Correspondence

OPEN ACCESS

molecular systems biology

Do genome-scale models need exact solvers or clearer standards?

Ali Ebrahim[1], Eivind Almaas[2], Eugen Bauer[3], Aarash Bordbar[4], Anthony P Burgard[5], Roger L Chang[6], Andreas Dräger[1,7], Iman Famili[8], Adam M Feist[1], Ronan MT Fleming[3], Stephen S Fong[9], Vassily Hatzimanikatis[10], Markus J Herrgård[11], Allen Holder[12], Michael Hucka[13], Daniel Hyduke[14], Neema Jamshidi[15,16], Sang Yup Lee[11,17], Nicolas Le Novère[18], Joshua A Lerman[1], Nathan E Lewis[19], Ding Ma[20], Radhakrishnan Mahadevan[21], Costas Maranas[22], Harish Nagarajan[5], Ali Navid[23], Jens Nielsen[11,24], Lars K Nielsen[25], Juan Nogales[26], Alberto Noronha[3], Csaba Pal[27], Bernhard O Palsson[1], Jason A Papin[28], Kiran R Patil[29], Nathan D Price[30], Jennifer L Reed[31], Michael Saunders[20], Ryan S Senger[32], Nikolaus Sonnenschein[11], Yuekai Sun[33] & Ines Thiele[3]

Mol Syst Biol. (2015) 11: 831
Comment on: **L Chindelevitch *et al*** (October 2014)
See reply: **L Chindelevitch *et al*** (in this issue)

Constraint-based analysis of genome-scale models (GEMs) arose shortly after the first genome sequences became available. As numerous reviews of the field show, this approach and methodology has proven to be successful in studying a wide range of biological phenomena (McCloskey *et al*, 2013; Bordbar *et al*, 2014). However, efforts to expand the user base are impeded by hurdles in correctly formulating these problems to obtain numerical solutions. In particular, in a study entitled "An exact arithmetic toolbox for a consistent and reproducible structural analysis of metabolic network models" (Chindelevitch *et al*, 2014), the authors apply an exact solver to 88 genome-scale constraint-based models of metabolism. The authors claim that COBRA calculations (Orth *et al*, 2010) are inconsistent with their results and that many published and actively

1 Department of Bioengineering, University of California, San Diego, CA, USA. E-mail: aebrahim@ucsd.edu
2 Department of Biotechnology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
3 Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belval, Luxembourg
4 Sinopia Biosciences Inc., San Diego, CA, USA
5 Genomatica, Inc., San Diego, CA, USA
6 Department of Systems Biology, Harvard Medical School, Boston, MA, USA
7 Center for Bioinformatics Tuebingen (ZBIT), University of Tuebingen, Tübingen, Germany
8 Intrexon, Inc., San Diego, CA, USA
9 Department of Chemical and Life Science Engineering, Virginia Commonwealth University, Richmond, VA, USA
10 Laboratory of Computational Systems Biotechnology, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
11 The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark
12 Department of Mathematics, Rose-Hulman Institute of Technology, Terre Haute, IN, USA
13 Department of Computing and Mathematical Science, California Institute of Technology, Pasadena, CA, USA
14 Department of Biological Engineering, Utah State University, Logan, UT, USA
15 Department of Radiology, University of California, Los Angeles, CA, USA
16 Institute of Engineering in Medicine, University of California, San Diego, CA, USA
17 Department of Chemical and Biomolecular Engineering (BK21 Plus Program), Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
18 Babraham Institute, Cambridge, UK
19 Department of Pediatrics, University of California, San Diego, CA, USA
20 Department of Management Science and Engineering, Stanford University, Stanford, CA, USA
21 Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, Ontario, Canada
22 Department of Chemical Engineering, Pennsylvania State University, University Park, PA, USA
23 Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, CA, USA
24 Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden
25 Australian Institute for Bioengineering & Nanotechnology (AIBN), The University of Queensland, Brisbane, Queensland, Australia
26 Department of Environmental Biology, Centro de Investigaciones Biológicas (CSIC), Madrid, Spain
27 Synthetic and Systems Biology Unit, Biological Research Center, Szeged, Hungary
28 Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, USA
29 European Molecular Biology Laboratory, Heidelberg, Germany
30 Institute for Systems Biology, Seattle, WA, USA
31 Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, WI, USA
32 Department of Biological Systems Engineering, Virginia Tech, Blacksburg, VA, USA
33 Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA
DOI 10.15252/msb.20156157

