

Application of string similarity ratio and edit distance in automatic metabolite reconciliation comparing reconstructions and models

Martins Mednis^{1*}, Maike K. Aurich²

¹Biosystems group, Department of Computer Systems, Latvia University of Agriculture, Liela iela 2, LV3001, Jelgava, Latvia

²Center for Systems Biology, University of Iceland, Sturlugata 8, IS101, Reykjavik, Iceland

*Corresponding author

Martins.mednis@llu.lv

Received: 14 November 2012; accepted: 27 November 2012; published online: 28 November 2012.

This paper has supplementary material.

Abstract: Increasing numbers of biochemical network models become available and reuse of these models is becoming more common. As a consequence, tools to compare models are needed. Comparison can be difficult because model builders often use different standards during reconstruction, metabolite formulas are not always indicated, IDs and names of metabolites are different, and models are stored in different formats (SBML, COBRA and others).

Herein, a model comparison algorithm for SBML and COBRA format models is presented, called ModeRator. Precondition for correct matching of reactions is the comparison of the participating metabolites.

ModeRator is based on the comparison of metabolite names as text strings. An automatic three level filtering approach is implemented in the software, which rejects pairs of potentially equal metabolites and builds an opinion about metabolite pairs with high similarity in metabolite names.

ModeRator was applied to two test cases, comparing two models of each, *E.coli* and *S.cerevisiae*. Matches of the automatic mapping were manually inspected and compared with the automatic predictions.

Automatic metabolite mapping of *E.coli* models (1314 and 1704 metabolites) comparing only identifiers revealed a high number of accordant metabolites. Both models originate from the same source (BioCyc database). No significant difference between automatic mapping and manual curation are observed.

For the comparison of two *S.cerevisiae* models (679 and 1061 metabolites), three level filtration by metabolite name is used. The discrepancy between manual curated predictions and ModeRator predictions was 7%.

Keywords: Biochemical networks, reconstruction, models, metabolite mapping, pairwise comparison.

1. Introduction

The function of cells is based on complex networks of interacting chemical reactions carefully organized in space and time. These biochemical reaction networks produce observable cellular functions (Palsson, 2006). Reconstruction of a biochemical network is the assembly of the components and their interconversions that appear within an organism, based on the genome annotation and the bibliome (Lewis et al., 2009). Reconstruction based models can be used for the analysis of network capabilities, prediction of cellular phenotypes and in silico hypothesis generation (Lewis et al., 2009).

The first fully sequenced genome was that of *H.influenzae* in 1995 (Fleischmann et al., 1995), which enabled the first reconstruction of a genome-scale metabolic network in 1999 (Edwards and Palsson, 1999). Thanks to the accelerating speed of sequencing techniques, it is now possible to reconstruct the network of biochemical reactions in many organisms. The reconstruction process for metabolic networks is well developed (Schellenberger et al., 2011; Thiele and Palsson, 2010) and implemented for a number of different organisms (Palsson, 2006; Schellenberger et al., 2010), including human (Duarte et al., 2007). Several of these networks are available online: Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Kanehisa et al., 2011), EcoCyc

(Keseler et al., 2011), BioCyc (Karp et al., 2005; Karp et al., 2010) and metaTIGER (Whitaker et al., 2009).

The available reconstructions and models are growing both in number and size (number of interactions within reconstruction or model). Often multiple reconstructions exist for an organism. The need to compare models or to couple them as parts of larger models has been noted by Radulescu (Radulescu et al., 2008). Manual comparison is very time-consuming especially in case of genome-scale reconstructions consisting of thousands of reactions and metabolites. The demand for a method to relate different models has been pointed out elsewhere (Gay et al., 2010).

Genome-size reconstructions are available in different formats. Mostly they are in form of 1) plain text files, 2) SBML (Systems Biology Markup Language (Hucka et al., 2003)) file format and 3) spreadsheets (including COBRA format (Schellenberger et al., 2011)).

Due to different formats of models and different approaches to standardization of metabolites and reactions, the comparison of reconstructions and models is complicated. Tools exist that enable visual comparison of the model structure, (Boele et al., 2012; Kostromins and Stalidzans, 2012; Schellenberger et al., 2010) or to compare parameters of the structure (Rubina and Stalidzans, 2010; Yamada and Bork, 2009).

