Baltic J. Modern Computing, Vol. 4 (2016), No. 4, pp. 1038–1057
http://dx.doi.org/10.22364/bjmc.2016.4.4.27

Comparative Analysis of Biochemical Network Reconstructions

Martins MEDNIS

Latvia University of Agriculture, Liela iela 2, Jelgava, Latvia

martins.mednis@llu.lv

Abstract. Reconstruction of genome-scale metabolic network is a result of assembling various information sources about all biochemical reactions expected in the metabolic network of interest. Despite the efforts of leading bio-models databases to make comparison of biochemical networks by elements names obsolete, interest from researchers in using string similarity metrics in comparison of metabolites names has been growing.
Multiple challenges in comparison of reconstructions are discussed in this article and an insight into current approach of metabolic model comparison has been given. The discussion of challenges and attempts to solve them are followed by the author's proposed algorithm for models comparison that can be particularly useful in case of reconstructions. Special attention is given to the use of metabolites names and chemical formulas. The author's proposed algorithm has been implemented in a software tool *ModeRator*. The article is concluded with use cases of the comparison algorithm and the software tool.

Keywords: Model comparison, metabolites similarity, pairwise comparison

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) and therefore the functions of such networks are hard for the human mind to comprehend (Palsson, 2006).

During the last decade, over 50 genome-scale reconstructions have been built for various organisms. Despite the growing number of reconstructions and models in databases such as JWS (Snoep and Olivier, 2003; Van Gend et al., 2007) or Biomodels (Le Novère et al., 2006), the computational analysis has been rarely applied to comparisons between multiple organisms. The main reason for this is existence of differences between reconstructions that are inherited from the respective reconstruction processes of the organisms to be compared (Oberhardt et al., 2011).

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. The process of iterative model building promises to accelerate the biological discovery, product development, and process design (Palsson, 2006; Ideker et al., 2001). Consequently, the need for analysis, comparison, and merge of biomodels is growing. The demand for a method to relate different models (Gay et al., 2010) and compare or to couple them as parts of larger models has been noted by Radulescu et al., (2008).

The disuse of strict standardization in identification of metabolites and reactions leads to problematic reuse of models. Single metabolite can have multiple ways of notation. The use of synonyms worsens the problem. The differences in reconstructions annotations lead to the current situation where a number of biochemical network models of the same organism exist, but there is no way to inspect (in a reasonable time) how much they overlap, what parts do they have in common or is one model a subset of the other.

The currently available software solutions for automated comparison of reconstructions and models rely on elements identifiers and can recognize the identity of identically annotated elements. Therefore all could-be-equal elements with non-comparable or different type identifiers have to be pairwise inspected manually by a competent biologist. In case of genome-scale reconstructions checkable pairs of metabolites and reactions can reach several millions. Therefore biologists need computational help to reduce the manual work. However, there is a lack of automated solution that could handle the comparison of genome-scale reconstructions with poor or differently styled annotation.

1 The challenge of reconstruction comparison

The reconstruction of genome-scale metabolic network is a result of assembling various information sources about all the biochemical reactions expected in the metabolic network of interest (Palsson, 2006).

Many efforts in biology are inspired by the observation that different species have many common properties and molecular mechanisms (Bruggeman and Westerhoff, 2007). For instance, glycolysis process takes place in all the known organisms. The similarity of organisms and modules of biochemical networks justifies necessity of reconstruction comparison between different organisms, and not just different reconstructions or models of one organism.

Since 1997, over close to hundred genome-scale reconstructions for various organisms, including human, have been built. Human reconstruction *Recon2* (Thiele et al., 2013) containing 7440 reactions and 5063 metabolites was able to predict with 77% accuracy compared to experimental data changes of metabolite biomarkers 49 inborn errors of metabolism.

The growing number of available reconstructions can be used as integrated knowledge building new models or reconstructions for the process or organism of interest. To utilize the knowledge stored in model databases the models have to be compared to find their level of agreement and make use of highly reliable parts of existing models making use of existing knowledge.

The main reason for not applying computational analysis on comparisons between multiple organisms are the differences between reconstructions that are inherited from the respective reconstruction processes of the organisms to be compared (Oberhardt et al., 2011).

The overall purpose of reconstruction comparison is to find what reactions both reconstructions have in common. The information about common reactions can later be used by a biologist to make conclusions about common pathways.

The comparison of biochemical network reconstructions would be simple if all the reconstructions would be created according to a standard. That is not the case because different scientific groups in different countries with different traditions develop reconstructions over last 20 years. Several groups of challenges therefore are arising:

- different amount and quality of annotations;
- differences in metabolite description;
- differences in reaction notation;
- compartmentalization.

2 Current approach of metabolic model comparison

The possible problems with reconstruction comparison origin from the very beginning of the creation of reconstruction as process of reconstruction is based on analysis and combination of available information about biochemical reactions forming the network.

Automatically generated draft reconstructions may have comprehensive annotation, however, addition of information from various sources sooner or later spoils the initial consistency.