© 2015 The Authors. Published under the terms of the CC BY 4.0 license

used (Lee *et al*, 2007; McCloskey *et al*, 2013) genome-scale models do support cellular growth in existing studies only because of numerical errors. They base these broad claims on two observations: (i) three reconstructions (*i*AF1260, *i*IT341, and *i*NJ661) compute feasibly in COBRA, but are infeasible when exact numerical algorithms are used by their software (entitled MONGOOSE); (ii) linear programs generated by MONGOOSE for *i*IT341 were submitted to the NEOS Server (a Web site that runs linear programs through various solvers) and gave inconsistent results. They further claim that a large percentage of these COBRA models are actually unable to produce biomass flux. Here, we demonstrate that the claims made by Chindelevitch *et al* (2014) stem from an incorrect parsing of models from files rather than actual problems with numerical error or COBRA computations.

Calculating numerically accurate and thermodynamically consistent flux states

To prove the feasibility of biomass production in the chosen three models, along with some others, we used the same rational solver QSopt_ex (Applegate *et al*, 2007) to compute feasible flux states. Moreover, we used SymPy, a symbolic math library (Joyner *et al*, 2012), to show that the exactly computed feasible flux state has no numerical error. Furthermore, the computed optimal growth rate from QSopt_ex matched those computed by several floating-point solvers accessed via cobrapy (CPLEX, gurobi, glpk, and MOSEK) and the COBRA toolbox (gurobi and CPLEX) to well within a precision of 10^{-6}. Using linear programming problems generated by COBRA for *i*IT341 and a version of the model we constrained to produce no biomass, we observed consistent results between COBRA and the reputable solvers hosted on the NEOS server. These results unequivocally demonstrate that these COBRA models solve consistently with both rational and floating-point solvers. We were able to extend this analysis to show 23 out of 29 models that Chindelevitch *et al* (2014) claim to be "blocked" by FBA have solutions that produce biomass flux without numerical error (Table EV1). Thus, the authors' claim that exact arithmetic is necessary for consistency and reproducibility is inaccurate, along with their findings that these previously published and computed models do not produce biomass flux.

The authors further claim that even more models are "energy blocked" and cannot produce a feasible flux state to produce biomass without thermodynamically infeasible cycles (often referred to as type III loops). Using loopless FBA (Schellenberger *et al*, 2011a), we were able to compute solutions that produce biomass without using these loops. Moreover, we demonstrate that in the case that all reactions allow 0 flux (as is the case in the MONGOOSE formulation), all solutions with loops can be converted into solutions without loops and still produce biomass. As these solutions were obtained using an existing algorithm, the inability of MONGOOSE to identify such solutions is a limitation on the method used by MONGOOSE, not on the published reconstructions as stated by Chindelevitch *et al* (2014). In total, our analysis shows that for 51 out of 59 models, the claims made by MONGOOSE about model blockage are incorrect (Table EV1).

A call for clear standards in model formulation

While the article by Chindelevitch *et al* (2014) has a valid goal of computing flux states that have been diligently checked for numerical error and thermodynamically infeasible loops, its general conclusions about the current state of COBRA models are incorrect. While more new tools to ensure model quality are welcome, conventional checks with minimal computational overhead already exist, and are routinely employed by the community of flux balance analysis users to ensure that models produce numerically accurate and thermodynamically consistent flux states. We have identified the primary source of the differences between our computations and those reported by Chindelevitch *et al* (2014) to be difficulties with parsing reconstructions from published files and their conversion into computable models. Many of the models were read from reconstructions encoded as SBML files. The mechanism of encoding COBRA model information along with a reconstruction in SBML was originally defined by the COBRA toolbox (Schellenberger *et al*, 2011b), which we therefore consider the reference implementation. For example, as a part of the SBML encoding, boundary metabolites are written with their SBML boundary condition set to true for "exchange" reactions. This convention is meant to signify a system boundary where extracellular metabolites enter and leave the system. The parser developed by Chindelevitch *et al* (2014) to read models from SBML reconstructions ignores this distinction and therefore adds additional constraints to the model. These incorrectly added constraints block any metabolites from entering the system, causing the models to give infeasible growth solutions consistent with mass balance, because mass is not entering and therefore no growth is possible. Thus, erroneous results and conclusions reported by Chindelevitch *et al* (2014) resulted from incorrect parsing of SBML files, resulting in ill-formulated models and a misinterpretation of their calculations.

