GAP-FILLING WITH THE ATLAS OF BIOCHEMISTRY TO RESOLVE METABOLIC GAPS IN *E. coli*

E. Vayena[1,2], A. Chiappino-Pepe[1], N. Hadadi[1], H. Mohammadi[1], M. Ataman[1], J. Hafner[1], S. Pavlou[2], V. Hatzimanikatis[1*]

[1] Laboratory of Computational Systems Biotechnology (LCSB), Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland.

[2] Department of Chemical Engineering, University of Patras, Caratheodory 1, University Campus, GR-26504 Patras, Greece

(* vassily.hatzimanikatis@epfl.ch)

ABSTRACT

Advances in medicine and biotechnology rely on the further understanding of biological processes. Despite the technological advances and increasing available types and amounts of omics data, significant biochemical knowledge gaps remain uncharacterized. We necessitate methods that enable systematically analysing the growing sets of data and identifying the knowledge gaps. Several approaches have been developed during the past decades to identify missing metabolic annotations in genome-scale models (GEMs). However, these approaches suggest metabolic reactions within a limited set of already characterized metabolic capabilities.

In this study, we propose a workflow to identify, classify and characterize missing metabolic capabilities in GEMs using the ATLAS of Biochemistry. The ATLAS of Biochemistry, which involves more than 149,000 possible enzymatic reactions between known biological compounds, represents the upper bound of missing biochemistry and is hence a guide to fill the gaps. We apply our gap-filling approach to the latest genome-scale model of *Escherichia coli* (iML1515) and develop a database of top suggested biochemistry that can indicate its missing metabolic capabilities. Interestingly, some gaps cannot be filled with the ATLAS of Biochemistry and represent biochemical bottlenecks for further analysis. Overall, our approach will be a reference and valuable tool for the reconstruction and refinement of metabolic networks, and our results will accelerate experimental studies toward fully annotated genomes.

INTRODUCTION

Genome-scale models constitute knowledge databases that include biochemical, genetical and genomical information about a specific organism and represent all known metabolic capabilities of the organism [1]. GEMs are powerful tools in the fields of metabolic engineering and synthetic biology. In genome-scale models the reaction network of the organism is mathematically represented by a stoichiometric matrix. In this matrix each row represents a metabolite and each column represents a reaction. Upper and lower flux bounds are imposed on each reaction.

However, even the most well curated GEMs are incomplete and contain information gaps [2]. Missing information in a network may be in the form of gaps or in the form of orphan reactions [2]. Gaps are holes in the metabolic network caused by missing reactions. These gaps create dead – ends in the network since metabolites are either only produced or consumed. Orphan reactions account for reactions integrated in the metabolic network that are not assigned to any enzyme. Both types of gaps result in decreased predictive capability of the model. Hence, methods have been developed to fill these knowledge gaps. Gap-filling methods can serve as tools to improve the predictive performance of GEMs, so that GEMs can serve further research and facilitate industrial exploitation of the organism. In addition, gap-filling algorithms suggest different metabolic pathways to complete the same metabolic task, that differ in terms of substrate utilization, byproducts and biomass yield, and thus, reveal possible targets for metabolic engineering and synthetic biology.

Several gap-filling algorithms have been developed the past decade in order to curate GEMs [2][4]. The main limitation of these approaches is the fact that they use universal databases of reactions to fill the gaps. Here we propose a workflow to gap fill GEMs using the ATLAS of Biochemistry, a well curated reaction repository generated by Hatzimanikatis et al. [3]. Contrary to already existing databases, ATLAS includes novel enzymatic reactions, that account for all theoretically possible ways to link two metabolites and can therefore be considered as the upper bound of biochemistry. Thus, gap-filling a GEM using ATLAS can lead to the integration of reactions that are not yet discovered but may be more biologically relevant than already documented reactions. Apart from adding new metabolites to the model, ATLAS can introduce new ways to connect metabolites already existing in the wildtype model, altering at a minimum level its biochemistry. By combining ATLAS-only reactions and already known reactions, novel sets of reactions can be used to fill in knowledge gaps in GEMS. In addition, gap-filling solution suggested by this method could be applied in metabolic engineering.

The workflow is used to gap-fill the metabolic network of *Escherichia coli* [5] and the results are tested under thermodynamic constraints [6] and ranked depending on their effect on the performance and biochemistry of the original metabolic network. The Atlas of Biochemistry is compared to the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [7], which we consider to be the reference for known biochemical reactions and metabolites, regarding its performance as a reaction pool for gap-filling.