Additionally, there are several software tools with functionality that is more or less related to a comparison of models. These include: Tools-4-Metatool (Xavier et al., 2011), Compare Subsystems (Oberhardt et al., 2011), SemanticSBML (Krause et al., 2010), COBRA (Becker et al., 2007), FAME (Boele et al., 2012) and MetRxn (Kumar et al., 2012). However, tools are needed to compare metabolites and reactions such that the scope of the models can be compared. ModeRator (Mednis et al., 2012) is a software explicitly developed for comparison of models.

The algorithm analyses the similarity of chemical formulas and names of metabolites. In case of identical formulas and names the metabolites are automatically considered to be identical. In the remaining cases a similarity rate is calculated and manual curation is needed. The algorithm is tested comparing pairs of E.coli and yeast (or S.cerevisiae) models.

2. Materials and methods

2.1. Comparison criteria

In order to decide whether two reactions are the same or different, one has to compare the involved metabolites. Mapping of metabolites means explicitly defining that metabolite “abc” from one reconstruction is the same as metabolite “xyz” from the other reconstruction. Technically, it can be achieved by assigning the same ID to two metabolites that are believed to be the same.

In case the two reconstructions to be compared come from different sources or authors, it is likely that the elements in both reconstructions do not share common identifiers. In other words, elements, like compartments, metabolites and reactions will have different IDs. In SBML files there is a unique id value for each element, but in COBRA models abbreviations serve as unique identifiers.

The problem is to evaluate and decide if two elements with different IDs are the same or not. If the IDs cannot be used for identifying an element, another property must be used. For metabolites the chemical formula is an alternative.

Chemical metabolite formulas are mandatory in COBRA models and enable some of the core functionalities of COBRA. They are stored as clear text (such as H₂O).

In comparison, inclusion of chemical formulas is not common practice in SBML models. Even more -- the SBML format does not support chemical formulas in clear text.

Another disadvantage for using chemical formulas as identifiers for metabolites is the existence of isomers, i.e. molecules with the same chemical formula but with different

spatial structures. Well known examples of isomers are glucose and fructose.

Another option is to compare elements, e. g. metabolites by names. Since the terminology is not standardized, it is unlikely that different authors will name elements absolutely identical. Fuzzy string comparison algorithms can be used to compare strings. Such an algorithm calculates the similarity of two given text strings - metabolite names in our case.

Various implementations for calculating a similarity ratio (Levenshtein, 1966; Ratcliff and Metzener, 1988) between two text strings are available.

A ratio usually is a floating point number ranging from 0.0 to 1.0 indicating similarity of two given sequences (or strings). The ratio “0” means that two strings have nothing in common. For example such strings would be “ABCD” and “EFGH”. These particular strings do not share a single common character. The ratio “1.0” means that two given strings are absolutely identical. The closer to 1.0, the more similar two given strings are.

The distance between two strings is the number of steps needed to transform string A into string B. The distance sometimes is also called edit distance. For example the distance between “abcd” and “aZcZ” is 2. That's because there are only two edits needed to change “abcd” into “aZcZ” -- two replace operations.

The threshold is a measurable value that serves the purpose of filtering elements by a value of its property. For instance, threshold of 5, means that any value that is below (or above) 5 is filtered and not passed further in the algorithm. The threshold of string similarity ratio (and/or distance) means that all pairs of strings that are not similar enough, are filtered out.

The Table 1 shows examples of ratios and distances for different string pairs. Given examples show an interesting trend: string length has an impact on ratio. On longer strings the impact is smaller, but on shorter strings the impact is higher.

The authors of SBMLmerge software (Schulz et al., 2006) are using only ratios (Ratcliff and Metzener, 1988) without taking into account edit distances to map metabolite names. As it is shown on Table 1 this approach works well only with long metabolite names. In order to find two similar names that are short in length, the threshold of ratio needs to be lowered. However, lowering the threshold increases the risk of false positives. This means that on low threshold two names can be reported as similar, although they are different metabolites, like “D-glutamate” and “L-glutamate”

Table 1

Example of Levenshtein ratio and edit distance variations for different strings.