From the software tools surveyed, currently only *Tools-4-Metatool* (Xavier et al., 2011), *Compare Subsystems* (Oberhardt et al., 2011), *SemanticSBML* (Krause et al., 2010), *COBRA* (Becker et al., 2007), *The FAME* (Boele et al., 2012), *MetRxn* (Kumar et al., 2012), *BudHat* (Waltemath et al., 2013), *PINT* (Wang et al., 2010) and *MEMOSys* (Pabinger et al., 2011) provide functionality that is related to the comparison of models. Software tools mostly rely on internal or external identifiers, like *KEGG ID* and *ChEBI ID* and 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.

If the identifiers of reconstruction elements (compartments, metabolites and reactions) can be directly used to correctly identify elements across different reconstructions, then the whole comparison problem can be reduced to the comparison of two metabolism graphs. However, in the real-world applications the internal identifiers cannot be used to identify elements across reconstructions.

No software tool that could handle flexible comparison of genome-scale reconstructions with poor or differently styled annotation have been found during the survey.

3 The proposed algorithm

The overall purpose of reconstruction comparison is to find what reactions both reconstructions have in common. The information about common reactions can later be used by a biologist to make conclusions about common pathways which are formed by a series of reactions.

Usually the elements of metabolic network are metabolites and enzymes – metabolites react with each other with help of an enzyme producing other metabolites. Elements of reconstruction are data lists describing **metabolites**, **reactions** and **compartments**.

Logical order of steps needed to compare two biochemical reconstructions is:

1. compare and map compartments,
2. compare and map metabolites within compartments;
3. compare reactions.

Reactions can be compared only after the involved metabolites have been compared and mapped. Since metabolites may reside in different compartments, it is important to map compartments as well. Recognition and comparison of metabolites has attracted attention also from other researchers including Qi and Ozsoyoglu, 2013; Qi et al., 2014; Thavappiragasam et al., 2014. In this paper, the author focuses on cases where entities external identifiers, like *KEGG ID* and *ChEBI ID* can not be used in reconstruction comparison.

Depending on the source of the reconstruction and the file format, different set of additional information bits is available.

The following entities of reconstructions are compared:

- Metabolite comparison is based on their names that are provided in the reconstruction file. Information about compartments and chemical formulas is used to strengthen or weaken automatic decision about equality.
- Reactions are compared on their equations (reversibility, metabolites and their stoichiometry). Information about E.C. and GPR numbers is used to strengthen or weaken automatic decision about equality.

3.1 Comparison of metabolites

The pairwise comparison of metabolites means that each metabolite from one reconstruction is compared with each metabolite from the other reconstruction. The number of comparison operations needed equals $m \times n$, where n and m are numbers of metabolite count in reconstructions that are compared - so reconstructions each containing thousand metabolites will require one million comparison operations.

The algorithm to compare two individual metabolites is summarized in Figure 1 and is applied to each pair of metabolites. When algorithm ends with “Discard the

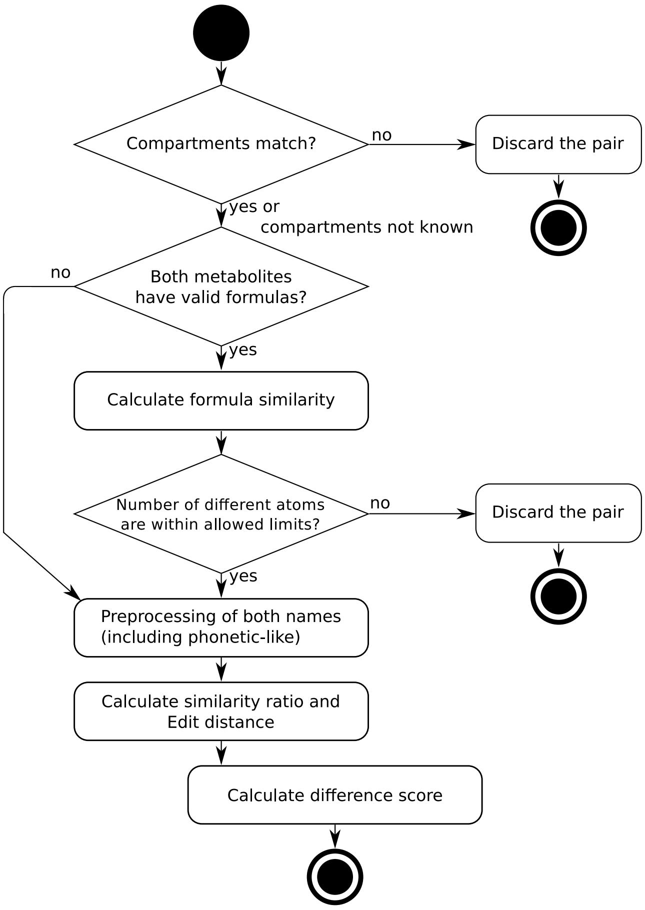

Compartments match?
no
Discard the pair
yes or
compartments not known
no
Both metabolites have valid formulas?
yes
Calculate formula similarity
Number of different atoms are within allowed limits?
no
Discard the pair
yes
Preprocessing of both names (including phonetic-like)
Calculate similarity ratio and Edit distance
Calculate difference score

Fig. 1: Algorithm of processing a pair of metabolites

pair" action, the particular metabolite pair will not be passed further to the mapping algorithm.