Part of the issue, however, rests with difficulties associated with encoding models in a consistent format between different labs and software packages. As is the practice in the field, we contacted the authors of the models that we could not solve in order to resolve the differences; after all, the models had been used to perform COBRA computations in their respective publications. In these cases, the authors were able to supply a "fixed" SBML file after correcting errors in the SBML encoding in their respective codebases. An example of one such error was the presence of both "CO2" and "co2" as metabolites in the SBML file for *i*VS941 (Satish Kumar *et al*, 2011). While the GAMS software used in simulating that model is case-insensitive and correctly creates one constraint, parsing the file in other packages (such as the COBRA toolbox, cobrapy, and MONGOOSE) incorrectly created two separate constraints for the uppercase and lowercase versions. Therefore, an inadvertent error in a file-encoding led to different mathematical models in different software tools, and working with the authors of the original model was necessary to resolve the differences. Out of the 88 models attempted by Chindelevitch *et al* (2014), we were able to solve 80, and 9 of these required modifications to fix encoding errors. We attempted to parse 6 of the remaining 8 reconstructions. While the models we parsed from these reconstructions did not solve, this result was still consistent between floating-point and exact solvers.

 © 2015 The Authors

This situation is a symptom of the well-known issue with interoperability of reconstructions between different laboratories and software packages in constraint-based modeling (Ravikrishnan & Raman, 2015). We believe we can improve upon these issues by better adhering to the standard practices of openness and reproducibility (Dräger & Palsson, 2014). We believe the community needs to standardize on the most recent version of the flux balance constraints (fbc) extension to SBML as the single well-specified format to reliably encode reconstructions, as strict use of fbc version 2 was specifically designed to build genome-scale models unambiguously [SBML-flux Working Group, 2014 SBML Flux Balance Constraints (fbc), http://sbml.org/Documents/Specifications/SBML_Level_3/Packages/Flux_Balance_Constraints_(flux) (Accessed June 13, 2015)]. Therefore, we propose that new reconstructions be published as validated SBML+fbc files and that the authors of existing reconstructions convert them into this format. Moreover, in the interests of reproducibility, studies including flux balance analysis on these genome-scale models should strive to make their code easily reproducible. The models and code used in this study are available as Dataset EV1 and also at https://github.com/opencobra/m_model_collection.

Expanded View for this article is available online: http://msb.embopress.org

Acknowledgements
We thank Leonid Chindelevitch for extensive discussions and for sharing results obtained with the MONGOOSE platform for comparison with solutions obtained with COBRA software.

Author contributions
AE wrote the code and assembled the models included in Dataset EV1. All of the authors contributed to the design, approach, and written manuscript. Subsequent authors are arranged alphabetically by last name.

References
Applegate DL, Cook W, Dash S, Espinoza DG (2007) Exact solutions to linear programming problems. *Oper Res Lett* 35: 693 – 699

Bordbar A, Monk JM, King ZA, Palsson BO (2014) Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* 15: 107 – 120

Chindelevitch L, Trigg J, Regev A, Berger B (2014) An exact arithmetic toolbox for a consistent and reproducible structural analysis of metabolic network models. *Nat Commun* 5: 4893

Dräger A, Palsson BØ (2014) Improving collaboration by standardization efforts in systems biology. *Front Bioeng Biotechnol* 2: 61

Joyner D, Čertík O, Meurer A, Granger BE (2012) Open source computer algebra systems: SymPy. *ACM Commun Comput Algebra* 45: 225 – 234

Lee KH, Park JH, Kim TY, Kim HU, Lee SY (2007) Systems metabolic engineering of *Escherichia coli* for L-threonine production. *Mol Syst Biol* 3: 149

McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Mol Syst Biol* 9: 661

Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? *Nat Biotechnol* 28: 245 – 248

Ravikrishnan A, Raman K (2015) Critical assessment of genome-scale metabolic networks: the need for a unified standard. *Brief Bioinform* doi: 10.1093/bib/bbv003

Satish Kumar V, Ferry JG, Maranas CD (2011) Metabolic reconstruction of the archaeon methanogen Methanosarcina Acetivorans. *BMC Syst Biol* 5: 28

Schellenberger J, Lewis NE, Palsson BØ (2011a) Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophys J* 100: 544 – 553

Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR, Palsson BØ (2011b) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* 6: 1290 – 1307

License: This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.