String A	String B	Ratio	Distance
Bicarbonate	bicarbonate	0.9	1
Glucose-6-phosphate	Glucose six phosphate	0.8	5
Glucose-6-phosphate	Glucose-six-phosphate	0.9	3
L-tryptophanyl-tRNA ^{trp}	L-Tryptophanyl-tRNA(trp)	0.86	4
L-lysine	L-Lysine	0.87	1
D-glutamate	L-Glutamate	0.81	2
abcd	aZcZ	0.5	2

2.2. Three level filtering

The comparison of metabolite names of large models can relate several metabolites of model M1 to one metabolite of model M2. The second level filters out multiple links to the model M1 (Fig. 1). The third level filters out multiple links to

the model M2. During first level filtering the strength of a link between all possible metabolite pairs between model M1 and M2 is calculated. During the third level filtering the relations “many to one” are compared and the pair with the highest similarity is nominated as mapped metabolites.

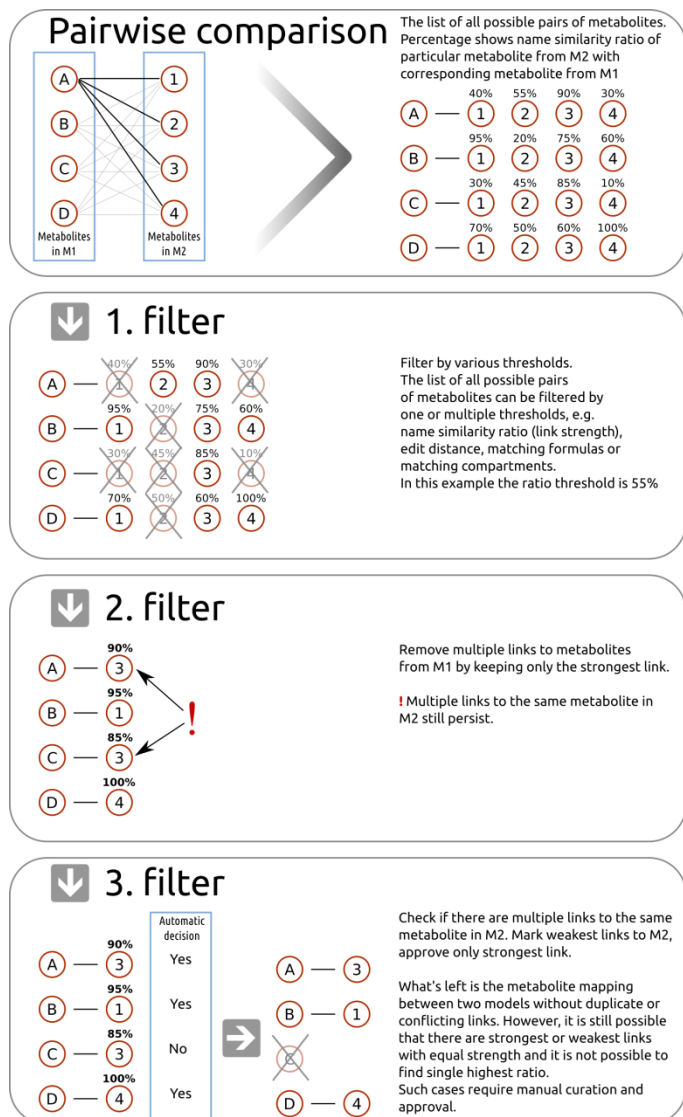


Fig. 1. Schematic example of three level filtering approach. Metabolites of model M1 are marked by letters and metabolites of model M2 are marked by numbers. Name similarity ratio (strength of links) is measured in percents.

The above mentioned approach is implemented in the most recent (2.5.4) version of ModeRator - previously published software tool for model comparison (Mednis et al., 2012).

Two pairs of models were compared: Escherichia coli models “ecol199310cyc” and “ecol316407cyc” from BioCyc database (<http://www.biocyc.org/>) and Saccharomyces cerevisiae models iND750 developed by Natalie Duarte (Duarte et al., 2004) and iLL672 developed by Lars Kuepfer (Kuepfer et al., 2005).