METHODS

Here we present our gap-filling approach (Figure 1), which comprises five main steps. The workflow aims to produce an expanded metabolic network, that contains the native biochemistry of the organism and information originated from the ATLAS of Biochemistry, by connecting the metabolic network of study with ATLAS.

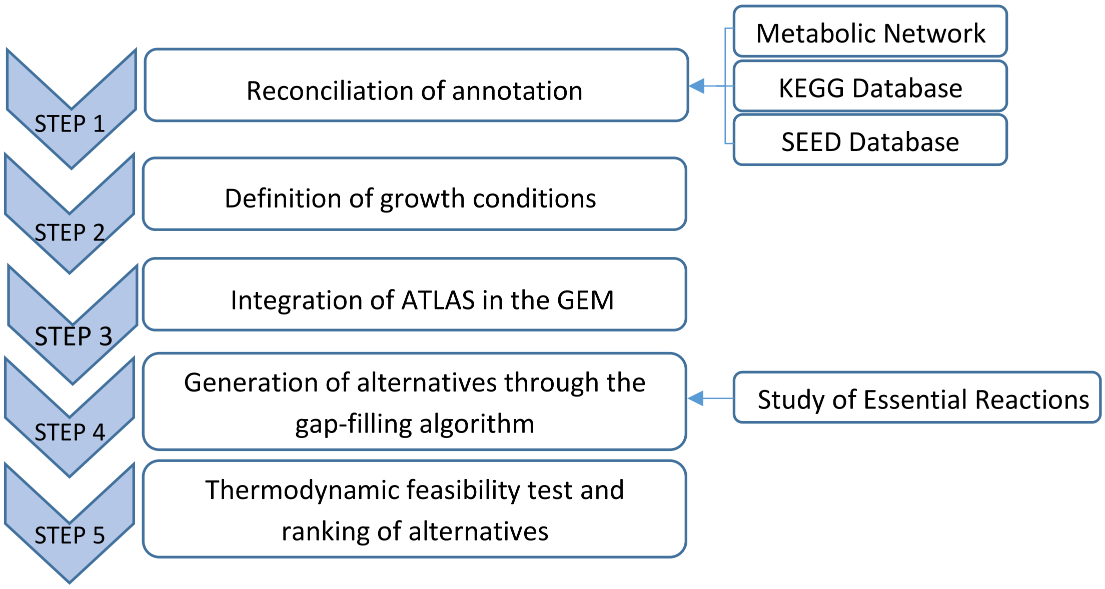
Metabolic Network
KEGG Database
SEED Database
STEP 1
Reconciliation of annotation
STEP 2
Definition of growth conditions
STEP 3
Integration of ATLAS in the GEM
STEP 4
Generation of alternatives through the gap-filling algorithm
Study of Essential Reactions
STEP 5
Thermodynamic feasibility test and ranking of alternatives

Figure 1. *The gap-filling workflow.*

STEP 1 In order to properly connect the metabolic network of the organism of study and ATLAS, it is essential that the same annotation is used, that means that there must be a match in the metabolite IDs. Since ATLAS is based on the KEGG database, the metabolites of the model are mapped to the

KEGG database. The metabolites are also mapped to the SEED database so that the calculation of the thermodynamic properties of the model is possible.

STEP 2 Since the gap-filling workflow is based on reaction essentiality, it is necessary to further explore under what conditions a reaction is essential for biomass production. Substrate availability is a factor that can greatly affect the phenotype of an organism, since the presence or the absence of different compounds in the growth medium triggers different metabolic pathways and therefore, growth conditions must be well defined. This fact evinces the need of gap-filling a GEM under different growth conditions in order to better curate the gaps in the metabolism of an organism.

STEP 3 The third step is to integrate the ATLAS of biochemistry in the metabolic network we aim to gap-fill. Once the model and the database are merged, a reaction essentiality analysis is carried out for the original GEM and for the expanded one. The reactions that appear to be essential for the original GEM but not for the expanded one, are referred to as rescued reactions, constitute the target reactions for the gap-filling algorithm. The essentiality test of the merged model can unveil reactions of that are of great importance for the survival of the organism. Given the fact that these reactions are indispensable for the organism, they constitute targets for further study and can even be appointed as candidate targets for the development of new medicine.

STEP 4 After the rescued reactions are identified, the gap-filling algorithm attempts to find alternative sets of reactions that could substitute for the rescued reactions. More specifically, the algorithm knocks out each one of the rescued reactions, one at a time, and using a mixed integer linear programming (MILP) formulation attempts to restore the functionality of the metabolic network by allowing flux through a minimum set of reactions. The algorithm returns the sets of solutions of the minimum size and the sequent size.