To calculate the similarity of metabolites names, author has chosen to use Gestalt Pattern Matching algorithm by Ratcliff and Metzener, 1988. This algorithm is available in Python standard module *difflib*. To calculate Edit distance Levenshtein, 1966 algorithm from *pyLevenshtein* library has been used.

The *similarity ratio* and *edit distance* independently characterize *the similarity* of any two metabolite names. The two metrics have different scopes: the *similarity ratio* is a percentage in floating point format, but the *edit distance* is an integer starting from 0.

In metabolite comparison algorithm these two metrics are combined into one. If the *similarity ratio* (R) clearly characterizes the *similarity* of given names, then the expression $(1 - R)$ calculates the *dissimilarity*. The *edit distance* already characterizes the dissimilarity of given names, therefore, the division of *edit distance* and the length of the shortest name, is still a number that characterizes the dissimilarity, but within the scope that is similar to that of the *ratio*.

Not division with the length of the shortest nor the longest name can guarantee a result between 0 and 1. The division with length of the shortest name will always be a greater number then the division with length of the longest, and it is essential for short metabolite names.

The combined *Difference score* of dissimilarity is presented in the Equation (1).

$$D = \mathcal{A} \cdot (1 - R) + \mathcal{B} \cdot \left(\frac{E}{L}\right), \tag{1}$$

where:
D : difference score
R : similarity ratio
E : edit distance
L : the length of the shortest name
$\mathcal{A}$: coefficient to affect the impact of similarity ratio
$\mathcal{B}$: coefficient to affect the impact of edit distance

The difference score is a sum of two dissimilarity metrics (1). This new summed metric is calculated for each metabolite pair. The essence of the equation is that for long metabolite names, the summed dissimilarity consists mainly of the *ratio* component, but for short metabolite names, the main contributor is the *edit distance* component. Examples of various names and corresponding difference score are given in the Table 1. For two identical names the calculated difference score is 0 (zero).

Table 1: Example of similarity ratio and distance variations for different strings. *Edit distance* algorithm by Levenshtein, (1966), *similarity ratio* algorithm by Ratcliff and Metzener, (1988) and *Difference score* algorithm by the Author.

String A	String B	Ratio*	Dist.**	D-score	Same?
ATP	ADP	0.66	1	0.67	no
D-glutamate	D-Glycerate	0.64	4	0.72	no
Glucose-6-phosphate	Glucose six phosphate	0.8	5	0.44	yes
Glucose six phosphate	L-Tryptophanyl-tRNA(trp)	0.31	22	1.74	no
Pyridoxal phosphate	Pyridoxal phosphate	1	0	0	yes

D-score: difference score, lower is better.
Ratio*: Similarity ratio.
Dist.**: Edit distance.

Phonetic-like preprocessing. Certain symbols in metabolite names can have different impact to biological meaning (see Table 2). For instance, special characters do not play significant role in the meaning of particular metabolite name, while numbers can change the biological meaning completely. A procedure is proposed to obfuscate characters with small impact on the meaning and to increase impact of numbers in the metabolite names. As seen in the Table 2, the phonetic-like preprocessing decreases the difference score for the first pairs (Aspartate), but increases it for the second pairs (Trihydroxypropane). In both cases the result improves chances to match equal metabolites and avoid matching of unequal.

Table 2: Two examples showing *raw* and *phonetically* processed metabolite names.

Name A	Name B	Phonetic	D-score	Same?
Aspartate-(L)	Aspartate L	raw	0.44	yes
AspartateL	AspartateL	processed	0	
3-Trihydroxy-propane	2-Trihydroxy-propane	raw	0.11	no
threeTrihydroxy-propane	twoTrihydroxy-propane	processed	0.35	

Comparison of metabolites formulas Chemical formulas can be used to verify that a particular pair of metabolites truly contains the same metabolites or not. If both formulas are available, the basic solution would be to compare formulas "as they are". If one or both formulas are not available, the decision can not be made. The *formula similarity* metric is given in the Equation (2). The formula comparison algorithm calculates how many atoms are different between two formulas.

$$F = \frac{1}{2}\left(1 - \frac{minH}{maxH}\right) + \frac{1}{2} + O \tag{2}$$

where:
F : formula similarity index
minH : smallest number of hydrogen atoms
maxH : greatest number of hydrogen atoms
O : number of other differing atoms

The formula similarity index is used as a multiplier for *Difference score*. The essence of Equation (2) is the following:

- for two equal formulas the equation will produce value 0.5 and therefore it will reduce the previously calculated *Difference score* by half;
- for formulas where only count of hydrogen atoms are different the produced value will be between 0.5 and 1.0 and therefore the *Difference score* will be slightly decreased (enhanced);
- if other atoms are different among the formulas, their count is added to the *formula similarity* and therefore the *Difference score* will be increased (degraded).

Examples of different formulas comparison is given in Table 3.

The conjunction of the *Difference score* and the *formula similarity* is given in Equation (3)

$$S = (D + C) \cdot F \tag{3}$$

Table 3: Examples of similarity value for different chemical formulas.