3. Results and discussion

3.1. Comparison of E.coli models

E.coli models contain 1314 and 1704 metabolites (see the supplementary file “ecoli_metabolites”). Metabolites of the E.coli models were compared by their identifiers. Therefore filtering of multiple links was not applicable. Since E.coli models came from the same source, we expected to find equal annotation for same metabolites.

Automatic comparison using ModeRator revealed 1094 common metabolites with matching identifiers. Interestingly, not all metabolite pairs with matching IDs also had equal

metabolite names and chemical formulas. In fact, 31 of 1090 metabolite pairs had different names. Additionally, 30 pairs had different chemical formulas. For 391 metabolite pairs it was not possible to compare chemical formulas, because one or both formulas were missing.

Manual curation of the above mentioned 31 pairs with non equal names was performed using databases (Ecocyc (Keseler et al., 2011), Metacyc (Caspi et al., 2012), and E. coli Metabolome database (Guo et al., 2012)). In total, 30 metabolite pairs were approved during manual curation. One pair was left without decision.

3.2. Comparison of S.cerevisiae models

The reconstruction of Yeast metabolism iLL672 (Kuepfer et al., 2005) is based on the previous reconstruction iFF708 (Förster et al., 2003). iFF708 covered two main compartments, cytosol and mitochondria. Metabolites located in the mitochondria for a specific reaction terminate with an „m“ (Förster et al., 2003). This differentiation between cytosolic and mitochondrial metabolites persists in iLL672 (Kuepfer et al., 2005). Metabolite localisation was ignored during manual comparison (strings were matched ignoring the terminal „m“). The compared S. cerevisiae models contain 679 and 1061 metabolites.

Thresholds for name similarity ratio and edit distance were chosen low enough to minimize false negatives when two identical metabolites with not-so-identical names are filtered from results. The threshold for similarity ratio was 68% and the threshold for edit distance was 15 edits.

Automatic comparison by identifiers revealed only one metabolite - “Acy-carrier protein”, so we had to compare metabolites by names. From total 447 returned matches (pairs), 248 pairs with identical metabolite names (ratio 100%) were automatically approved. 199 pairs received a non 100% ratio. During third level filtering 152 of them were automatically approved (considered as mapped metabolites being equal) and 47 automatically disapproved (considered as being different metabolites).

Manual curation was only based on metabolite names. For the manual comparison of iLL672 and iND750, BIGG database (Schellenberger et al., 2010) was used. It captures detailed information about iND750 metabolites. Using this resource, many metabolite matches could be either confirmed or verified. In case the match could not be resolved, we additionally queried the Yeast metabolome database (YMDB, <http://www.ymdb.ca/>) (Jewison et al., 2012)). Using synonyms listed in YMDB, we again queried BIGG database and derived a number of alternative matches of iLL672 metabolites to iND750 metabolites (alternative matches provided in the column „Comments” in supplementary file “yeast_metabolites”).

In case a typing error was likely, match was confirmed and a note was made in the column „Comments” (See supplementary materials file “yeast_metabolites”).

199 pairs with non 100% match ($68 \leq \text{ratio} (\%) < 100$) were curated. Manual curation and automatic predictions agreed in 133 cases, 64 were different and in two cases decision could not be made based on the provided information.

Many differences were found through excluding or including „“ , “-”, or „spaces“, and terms of stereoisomerism into metabolite names.

Table 2

Manual curation vs. automatic comparison. „Disagree on SAME“ means that manual curation decision is „SAME“, but ModeRator decides otherwise.

Manual curation vs. automatic comparison	Cases	Comment
Identical names	250	Manual curation was not necessary
Agree on decision „SAME metabolites“	124	Manual and automatic curation decide that metabolites are the same
Agree on DIFFERENT	42	Manual and automatic curation decide that metabolites are different
Disagree on SAME	9	Manual curation decides that metabolites are the same
Disagree on DIFFERENT	22	Manual curation decides that metabolites are different

4. Conclusion

Different formats and description standards of models for formulas (charged/uncharged), names and abbreviations of metabolites complicate the comparison of reconstructions and models.

The in ModeRator implemented three level filtering approach can perform automatic metabolite matching as part of a model comparison task. The three level filtering serves as decision support system that can accelerate subsequent manual efforts. The algorithm can find identical metabolites which are declared as mapped metabolites without manual curation. It also filters out metabolite pairs with low similarity (levels of similarity thresholds can be adjusted) decreasing the number of metabolites that have to be manually inspected and thus decreases the time needed for manual curation.