STEP 5 The sets of alternative solutions can differ in the number of reactions, the compounds and the enzymes they integrate. Each set of reactions leads to model predictions that may match or differ from the predictions of the original model. Thus, the alternative solutions are evaluated and ranked. The solutions that result in a thermodynamically infeasible metabolic network are rejected. The rest of the solutions are ranked according to their effect on the original GEM, as far as its biochemistry and performance are concerned. The larger the number of different non-native metabolites and enzymes are added by an alternative to the model, the lower the alternative is ranked. In addition, alternatives that greatly increase or decrease the maximum growth yield or the thermodynamic growth yield are ranked lower than those that have little or no effect at all on the performance of the original GEM. Another criterion is the number of the reactions that are used to replace each rescued reaction. Usually organisms do not favor larger pathways since they normally require more substrate and cofactors and are in general less fast and more difficult to realize than shorter pathways.

RESULTS

In this project the metabolic network of *E. coli* was expanded based on the compounds that already existed in the iML1515 model (1,169 unique metabolites across all compartments). Two databases (Table 1), a subset of the KEGG database and a subset of ATLAS, were examined. The first database, Ecoli_mets_KEGG, contains only KEGG reactions that integrate metabolites exclusively coming from the wildtype model. The second database, Ecoli_mets_ATLAS, consists of reactions, among metabolites coming from the wildtype model, that are part of ATLAS. By gap-filling using these two databases, we examined whether and at what level different connectivity between native metabolites can address effectively the knowledge-gap problem.

Table 1. *Information on the databases used for gap-filling.*

Database	Compounds	Reactions
Ecoli_mets_KEGG	810 SEED ID: 716	1,263
Ecoli_mets_ATLAS	778 SEED ID: 698	9,204

Figure 2 compares the ATLAS of Biochemistry and the KEGG database, regarding their performance as reaction pools. After the thermodynamic feasibility analysis, the Ecoli_mets_ATLAS database can rescue 200 out of the 355 essential reactions. 84% of the reactions save by this database can be substituted with 2 or more alternative solutions sets, while the heuristic returns more than 10 alternative solutions sets for 51% of the rescued reactions. These solution sets integrate both novel and already known reactions. On the other hand, the Ecoli_mets_KEGG database can rescue only 101 reactions and there is only one feasible solution for 70% of the rescued reactions. However, 14 out of these 101 reactions are not rescued by the subset of ATLAS (Table 2). These reactions can be substituted by only one reaction solution set. Further analysis of these 14 cases revealed that the exact reconstructions of these KEGG reactions are not included in ATLAS since the generalized rules that were used to generate ATLAS do not apply to the compounds they integrate, or they are reconstructed in 1-step or multi-steps biotransformations, that are not included in the subset of ATLAS used.

