Biosystems and Information Technology (2013) Vol.2(1) 1-5
DOI: http://dx.doi.org/10.11592/bit.130501

original scientific article

Automatic comparison of metabolites names: impact of criteria thresholds

Martins Mednis'[1*], Armands Vigants[2]

[1]*Biosystems Group, Department of Computer Systems, Latvia University of Agriculture, Liela iela 2, LV-3001, Jelgava, Latvia*
[2]*Institute of Microbiology and Biotechnology, Faculty of Biology, Latvia University, Kronvalda bulv. 4, LV-1586, Riga, Latvia*
**Corresponding author*
martins.mednis@llu.lv

Received: 15 May 2013; accepted: 29 May 2013; published online: 29 May 2013.
This paper has supplementary material.

Abstract: *The growing number of stoichiometric reconstructions and models tends to change the model building process. Instead of creating a new model from scratch scientists can look at the earlier created relevant models to assess the opinion and consensus level of other modellers. Several initiatives have been performed to build consensus models for particular organisms following this approach. One of possible improvements in the model development taking into account earlier developed ones is automated comparison of models. That is enabled by the fact that models usually are in a computer readable format to be simulated. Still there are some problems like different ways of naming metabolites, models without formulae of metabolites, different approach in definition of compartments and other peculiarities of different research groups at different times.*

There are several software tools that offer reconciliation or mapping of metabolites in models to assess the similarity of models. It is computationally trivial to find metabolite pairs with identical names in two models. Still very often the comparison algorithms lack flexibility and some work should be done by manual curation of metabolite pairs to recognize that the difference is caused by symbols like brackets, quotes, apostrophes, spaces, upper/lower case letters or similar ones.

The proposed approach suggests combination of automated comparison with manual curation: most of possible metabolite pairs are rejected by computer leaving just the most similar metabolite pairs for manual comparison. The elasticity in metabolite name comparison is introduced using Levenstein similarity ratio and Levenstein edit distance. Application of these criteria with different acceptance thresholds is analyzed comparing two models of Saccaromyces cerevisiae with 681 and 1063 metabolites. The results are compared with manually approved pairs of matching metabolites.

Keywords: fuzzy string comparison, edit distance, similarity ratio, metabolic networks.

1. Introduction

The molecular processes in cells form a huge network, which makes detailed mathematical modeling and simulation extremely difficult (Schulz et al., 2006). Genome-scale reconstructions of metabolic networks and stoichiometric models may contain thousands of metabolites and reactions (Thiele et al., 2013). The functions of such networks are hard for the human mind to comprehend (Palsson, 2006). The process of iterative model building (Thiele and Palsson, 2010) promises to accelerate biological discovery, product development, and process design (Ideker et al., 2001; Palsson, 2006). The increasing knowledge base of living organisms leads to even more complex biochemical models and scientists often decide to model only a part of genome, not the whole metabolism (Mednis and Aurich, 2012).

It is wise to look around and check what other models of that particular organism exists before creating a new model of an organism or it's part. If more than one model is available, it is important to evaluate them and choose as starting point the most comprehensive model, the one with least inconsistencies, intersection or merge of models. The published genome-scale reconstructions of the same organism should be carefully compared to avoid misleading conclusions. Consequently, the need for analysis, comparison, intersection and merge of biomodels is growing. The demand for a method to relate different models has been pointed out (Gay et al., 2010; Radulescu et al., 2008).

Metabolites could be compared by chemical formula and name. Due to identical formulas in case of isomers (Poggendorff, 1830) and the lack of formula in many models the comparison of metabolite names becomes irreplaceable. Still chemical formulas also can serve as additional criterion during comparison of models. If the formulas are available, it is possible to check if they are equal for the particular pair of metabolites. Still the main criterion remains metabolite name.