The two demonstrated model comparisons (E.coli and S.cerevisiae) demonstrate that the success rate strongly depends on similarity of metabolite descriptions of the compared models. For two models derived from the same source, automatic metabolite matching demonstrates very good performance and manual curation may be needed just for few metabolites.

Acknowledgments

This work is funded by a project of European Structural Fund Nr. 2009/0207/1DP/1.1.1.2.0/09/APIA/VIAA/128 “Latvian Interdisciplinary Interuniversity Scientific Group of Systems Biology” www.sysbio.lv.

References

- Becker, S. a, Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., and Herrgard, M. J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature protocols*, 2(3), 727–38. <http://dx.doi.org/10.1038/nprot.2007.99>
- Boele, J., Olivier, B. G., and Teusink, B. (2012). FAME, the Flux Analysis and Modeling Environment. *BMC systems biology*, 6(1), 8. <http://dx.doi.org/10.1186/1752-0509-6-8>
- Caspi, R., Altman, T., Dreher, K., Fulcher, C. a, Subhraveti, P., Keseler, I. M., Kothari, A., et al. (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research*, 40(Database issue), D742–53. <http://dx.doi.org/10.1093/nar/gkr1014>
- Duarte, N. C., Becker, S. a, Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., et al. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6), 1777–82. <http://dx.doi.org/10.1073/pnas.0610772104>
- Duarte, N. C., Herrgård, M. J., and Palsson, B. Ø. (2004). Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome research*, 14(7), 1298–309. <http://dx.doi.org/10.1101/gr.2250904>
- Edwards, J. S., and Palsson, B. O. (1999). Systems properties of the Haemophilus influenzae Rd metabolic genotype. *The Journal of Biological Chemistry*, 274(25), 17410–17416. <http://dx.doi.org/10.1074/jbc.274.25.17410>
- Finney, A., and Hucka, M. (2003). Systems biology markup language: Level 2 and beyond. *Biochemical Society Transactions*, 31(Pt 6), 1472–1473. <http://dx.doi.org/10.1042/BST0311472>
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223), 496–512. <http://dx.doi.org/10.1126/science.7542800>
- Förster, J., Famili, I., Fu, P., Palsson, B. Ø., and Nielsen, J. (2003). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome research*, 13(2), 244–53. <http://dx.doi.org/10.1101/gr.234503>
- Gay, S., Soliman, S., and Fages, F. (2010). A graphical method for reducing and relating models in systems biology. *Bioinformatics*, 26(18), i575–i581. <http://dx.doi.org/10.1093/bioinformatics/btq388>
- Guo, a. C., Jewison, T., Wilson, M., Liu, Y., Knox, C., Djoumbou, Y., Lo, P., et al. (2012). ECMDDB: The E. coli Metabolome Database. *Nucleic Acids Research*, 1(10), 1–6. <http://dx.doi.org/10.1093/nar/gks992>
- Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, a. P., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4), 524–531. <http://dx.doi.org/10.1093/bioinformatics/btg015>
- Jewison, T., Knox, C., Neveu, V., Djoumbou, Y., Guo, A. C., Lee, J., Liu, P., et al. (2012). YMDB: the Yeast Metabolome Database. *Nucleic acids research*, 40(Database issue), D815–20. <http://dx.doi.org/10.1093/nar/gkr916>
- Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1), 27–30. <http://dx.doi.org/10.1093/nar/28.1.27>
- Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(Database issue), D109–14. <http://dx.doi.org/10.1093/nar/gkr988>
- Karp, P. D., Ouzounis, C. a, Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., et al. (2005). Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic acids research*, 33(19), 6083–6089. <http://dx.doi.org/10.1093/nar/gki892>
- Karp, P. D., Paley, S. M., Kruppenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., et al. (2010). Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, 11(1), 40–79. <http://dx.doi.org/10.1093/bib/bbp043>
- Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muñiz-Rascado, L., Bonavides-Martinez, C., et al. (2011). EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic acids research*, 39(Database issue), D583–90. <http://dx.doi.org/10.1093/nar/gkq1143>
- Kostromins, A., and Stalidzans, E. (2012). Paint4Net: COBRA Toolbox extension for visualization of stoichiometric models of metabolism. *Biosystems*. <http://dx.doi.org/10.1016/j.biosystems.2012.03.002>
- Krause, F., Uhlendorf, J., Lubitz, T., Schulz, M., Klipp, E., and Liebermeister, W. (2010). Annotation and merging of SBML models with semanticSBML. *Bioinformatics (Oxford, England)*, 26(3), 421–2. <http://dx.doi.org/10.1093/bioinformatics/btp642>
- Kuepfer, L., Sauer, U., and Blank, L. M. (2005). Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome research*, 15(10), 1421–30. <http://dx.doi.org/10.1101/gr.3992505>
- Kumar, A., Suthers, P. F., and Maranas, C. D. (2012). MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC bioinformatics*, 13(1), 6. <http://dx.doi.org/10.1186/1471-2105-13-6>
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady*, 10(8), 707–710. Retrieved from <http://profs.sci.univr.it/~liptak/ALBioinfo/files/levenshtein66.pdf>
- Lewis, N. E., Jamshidi, N., Thiele, I., and Palsson, B. Ø. (2009). Metabolic systems biology: a constraint-based approach. In R. Meyers (Ed.), *Encyclopedia of Complexity and Systems Science* (p. 5535). New York: Springer.