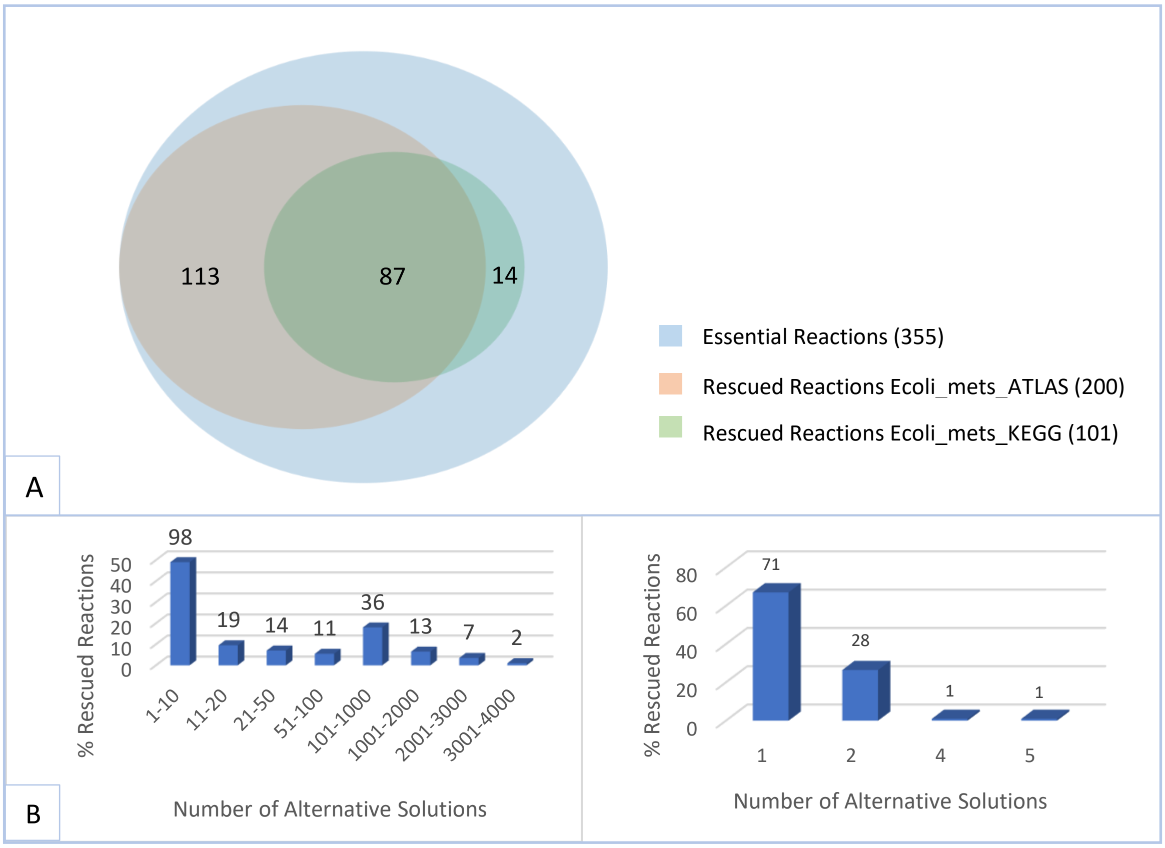

Figure 2. *A) Venn diagram comparing the two databases used for gap-filling the metabolic network of E. coli. B) Comparison of the number of alternative solutions per rescued reaction (Ecoli_mets_ATLAS on the left, Ecoli_mets_KEGG on the right).*

Table 2. *Reactions rescued only by Ecoli_mets_KEGG database.*

Rescued Reaction ID	Gap-filling solution KEGG ID	Comments
IPPMIa	R10170	Lump of IPPMIa and IPPMIb
EX_cu2_e	R09735	Not reconstructed - Compound with complex structure
CU2tex		
CU2tpp		
MOCOS		
DHPS2	R03066	Not reconstructed
DHNAOT4	R05617	Not reconstructed - Compound with undefined structure
DHNCOAS	R04150	Not reconstructed
ENTCS	R07644	Not reconstructed - Compound with complex structure
SERASr		
DHBS		
GTPCI	R11072	Not reconstructed
G5SADs	R00708	1 step biotransformation
SULR	R10146	Biotransformation via one PubChem intermediate