Formula A	Formula B	Differences	F.sim	Is equal?
H2O	H2O	–	0.5	yes
H2O	H2O2	O1	1.5	no
C7H14N2O8P	Formula01	–	1	no decision
C3H4O10P2	C3H5O10P2	H1	0.625	yes
C7H14N2O8P	C7H14N2O9P2	O1, P1	2.5	no
C7H14N2O8P	C7H16N2O8P	H2	0.56	yes

where:
S : the final score of the difference
D : *Difference score*
F : formula similarity
C : free constant - a positive integer

The free constant (C) is important and should not be set to zero because if both metabolites names are identical and therefore *Difference score* already is zero, then different chemical formulas would make no impact to decrease the similarity of metabolites. For example, if, the C is 1 then for equal names and equal formulas the final score will be 0.5. The non-zero value of the final score leaves open space for additional multipliers that can be added later after further research.

3.2 Mapping of metabolites

The *mapping of metabolites* is a procedure that explicitly defines which metabolite from one reconstruction corresponds to which element in the other reconstruction.

Metabolite mapping between two networks can only take place after the individual comparison of metabolites. The *mapping* is a procedure that explicitly defines which element from one network corresponds to which element in the other network. The problem of metabolite mapping can be classified as *bipartite graph matching*.

The problem of metabolite mapping can be classified as *bipartite graph matching*. A matching in a graph is a subset of its edges, no two of which share an endpoint. Polynomial time algorithms are known for many algorithmic problems on matchings, including *maximum matching* (finding a matching that uses as many edges as possible), *maximum weight matching*, and *stable marriage*

The *Difference score* for each metabolite pair is used as a criterion in mapping - only pairs with lowest difference gets mapped.

The metabolite mapping algorithm solves the *stable marriage problem* (Gale and Shapley, 1962). The difference from Gale algorithm is that it is not always required or possible to produce a *stable marriage* between all pairs of metabolites between two reconstructions. The task is to pair only *equal* metabolites, not to make sure that no one is left unpaired. In case of uncertainty is also necessary to keep a number of *multi-engaged* metabolite pairs, because it is the biologist that makes the final approval of which metabolites from one reconstruction suit to which metabolites on other reconstruction. The matching algorithm provides suggestions in cases where it is not possible to create a match automatically. Such cases appear quite often in real-world reconstructions. Also, reconstructions not necessarily have equal number of metabolites, and it is not always necessary to create a *stable marriage* for all metabolites even in equally sized reconstructions, because both reconstructions may cover different parts of genome, which overlap for a certain degree.

3.3 Comparison of reactions.

Reaction comparison algorithm not only tells whether two reactions are equal or not. It calculates the difference – how many reactants differ in both reaction sides.

The filtering (ignoring) of common metabolites like water and hydrogen can give overall improvement on comparison of reactions. However, in cases when a researcher does not know what are the metabolites that should be ignored, the comparison that tolerates small differences is desirable.

It should be stressed that two reactions can be equal despite some missing metabolites if the reactions in reconstruction are not balanced. The tolerant approach with missing metabolites should be taken only in cases when reaction balance can not be verified.

3.4 Impact of metabolite similarity thresholds on the comparison of reactions.

Figure 2 shows how the number of mapped metabolites affects the number of found reactions. In this example two reconstructions of *C. acetobutylicum* by Salimi and Mandal, (2010) and McAnulty et al., (2012) were compared. The curves in the plot are:

- Mapped metabolites – the number of approved and mapped metabolites;
- Equal reactions – the number of found equal reactions;
- Tolerated (OR) – the number of found equal reactions where one missing reactant from substrates **or** products is tolerated. The similarity threshold is 51% – the percentage of matching reactants;
- Tolerated (AND) – the number of found equal reactions where one missing reactant from substrates **and** products is tolerated. The similarity threshold is 51% – the percentage of matching reactants;
- Reactions with mapped metabolites – the number of reactions containing at least one mapped metabolite;
- MPNVP reactions – maximal possible number of common reactions (the number of reactions in the smallest reconstruction);
- MPNVP metabolites – maximal possible number of common metabolites (the number of metabolites in the smallest reconstruction, taking compartment coverage into account);
- Manually appr. metabolites – the number of manually (by a biologist) approved metabolites after the automatic comparison and matching.

What is interesting, the number of reactions where at least one metabolite is reconciled is close to the maximal possible number of common reactions from the very