Comparison is an essential procedure before merging or intersecting two or more models. Some model comparison related functionality is proposed by existing software tools. The COBRA toolbox (Schellenberger et al., 2011) has two functions related to the search for duplicates and the comparison of two models. *CheckCobraModelUnique()* finds reactions and metabolites that are not unique. The second function *isSameCobraModel()* receives two models as input parameters and returns three outputs: *isSame* - "true" if all common fields are identical, otherwise "false"; *nDiff* - number of differences between the two models for each field.

The FAME (*Flux Analysis and Modeling Environment)* (Boele et al., 2012), is a browser-based graphical interface that allows users to build, edit, run, and visualize stoichiometric models. The Fame also has comparison and merge functionality. The merge facility can be used to merge an additional pathway into an existing model. When merging, one model must be designated the *master* model, and the other the *slave model*. This is to ensure that whenever a merging conflict exists, e.g. when a reaction exists in both models but with different constraints, the information in the master model will take precedence.

If reactions have the same reaction ID in both models (this is usually the case for identical reactions if both models are generated by the same source, such as FAME), they are assumed identical will be included only once. However, FAME has no way of knowing whether reactions are identical if neither reaction IDs nor species IDs follow the same convention - if this is the case, both versions of the same reaction will be included. If a reaction has been deleted from the master model, but exists in the slave model, it will be in the result of the merge operation (existing reactions overwrite deleted ones, as no record of deletions is kept).

The above described tools do not tolerate even small differences in metabolite names like brackets, quotes, apostrophes, spaces, upper/lower case letters and some more symbols which may be caused by the modelers style of defining metabolites. Therefore many pairs of identical metabolites may not be recognized leading to wrong conclusions about the similarity of models. Some other tools compare models in a more flexible and adaptive way.

Model SBMLmerge feature is provided in SemanticSBML (Krause et al., 2010). SBMLmerge (Schulz et al., 2006) first merges the lists of elements in the annotated input files. The resulting list is then searched for conflicting elements by pairwise comparison, based on the identifying attributes, including the annotations. If two conflicting elements are found, their describing attributes are compared. The values for these attributes can be identical for both conflicting elements, or they can differ in one or more values: if all attribute values are identical, the elements are assumed to have the same biological meaning.

If both networks (models) come from the same source or both models contain information about metabolites chemical formula, it is possible to skip reconciliation of metabolites. An approach of reaction comparison based on comparison of metabolites formulas has been described by Mednis (Mednis et al., 2012; Mednis et al., 2012). However, such approach is not suitable when metabolite chemical formulas are not available.

A software tool ModeRator (Mednis et al., 2012) for biochemical network comparison can perform metabolite mapping between two models based on name similarity and other criteria as well. The software tool also introduces three-level-filtering algorithm (Mednis and Aurich, 2012) which significantly reduces the amount of data that requires manual curation.

This article is devoted to the efficiency analysis of additional comparison criteria in application of mapping metabolite names. The impact of filtering by similarity ratio and edit distance, metabolite compartments and Three-level-filtering is discussed. Two *Saccharomyces cerevisiae* (Baker's yeast) models are compared to determine how the thresholds of name similarity criteria impact the number of manually approvable pairs.

2. Materials and methods

2.1. Pairwise comparison

The purpose of metabolite *mapping* or *reconciliation* (Oberhardt et al., 2011) is to find a corresponding metabolite in Network B for each metabolite from Network A (Fig. 1).

Comparing two lists of metabolites can be very laborious since it involves the screening of all possible element combinations. Therefore software automatically rejects invalid combinations leaving only the most similar pairs of metabolites for manual curation.

The algorithm of metabolites reconciliation starts with creation of Cartesian product from both lists of metabolites. The result is another list containing all possible pairs of metabolites (Fig. 1). The next step is to calculate name similarities for all metabolite pairs in this list.