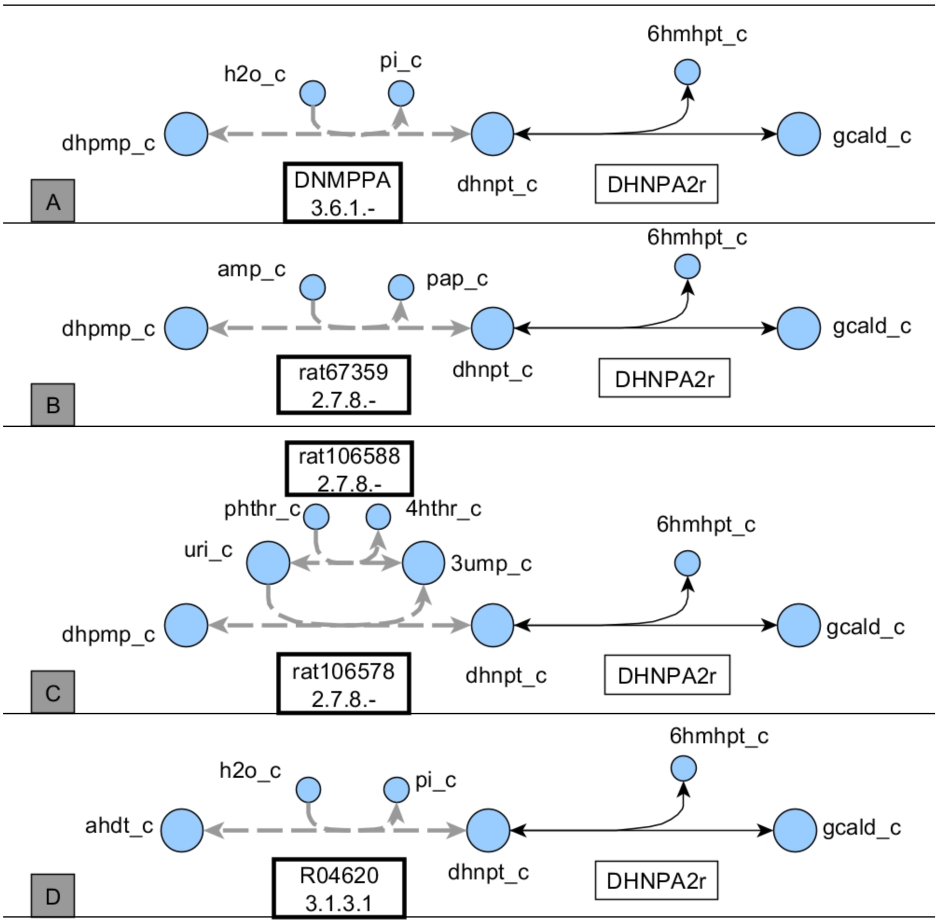

Figure 3. *(A) The reaction DNMPPA in the original network. (B) The alternative provided by the Ecoli_mets_ATLAS database with the highest score and (C) with the lowest according to our ranking system. (D) The only alternative provided by the Ecoli_mets_KEGG database.*

There are 87 reactions rescued by both databases. However, for 86 out of the 87 reactions the ATLAS subset provides more solutions than the KEGG subset. The Ecoli_mets_KEGG database provides two thermodynamically feasible solutions that substitute for the reaction Pyrroline-5-carboxylate reductase (P5CR), R01248 and R10507, while Ecoli_mets_ATLAS provides only one, rat9091 which is the exact reconstruction of R01248.

Here, we present an example of a reaction rescued by both databases. Dihydroneopterin monophosphate dephosphorylase (DNMPPA) (Figure 3.A) is an orphan reaction. It participates in the chorismite metabolism, the tetrahydrofolate biosynthesis pathways. When gap-filling the model with the Ecoli_mets_ATLAS database the heuristic returns 1814 thermodynamically feasible alternative solution sets (Figure 3.B and C). 82 out of the 1814 alternatives integrate only one reaction and do not require any non-native enzymes, while they have a matching biomass production with the original GEM, both with and without thermodynamic constraints. On the other hand, when using the Ecoli_mets_KEGG database the heuristic returns a single solution (Figure 3.D), that enables the gap-filled GEM to have a matching performance to the original one. Interestingly, the KEGG reaction used to gap-fill the GEM, R04620, is reconstructed in ATLAS through multiple 1-step biotransformations. These ATLAS reactions are part of the solutions sets proposed by the Ecoli_mets_ATLAS database, possessing highly rated positions among the alternatives.

CONCLUSIONS

The ATLAS of Biochemistry is proved to be a powerful tool in the gap-filling field. Integrating more than 149,000 novel and already known enzymatic reactions, ATLAS provides a wider range of gap-filling solutions. The subset of ATLAS used in this study can rescue 200 out of 355 essential reactions in iML1515, while the corresponding subset of KEGG can rescue only 101 reactions. After the thermodynamic analysis, 84% of the reactions rescued by the Ecoli_mets_ATLAS database can be substituted by 2 or more alternative solution sets, while there is only one feasible solution for 70% of the reactions rescued by the Ecoli_mets_KEGG database. Integrating additional layers of information from ATLAS to the GEM is expected to suggest reactions to fill more gaps in the metabolism of *Escherichia coli.* Gap-filling GEMs with the ATLAS of Biochemistry and our gap-filling algorithm can accelerate experimental studies towards fully curated metabolic network reconstructions and reveal new pathways in the scope of retrobiosynthesis.

REFERENCES

[1] O'Brien, E. J., Monk, J. M., & Palsson, B. O. (2015). *Cell, 161*(5), 971-987.

[2] Orth, J. D., & Palsson, B. (2010). *Biotechnol Bioeng, 107*(3), 403-412.

[3] Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A., & Hatzimanikatis, V. (2016). *ACS Synth Biol, 5*(10), 1155-1166.

[4] Pan, S., & Reed, J. L. (2018). *Curr Opin Biotechnol, 51*, 103-108.

[5] Monk, J., Lloyd, C., Brunk, E., Mih, N., Sastry, A., King, Z., et al. (2017). *Nature Biotechnology, 35*(10), 904-908.

[6] Salvy, P., Fengos, G., Ataman, M., Pathier, T., Soh, K. C., & Hatzimanikatis, V. (2018). *Bioinformatics*

[7] Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., & Tanabe, M. (2018). *Nucleic Acids Res*.