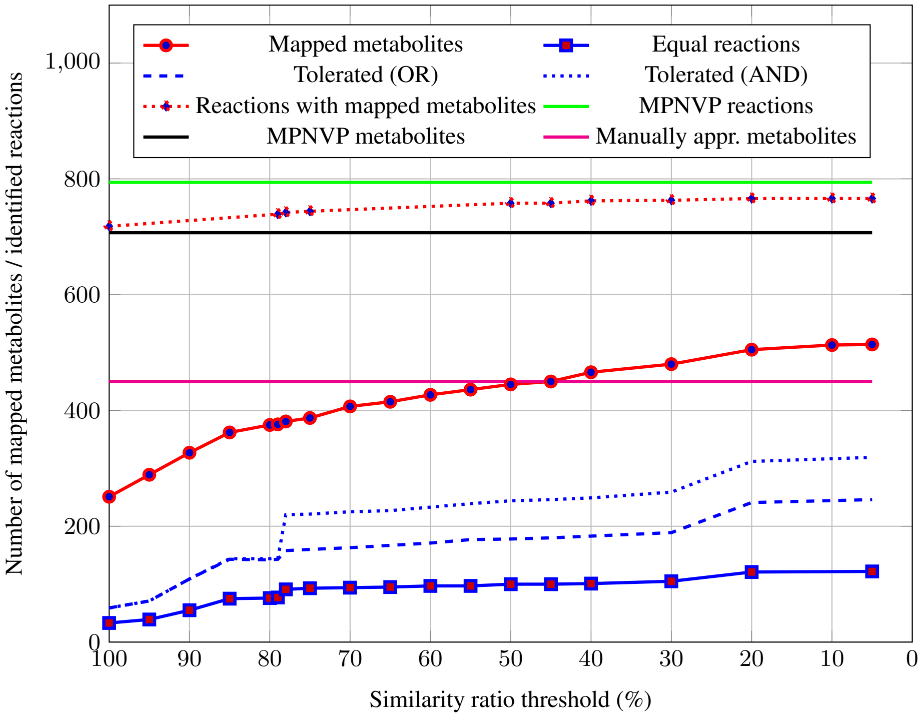


Fig. 2: Impact of mapped metabolites on the comparison of reactions

beginning. However, the number of equal reaction where it is possible pinpoint equal reactions barely reaches 15% of theoretically possible common reactions.

Figure 2 clearly shows number of things that have to be taken into account:

- automatic mapping of metabolite pairs with tolerated formulas can lead to *false positive* mapping of some metabolite;
- even knowing formulas and compartments for all metabolites does not guaranty correct metabolite matching;
- automatic mapping of metabolites (without manual approval) can lead to *false positive* results in comparison of reactions.

4 ModeRator - a software tool for comparison

The software tool *ModeRator* has been made according to the object-oriented paradigm. Reconstructions that are loaded into *ModeRator* become objects that have methods that enable their comparison with other reconstructions. The code of *ModeRator* is organized in many classes, but the core of the inner data model consists of just four classes: `st_model`, `metabolite`, `reaction` and `reactant`.

Handling of different file formats. To compare reconstructions in different file formats, *ModeRator* converts them to inner data model. Constructor classes for two importers have been implemented: for COBRA reconstructions in MS Excel spreadsheets and for SBML models. The constructors deal with specifics of particular file format.

The use of libSBML (Bornstein et al., 2008) enables convenient way of SBML model conversion to *ModeRator* inner data format. SBML files prior to Level 3 does not support storing of chemical formulas. *ModeRator* can process three different non-standard patterns of storing chemical formulas in SBML files: directly in the notes field, in paragraph in the notes field, in the metabolite's name field after the actual name. A special algorithm in *ModeRator* scans `name` and `notes` fields, splits them by various delimiters, and tries to parse splitted parts as chemical formulas. If the algorithm succeeds, it assumes that the formula is found.

In COBRA models, reactions are stored as strings in spreadsheet cells. Metabolites in one sheet, reactions in another sheet, and a set of columns with additional data. In order to read COBRA compatible MS Excel files, *ModeRator* makes use of Python `xlrd` library. Therefore, the rest of reading COBRA models involves only string processing. The importer of COBRA models deals with inconsistency of reaction string formatting.

A peculiarity of COBRA models is that there is no list of compartments. The compartment identifier (usually name) is indicated in a column beside other information about metabolites. Therefore the list of compartments is created dynamically while importing the list of metabolites. There can be situations where compartment is not indicated in a dedicated column but in brackets as a part of the metabolite abbreviation, for instance, `ADP[c]` or `H2O[m]`. A workaround for such cases has been implemented in *ModeRator* – if there are less than 2 compartments, *ModeRator* will try to guess them from metabolite names.

The Graphical User interface. The functions of *ModeRator* are arranged in consecutive tabs. For instance, in the first tab user can import two biochemical reconstructions. Other tabs are dedicated for comparison of metabolites or reactions.

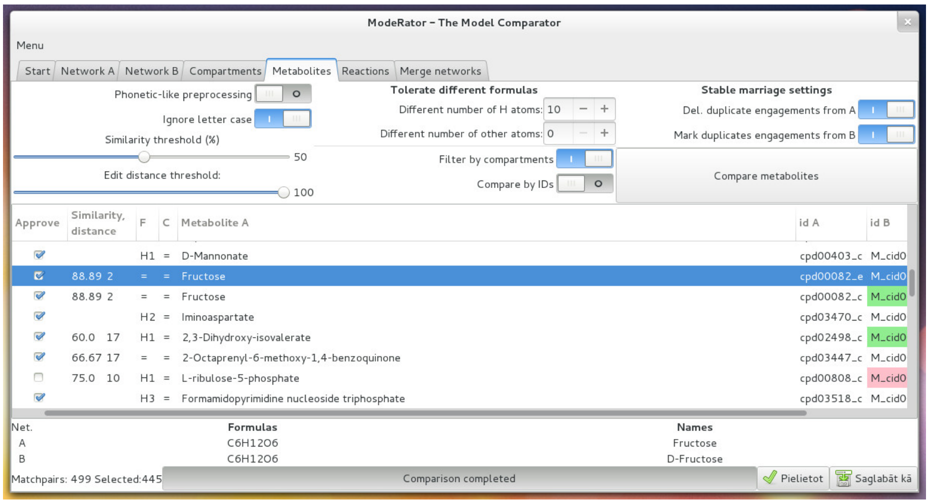
Filtering of metabolites. The presence of chemically unbalanced reactions makes the identification of equal reactions across multiple reconstructions harder. An option to equalize balanced and unbalanced reactions is to filter (ignore) specific metabolites, like water and hydrogen from all reactions.