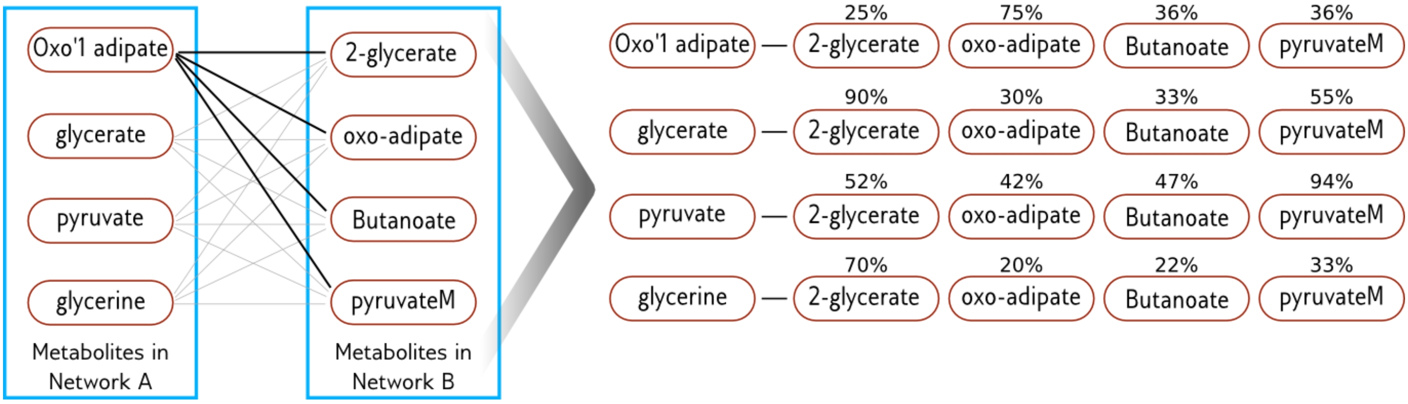
Oxo'1 adipate
glycerate
pyruvate
glycerine
Metabolites in Network A
2-glycerate
oxo-adipate
Butanoate
pyruvateM
Metabolites in Network B
Oxo'1 adipate — 2-glycerate 25% — oxo-adipate 75% — Butanoate 36% — pyruvateM 36%
glycerate — 2-glycerate 90% — oxo-adipate 30% — Butanoate 33% — pyruvateM 55%
pyruvate — 2-glycerate 52% — oxo-adipate 42% — Butanoate 47% — pyruvateM 94%
glycerine — 2-glycerate 70% — oxo-adipate 20% — Butanoate 22% — pyruvateM 33%

Fig. 1. **Pairwise comparison before Three-level-filtering.**

1st filter

The 1st filter is a set of user defined thresholds for various criteria. In the example discussed in this section, the criterion is name similarity ratio. We used Levenstein edit distance (Levenshtein, 1966) and similarity ratio implementation in pylevenstein (Mulligan, 2013). The threshold in this example is 55%. It means that each pair where metabolites names similarity ratio is below the threshold will be automatically discarded leaving only the most relevant results (Fig. 2). The edit distance can be used as a criterion. In such case, the algorithm automatically discards pairs with edit distance above the threshold.

The user can define also the compartment threshold. Filtering *by compartments* takes into account the information about metabolites compartments and will automatically discard those metabolite pairs where the information about their compartments is conflicting. In a real-world example a pair of *glucose[cytosol]* and *glucose[extracellular]* would be discarded, but a pair with *L-lysine[cytosol]* and *L-lysine[cytosol]* would be spared because the compartments are matching.

