, Nazem-Bokaee H (2012) Resolving Cell Composition Through Simple Measurements, Genome-Scale Modeling, and a Genetic Algorithm, Systems Metabolic Engineering, Humana Press: 85-101.
10. Burgard AP, Pharkya P, Maranas CD (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnology and bioengineering 84: 647-657.
11. Rocha I, Maia P, Evangelista P, Vilaca P, Soares S, et al. (2010) OptFlux: an open-source software platform for *in silico* metabolic engineering. BMC Syst Biol 4: 45.
12. Yen JY, Nazem-Bokaee H, Freedman BG, Athamneh AI, Senger RS (2013) Deriving metabolic engineering strategies from genome-scale modeling with flux ratio constraints. Biotechnology journal 8: 581-594.
13. Schellenberger J, Park JO, Conrad TM, Palsson BO (2010) BiGG: A Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11: 213.
14. Pabinger S, Snajder R, Hardiman T, Willi M, et al. (2014) MEMOSys 2.0: an update of the bioinformatics database for genome-scale models and genomic data. Database (Oxford): bau004.
15. Kumar A, Suthers PF, Maranas CD (2012) MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. BMC Bioinformatics 13: 6.
16. Arkin AP, Stevens RL, Cottingham RW, Maslov S, Henry CS, et al. (2016) The DOE Systems Biology Knowledgebase (KBase). bioRxiv.
17. Keating SM, Bornstein BJ, Finney A, Hucka M (2006) SBMLToolbox: an SBML toolbox for MATLAB users. Bioinformatics 22: 1275-1277.
18. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, et al. (2016) PubChem Substance and Compound databases. Nucleic Acids Res 44: D1202-D1213.

19. Bolton E, Wang Y, Thiessen PA, Bryant SH (2008) PubChem: Integrated Platform of Small Molecules and Biological Activities, Annual Reports in Computational Chemistry, American Chemical Society.

20. Schellenberger J, Que R, Fleming RM, Thiele I, Orth JD, et al. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature protocols 6: 1290-1307.

21. Kanehisa M, Goto S (2008) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30.

22. Krieger CJ, Zhang P, Mueller LA, Wang A, et al. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 32: D438-D442.

23. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature biotechnology 28: 977-982.

24. David H, Ozcelik IS, Hofmann G, Nielsen J (2008) Analysis of *Aspergillus nidulans* metabolism at the genome-scale. BMC Genomics 9: 163.

25. Henry CS, Zinner JF, Cohoon MP, Stevens RL (2009) iBsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations. Genome Biol 10: R69.

26. Shinfuku Y, Sorpitiporn N, Sono M, Furusawa C, et al. (2009) Development and experimental verification of a genome-scale metabolic model for *Corynebacterium glutamicum.* Microb Cell Fact 8: 43.

27. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3: 121.

28. Mahadevan R, Bond DR, Butler JE, Esteve-Nunez A, Coppi MV, et al. (2006) Characterization of metabolism in the Fe(III)-reducing organism *Geobacter sulfurreducens* by constraint-based modeling. Appl Environ Microbiol 72: 1558-1568.

29. Gonnerman MC, Benedict MN, Feist AM, Metcalf WW, Price ND (2013) Genomically and biochemically accurate metabolic reconstruction of *Methanosarcina barkeri* Fusaro, iMG746. Biotechnology journal 8: 1070-1079.

30. Puchalka J, Oberhardt MA, Godinho M, Bielecka A, Regenhardt D, et al. (2008) Genome-scale reconstruction and analysis of the *Pseudomonas putida* KT2440 metabolic network facilitates applications in biotechnology. PLoS computational biology 4: e1000210.

31. Duarte NC, Herrgard MJ, Palsson BO (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. Genome Res 14: 1298-1309.

32. Milne CB, Eddy JA, Raju R, Ardekani S, Kim P, et al. (2011) Metabolic network reconstruction and genome-scale model of butanol-producing strain *Clostridium beijerinckii* NCIMB 8052. BMC Syst Biol 5: 130.

33. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature biotechnology 28: 977-982.

34. Teusink B, Wiersma A, Molenaar D, Francke C, et al. (2006) Analysis of growth of *Lactobacillus plantarum* WCFS1 on a complex medium using a genome-scale metabolic model. Journal of Biological Chemistry 281: 40041-40048.

35. Nogales J, Gudmundsson S, Knight EM, Palsson BO, Thiele I (2012) Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. Proc Natl Acad Sci U S A 109: 2678-2683.

36. Pan P, Hua Q (2012) Reconstruction and *in silico* analysis of metabolic network for an oleaginous yeast, *Yarrowia lipolytica*. PLoS One 7: e51535.