Metabolite filtering feature is implemented in *ModeRator*. To filter a metabolite the user has to find it the list and enable filtering of particular metabolite by placing a tick. A quick search function is also available.

Since the *ModeRator* can also be used to generate graph drawings of the metabolism, in some cases it may be useful to filter other metabolites, like *CO2* to produce more transparent picture with less arrows.

Comparison of metabolites. The tab for metabolite comparison and mapping is shown in Figure 3. The metabolite names similarity and edit distance thresholds are set with graphical sliders. Phonetic-like name preprocessing can be enabled or disabled.

The GUI allows user to set various thresholds, like metabolites names similarity, allowed edit distance, the tolerance of formulas and filtering by compartments. The Author's proposed *Difference score* is used to weight matched metabolite pairs. It is also



Phonetic-like preprocessing: O
Ignore letter case: I
Similarity threshold (%): 50
Edit distance threshold: 100

Tolerate different formulas
Different number of H atoms: 10
Different number of other atoms: 0
Filter by compartments: I
Compare by IDs: O

Stable marriage settings
Del. duplicate engagements from A: I
Mark duplicates engagements from B: I

Approve	Similarity, distance	F	C	Metabolite A	id A	id B
☑		H1	=	D-Mannonate	cpd00403_c	M_cid0
☑	88.89 2	=	=	Fructose	cpd00082_e	M_cid0
☑	88.89 2	=	=	Fructose	cpd00082_c	M_cid0
☑		H2	=	Iminoaspartate	cpd03470_c	M_cid0
☑	60.0 17	H1	=	2,3-Dihydroxy-isovalerate	cpd02498_c	M_cid0
☑	66.67 17	=	=	2-Octaprenyl-6-methoxy-1,4-benzoquinone	cpd03447_c	M_cid0
☐	75.0 10	H1	=	L-ribulose-5-phosphate	cpd00808_c	M_cid0
☑		H3	=	Formamidopyrimidine nucleoside triphosphate	cpd03518_c	M_cid0

Net.	Formulas	Names
A	C6H12O6	Fructose
B	C6H12O6	D-Fructose

Matchpairs: 499 Selected:445

Fig. 3: Comparison of metabolites in *ModeRator*

possible to configure metabolite matching (*Three-level-filtering*) algorithm by disabling second and third filtering round.

Depending on the size of the reconstructions, comparison settings and user's computer the comparison process can take from few seconds to several minutes.

The user with biological knowledge makes the final decision ticking the first column whether automatically matched metabolite pairs truly are equal metabolites. Usually a manual curration and help from colleagues is needed. For that reason user can export the results of automatic metabolite matching to CSV file. After manual curration of automatically matched metabolites user has to apply metabolite mapping by pressing the *Apply* button.

Comparison of reactions. There are two methods for comparison of reactions. Comparison *by metabolite IDs* compares reactions based on metabolite mapping or internal identifiers. Comparison *by metabolite formulas* is applicable in cases when it is not possible to match metabolites by their names, but metabolite formulas are available in both reconstructions.

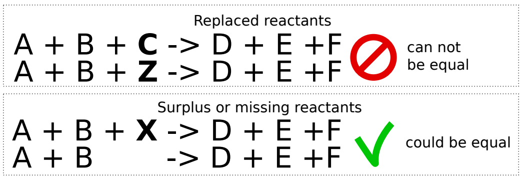


Fig. 4: Acceptable and not acceptable differences between reactions.

The *ModeRator* can match reactions where some metabolites are not mentioned in equations (see Figure 4). *Missing metabolites tolerance* settings allow user to set maximum number of allowed missing metabolites for each side and the overall tolerance limit. The *Limit* limits tolerance to certain length of reaction. The *length* is the number of involved reactants. By default the overall *Limit* set to “2” thus excluding transport reactions and other short reactions from tolerance settings influence.

Software dependencies and availability. The software has been tested on a number of free operating systems including *Ubuntu 12.10*, *Fedora 21* and *Debian 7*. The *ModeRator* can be downloaded from Biosystems Group homepage http://biosystems.lv/moderator2/. The website also provides sample files and documentation. The *ModeRator* is written in *Python*.

The recommended method for new users willing to avoid manual installation of all dependencies is to use *ModeRator* in a virtual environment. The download page provides OVA[1] package containing *xUbuntu* with the latest *ModeRator* pre-installed. OVA files can be opened with virtualization software, like *Virtualbox* and *VMware* on all major operating systems.