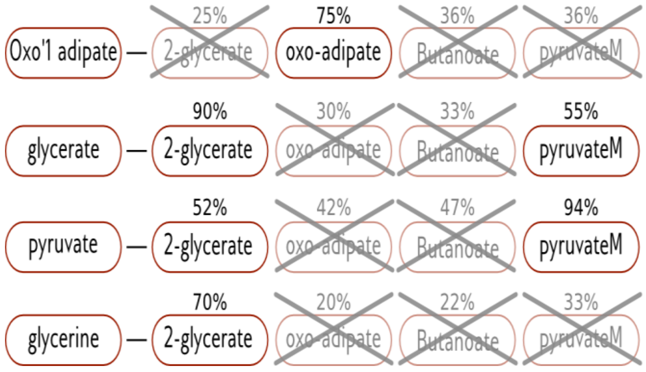

Fig. 2. **Three-level-filtering: 1st filfer.**

2nd filter

The 2nd filter discards pairs with least similarity. It is possible that many pairs with the same metabolite have similarity higher than the threshold. In Fig. 3 the metabolite “glycerate” has *connection* with metabolite “2-glycerate” and “pyruvateM”. However, the similarity with metabolite “2-glycerate” is higher and therefore all other *connections* to metabolite “glycerate” are discarded. The algorithm iterates through all *connections* of each metabolite from Network A and discards pairs with similarity less than highest.

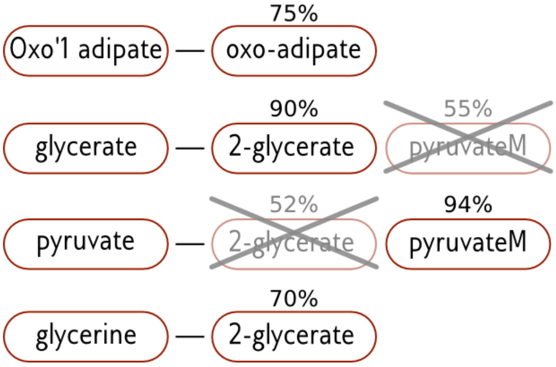

Fig. 3. **Three-level-filtering: 2nd filfer.**

3rd filter

The 3rd filter also discards pairs with least similarity. While the 2nd filter eliminates multiple connections to the same metabolite in Network A (Fig. 3), the 3rd filter does the same thing, but in opposite direction. It eliminates multiple connections to the same metabolite in Network B by discarding pairs with similarity ratio less than highest (Fig. 4).

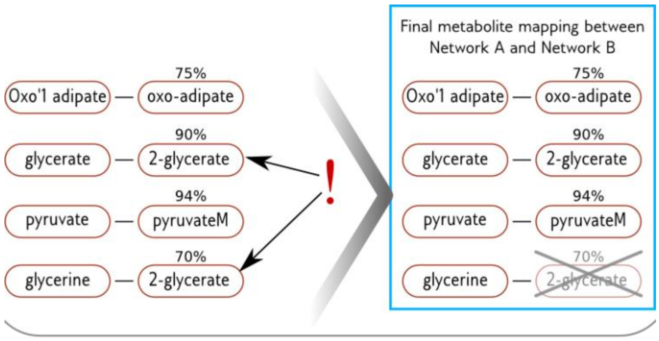

Fig. 4. **Three-level-filtering: 3rd filfer.**

2.2. Application of thresholds of different criteria

The threshold for similarity between two metabolite names is a number (set by the user) ranging from 0 to 100%. Examples of similar names and their similarity ratios are listed in Table 1.

Table 1.

Example of similarity ratio and edit distance variations for different strings

String A	String B	Similarity ratio	Edit 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-tRNAtrp	L-Tryptophanyl-tRNA(trp)	0.86	4
L-lysine	L-Lysine	0.87	1
D-glutamate	L-Glutamate	0.81	2

While the *three-level-filtering* algorithm does not solve the problem of fully automatic comparison, it can drastically reduce the amount of data that requires manual approval (Mednis and Aurich, 2012).

3. Results and discussion

In the experiments described in this article two *Saccharomyces cerevisiae* (Baker's yeast) models were compared. The models having 904 and 1268 reactions are based on iND750 (Duarte et al., 2004) and iLL672 (Kuepfer et al., 2005). Both models have located metabolites and reactions in compartments. In all experiments similarity ratio threshold was decreased step-wise from 100% to 5% with the step size 5% (except where it is noted).