- Mednis, M., Rove, Z., and Galvanauskas, V. (2012). ModeRator - a software tool for comparison of stoichiometric models. 7th IEEE International Symposium on Applied Computational Intelligence and Informatics (pp. 97–100). <http://dx.doi.org/10.1109/SACI.2012.6249983>
- Oberhardt, M. A., Puchalka, J., Martins Dos Santos, V. A. P., and Papin, J. A. (2011). Reconciliation of Genome-Scale Metabolic Reconstructions for Comparative Systems Analysis. (P. E. Bourne, Ed.) PLoS Computational Biology, 7(3), 18. <http://dx.doi.org/10.1371/journal.pcbi.1001116>
- Palsson, B. Ø. (2006). Systems Biology: Properties of reconstructed networks. Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511790515>
- Radulescu, O., Gorban, A. N., Zinovyev, A., and Lilienbaum, A. (2008). Robust simplifications of multiscale biochemical networks. BMC systems biology, 2(1), 86. <http://dx.doi.org/10.1186/1752-0509-2-86>
- Ratcliff, J. W., and Metzener, D. (1988). Pattern Matching: The Gestalt Approach. Dr Dobbs Journal, (July), 46.
- Rubina, T., and Stalidzans, E. (2010). Topological features and parameters of Biochemical Network Structures. International Industrial Simulation Conference (pp. 228–236). Budapest: EUROSIS.
- Schellenberger, J., Park, J. O., Conrad, T. M., and Palsson, B. Ø. (2010). BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC bioinformatics, 11, 213. <http://dx.doi.org/10.1186/1471-2105-11-213>
- Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature Protocols, 6(9), 1290–1307. <http://dx.doi.org/10.1038/nprot.2011.308>
- Schulz, M., Uhlendorf, J., Klipp, E., and Liebermeister, W. (2006). SBMLmerge, a system for combining biochemical network models. Genome informatics International Conference on Genome Informatics, 17(1), 62–71.
- Thiele, I., and Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols, 5(1), 93–121. <http://dx.doi.org/10.1038/nprot.2009.203>
- Whitaker, J. W., Letunic, I., McConkey, G. A., and Westhead, D. R. (2009). metaTIGER: a metabolic evolution resource. Nucleic Acids Research, 37(Database issue), D531–D538. <http://dx.doi.org/10.1093/nar/gkn826>
- Xavier, D., Vázquez, S., Higuera, C., Morán, F., and Montero, F. (2011). Tools-4-Metatool (T4M): Online suite of web-tools to process stoichiometric network analysis data from METATOOL. Biosystems, 1–4. <http://dx.doi.org/10.1016/j.biosystems.2011.04.004>
- Yamada, T., and Bork, P. (2009). Evolution of biomolecular networks: lessons from metabolic and protein interactions. Nature reviews. Molecular cell biology, 10(11), 791–803. <http://dx.doi.org/10.1038/nrm2787>