5 Use cases of *ModeRator*

The settings of *ModeRator* have to be adapted for particular cases depending on the type and quality of available information about reconstruction elements. Therefore the *ModeRator* settings for particular use cases are as different as the reconstructions are. Four pairs of biochemical network reconstructions were compared in three use cases. Different sets of information were available in reconstructions, hence different comparison settings were used. Table 4 lists all used reconstructions and comparison settings.

Table 4: Summary of comparison settings depending on the use case

Use case	**Organism**	Names	Identifiers	Formulas	Compartments	Variable charge	Phonetic	By mapped mets.	By formulas	Tolerated react.	Ignored mets.
		Metabolites						**Reactions**			
5.1	*E.coli*		✓		✓						
	S.cerevisiae	✓			✓						
5.2	*S.cerevisiae*	✓			✓		✓				
	C.acetobutylicum	✓		✓	✓	✓	✓	✓		✓	✓
5.3	*Z.mobilis*								✓		✓

[1] Open Virtual Appliance

The meaning of the settings are as follows:

- **Names** – metabolites were compared by names;
- **Identifiers** – metabolites were compared by internal identifiers;
- **Formulas** – metabolite formulas were available and were used for filtering after comparison by names;
- **Compartments** – information about compartments was available and was used for filtering after comparison by names;
- **Variable charge** – different number of hydrogen atoms was tolerated comparing formulas;
- **Phonetic** – phonetic-like preprocessing of metabolite names was used;
- **By mapped mets.** – the comparison of reactions was based on mapping of metabolites;
- **By formulas** – the comparison of reactions was based on chemical formulas of reactants;
- **Tolerated react.** – reactions with certain number of missing reactants were considered similar;
- **Ignored mets.** – specific metabolites, like water and hydrogen were ignored during the comparison of reactions.

5.1 Curated comparison of metabolites

Metabolites in two reconstructions of *E.coli*[2] containing 1314 and 1704 metabolites were compared by their identifiers. Manual curation was necessary for 31 pair with non equal names. In total, 30 metabolite pairs were approved during manual curation. One pair was left without decision because the available information was not enough for biologist to confirm or deny the identity of metabolites.

Metabolites in two reconstructions of *S.cerevisiae*[3] containing 1063 and 681 metabolites were compared by metabolite names. The threshold for similarity ratio was 68% and the threshold for edit distance was 15 edits. As the phonetic-like preprocessing and *difference score* was not used, the comparison in this use case was essentially based on similarity ratio. 723903 metabolite pairs were processed by computer. 400 metabolites were matched automatically. 376 (out of 447) metabolites were mapped after manual curation (Mednis and Aurich, 2012).

In this use case *ModeRator* version 2.5.5 was used.

5.2 Comparison of reactions after mapping of metabolites

In this use case metabolite pairs were weighted using the *difference score*, the edit distance threshold was set to 100 allowed edits. Phonetic-like preprocessing of metabolite names was enabled in some experiments.

[2] Both *E.coli* reconstructions were downloaded from *Biocyc* database

[3] *S.cerevisiae* reconstructions by Duarte et al., (2004) and Kuepfer et al., (2005)

Two reconstructions of S. cerevisiae. In this use case the same reconstructions of *S. cerevisiae* (see Use case 5.1) were compared. Unlike the previous use case, in this use case six comparison experiments with different settings were performed.

723903 **metabolite pairs** were processed by computer. Depending on the comparison settings 248 to 473 metabolites were matched automatically.

1146272 **reaction pairs** were processed by computer. Depending on the comparison settings 68 to 218 reactions were matched automatically.

Two reconstructions of Clostridium acetobutylicum. Two reconstructions of *C. acetobutylicum* by Salimi and Mandal, (2010) and McAnulty et al., (2012) containing 1134 and 707 metabolites and 1105 and 794 reactions were compared.

801738 **metabolite pairs** were processed by computer. Depending on the comparison settings 85 to 487 metabolites were matched automatically. 450 (out of 564, including metabolites with identical names) metabolites were mapped after manual curation by Dr.biol. Armands Vgants.

877370 **reaction pairs** were processed by computer. Depending on the comparison settings 109 to 449 reactions were matched automatically.

In this use case, the biologist approved 46 *C.acetobutylicum* metabolite pairs with names similarity under 50% including 24 metabolite pairs with similarity under 30% including 3 pairs with similarity under 15%. This allows to conclude that it is very difficult to set a *reasonable* threshold for name similarity, because the same metabolite may have quite different names. However, such cases are small part (¡5%) of all metabolite pairs that were automatically approved.

The similarity settings should be balanced with the costs of false-positive cases. In case of high importance of comparison results the threshold settings should be set at a level where all pairs even with low similarity would be analyzed by biologist spending more time, but gaining better confidence about the results. Even at low threshold settings of ModeRator the number of comparable pairs is heavily reduced and the data is well prepared for analysis by biologist.

5.3 Comparison of reactions skipping metabolite mapping

Two genome-scale reconstructions of *Zymomonas mobilis* by Lee et al., (2010) having 615 metabolites and 600 reactions, and Widiastuti et al., (2011) having 773 metabolites and 747 reactions were compared.