The total number of metabolite pairs to process is 723903 and is formed by a multiplication of metabolite numbers in both models – 1063 and 681. In this case the maximal possible number of valid pairs (MPNVP) can not exceed 681 as that is the number of metabolites in the smallest model. The effect of all the three filters at different similarity ratio thresholds (Fig. 5) indicate almost linear increase of filter 1 passing metabolite pairs. Threshold “0” means that 1st filter is not operating at all. Therefore the curve reaches the total number of metabolite pairs when threshold is low. The curves of 2nd and 3rd filter stop growing when the threshold is below 40%. Thus the 1st filter functions as a pre-filter and the realistic number of combinations for manual comparison is determined by 2nd and 3rd filter.

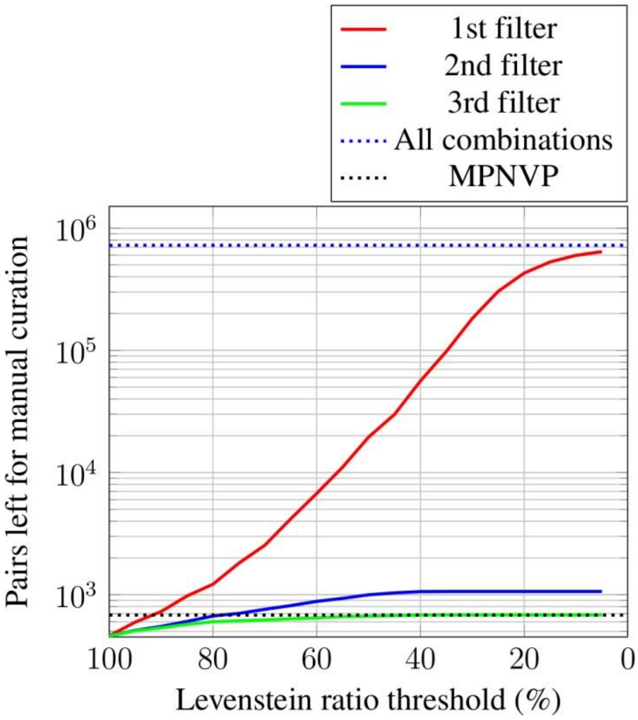

Fig. 5. **The effect of three-level-filtering**.

In the Fig. 5 the number of matched pairs after the 3rd filter at low similarity ratio threshold reaches MPNVP: maximal number of possible pairs. Introduction of compartment criterion reduces the number of mapped pairs to 60% of MPNVP (Fig. 6) demonstrating the high importance of correct compartment handling during comparison. Even at the similarity ratio threshold of 100% the number of metabolites with identical names reduces the number of matched pairs by half (48%) when compartments are taken into account.

Different criteria can be combined while comparing metabolite names. Additional introduction of edit distance to similarity ratio (compartments not taken into account) (Fig. 7) reduces the number of mapped pairs.

In the Fig. 7 legends *dist=100* means that edit distance threshold were 100. Since not one metabolite had a name longer than 100 characters, it can be assumed that this threshold is disabled. *Dist=20* means that for each particular pair of metabolites it is allowed to have 20 different characters. *Dist=5* means that only 5 edit operations are allowed to edit one name into another - if the edit distance for particular pair of metabolites is longer, the pair is automatically discarded.

In fact the edit distance criterion removes pairs with high similarity ratio if the number of different symbols exceeds the edit distance. Therefore the edit distance criterion gives effect when the similarity ratio drops below 80% because in case of high similarity usually the number of different letters is low. At some level of similarity ratio the edit distance criterion gives impact and prevents inclusion of metabolite pairs with high number of different letters.

The results revealed by automatic metabolite mapping were manually curated by a biologist. To reduce the amount of data for manual curation, the filtering by compartments were used (because such information was available). During the manual curation the biologist approved 407 metabolites out of 289 automatically mapped. Data used in manual curation is available in supplementary materials.

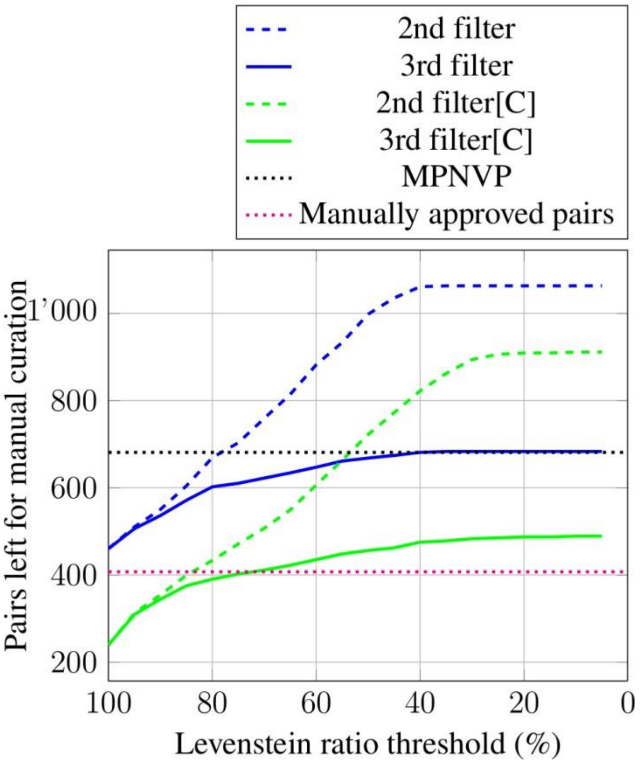

Fig. 6. **The effect of filtering by compartments on 2nd and 3rd filter.**

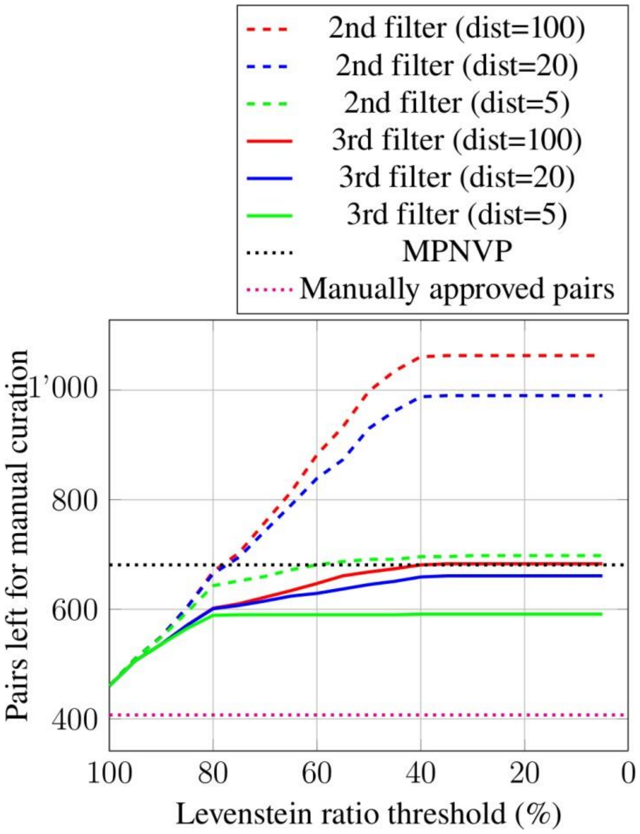

Fig. 7. **The effect of combined thresholds: similarity ratio and edit distance**

4. Conclusion

In case of two model comparison by metabolite names, the total number of combinations to be compared equals to the product of number of metabolites in both models which may lead to thousands or millions of candidate pairs. Therefore manual comparison is not a good alternative. The comparison by name similarity is a compromise between automatic and fully manual comparison. The major part of possible candidate pairs can be rejected automatically using simple, but laborious inspection. While the *three-level-filtering* algorithm does not solve the problem of fully automatic comparison, it can drastically reduce the amount of data that requires manual curation.

Low similarity ratio threshold values (down to 0%) leave all the work for manual comparison while high values (up to 100%) take into account only identical names. Therefore in case of rough comparison (for instance when many models have to be compared) the similarity ratio threshold should be kept high to reduce manual curation workload. In case of detailed comparison the similarity ratio threshold should be kept lower (50-60%) to avoid rejection of potential metabolite pairs.

Compartments should be taken into account mapping metabolites when possible to reduce the manual curation. The introduction of edit distance helps to filter away metabolite pairs with number of different symbols above the threshold. The effect of edit distance increase in case of long metabolite names.

The use of additional criteria (compartments, formulas, edit distance) can only improve the quality of automatic metabolites reconciliation, however, most of this data is often included in models.

Acknowledgements

This work and academic study is funded by a project “Support for doctoral studies in LUA” /2009/0180/1DP/1.1.2.1.2/09/IPIA/VIAA/017 agreement Nr. 04.4-08/EF2.D3.30

References

Boele, J., Olivier, B.G. & Teusink, B. (2012). FAME, the Flux Analysis and Modeling Environment. *BMC systems biology*, 6(1), p.8. http://dx.doi.org/10.1186/1752-0509-6-8

Duarte, N.C., Herrgård, M.J. & Palsson, B.Ø. (2004). Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome research*, 14(7), pp.1298–309. http://dx.doi.org/10.1101/gr.2250904

Gay, S., Soliman, S. & Fages, F. (2010). A graphical method for reducing and relating models in systems biology. *Bioinformatics*, 26(18), pp.i575–i581. http://dx.doi.org/10.1093/bioinformatics/btq388

Ideker, T. et al. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518), pp.929–34. http://dx.doi.org/10.1126/science.292.5518.929

Krause, F. et al. (2010). Annotation and merging of SBML models with semanticSBML. *Bioinformatics* (Oxford, England), 26(3), pp.421–2. http://dx.doi.org/10.1093/bioinformatics/btp642

Kuepfer, L., Sauer, U. & Blank, L.M. (2005). Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome research*, 15(10), pp.1421–30. http://dx.doi.org/10.1101/gr.3992505

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics-Doklady, 10(8), pp.707–710.

Mednis, M. & Aurich, M.K. (2012). Application of string similarity ratio and edit distance in automatic metabolite reconciliation comparing reconstructions and models. *Bit-Journal*, 1(1), pp.14–18. http://dx.doi.org/10.11592/bit.121102

Mednis, M., Brusbardis, V. & Galvanauskas, V. (2012). Comparison of genome-scale reconstructions using ModeRator. In 13th IEEE International Symposium on Computational Intelligence and Informatics. Budapest, pp. 79–84.

Mednis, M., Rove, Z. & Galvanauskas, V. (2012). ModeRator - a software tool for comparison of stoichiometric models. In 7th IEEE International Symposium on Applied Computational Intelligence and Informatics. Timisoara, pp. 97–100.

Mulligan, C. (2013). pylevenshtein. Available at: http://code.google.com/p/pylevenshtein/.

Oberhardt, M.A. et al. (2011). Reconciliation of Genome-Scale Metabolic Reconstructions for Comparative Systems Analysis P. E. Bourne, ed. *PLoS Computational Biology*, 7(3), p.18. http://dx.plos.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

Poggendorff, J.C. (1830). Annalen der Physik, J.A. Barth. Radulescu, O. et al., 2008. Robust simplifications of multiscale biochemical networks. *BMC systems biology*, 2(1), p.86. http://dx.doi.org/10.1186/1752-0509-2-86

Radulescu, O. et al. (2008). Robust simplifications of multiscale biochemical networks. *BMC systems biology*, 2(1), p.86.

Schellenberger, J. et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models: COBRA Toolbox 2 . 0. *Nature Protocols*, 6(9), pp.1290–1307. http://dx.doi.org/10.1038/nprot.2011.308

Schulz, M. et al. (2006). SBMLmerge, a system for combining biochemical network models. Genome informatics International Conference on *Genome Informatics*, 17(1), pp.62–71.

Thiele, I. et al. (2013). A community-driven global reconstruction of human metabolism. *Nature Biotechnology*, (September 2012).

Thiele, I. & Palsson, B.Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1), pp.93–121. http://dx.doi.org/10.1038/nprot.2009.203