The peculiarity of this pair of reconstructions is the lack of some metabolite names as well as use of several synonyms describing the same metabolite. That makes the usage of metabolite names problematic. On the other hand both reconstructions have formulas. This use case demonstrates the opportunity to compare reactions skipping metabolite comparison. That can be done to get fast similarity overview.

448200 reaction pairs were processed by computer. Depending on the comparison settings 93 to 277 reactions were matched automatically.

The use case demonstrates the flexibility of *ModeRator* software enabling direct reaction comparison skipping the metabolite comparison step. This kind of approach is reasonable only in case of corresponding peculiarities of data when there is limited information about metabolites while reactions are described in good quality.

Different confidence levels can be reached taking into account enzyme numbers, which, in combination with other data may give strong confidence about identity of reactions. Still, even there the same enzyme can catalyze several similar reactions. Variations of comparison parameters like reversibility of reactions and ignoring of water and hydrogen can change the comparison results significantly.

ModeRator version 2.2 with command-line interface (Mednis et al., 2012) was used in this use case.

5.4 Application of model comparison for determination of consensus level of models

Automated generation of an intersection of the two models combined with its structural analysis (Rubina and Stalidzans, 2010; Rubina and Stalidzans, 2012) can give fast indication about the agreement level between metabolic models of a particular organism in their overlapping part. The creation of intersection is one of the functionalities of ModeRator. Some of the structural parameters can be used to measure the agreement level between the models by analysis of their intersection. Intersection analysis of model pairs compared in subsections 5.1 and 5.2. illustrate two different cases: high agreement intersection model in case of *E.coli* and the low agreement intersection model in case of *S.cerevisiae*. The reason of high agreement of *E.coli* models is the fact that they are built by the same group of researchers and both models reflect the development of the *E.coli* models of a particular group of researchers (Rubina et al., 2013).

Applicability of some structural parameters for the determination of agreement level has been analyzed using software BINESA (Rubina and Stalidzans, 2013). A low agreement level of a model pair resulting in a fragmented, poor quality intersection model can be indicated by low values of average degree, average incoming degree, average outgoing degree and average number of the neighbors. A low agreement of the model pair can be recognized also by the distribution of the incoming and outgoing degrees of the metabolites: high percentage of the low inter-connectivity metabolites and low percentage of the hubs (more than ten links).

Conclusions

- Additional identifying information about reconstruction elements can be used to strengthen or weaken automatic decision about equality of two elements. However the sets of additional information rarely overlap.
- The approach of fuzzy string comparison works well with long metabolite names. Lowering the threshold involves higher risk of false positives to be found.
- Certain symbols in metabolite names can have different impact to biological meaning. For instance, special characters do not play significant role in the meaning of particular metabolite name, while numbers can change the biological meaning completely.
- In some cases it is still possible that proposed algorithm returns multiple mapping links for the same metabolite due to the lack of lowest difference score for a single pair of metabolites. Such cases require manual curation and approval.

- Tolerance for variable formula charge improves chances to find truly equal metabolites.
- Automatic mapping of metabolite pairs with tolerated chemical formulas can lead to *false positive* mapping of some metabolites.
- Even knowing formulas and compartments for all metabolites does not guaranty correct metabolite matching.
- Automatic mapping of metabolites (without manual approval) can lead to *false positive* results in comparison of reactions.

The following **future developments** can be proposed:

Despite the efforts of leading databases, like MetaCyc and KEGG to make comparison of biochemical networks (Altman et al., 2013) by elements names obsolete, the interest from other researchers in using string similarity metrics in comparison of metabolites names (Qi and Ozsoyoglu, 2013; Qi et al., 2014; Thavappiragasam et al., 2014) has been growing.

One of the directions of further research in comparison of biochemical reconstructions is better recognition of common reactions. Computer aided matching of metabolites is a good start, but apparently not enough to reliably find common reactions. This is indicated by a low number of identified common reactions. A reason for this could be that a truly common reactions may contain identified common metabolites along with the unidentified. The current version of ModeRator can report such possibly common reactions, however, lowering similarity threshold increases the number of false positives.

Another direction of further development is to solve a problem of compartmentalization in different scales. Most anatomical compartments are separated from each other by phospholipid membranes. In a simpler reconstruction, for example, fluids of the body are divided into two compartments: fluids in cells and fluids outside cells. In a more detailed reconstruction cells themselves may have internal compartments, like, nucleus, mitochondria, Golgi apparatus or cytosol. This problem when compartments of reconstructions differ in granularity is aimed to be addressed in future releases of software tool ModeRator.

Acknowledgments

The author is acknowledging the financial support received from the Education and Science Administration of Latvia and European Social Fund within the scientific research project "Development of algorithms and methods for biomolecular networks analysis and visualization" (Ref. No. 258/2012 R119).

References

Altman, T., M. Travers, A. Kothari, R. Caspi, and P. D. Karp (2013). "A systematic comparison of the MetaCyc and KEGG pathway databases". In: *BMC Bioinformatics* 14.1, p. 112.

Becker, S. a., A. M. Feist, M. L. Mo, G. Hannum, B. Ø. Palsson, and M. J. Herrgard (2007). "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox." In: *Nature protocols* 2.3, pp. 727–38.

Boele, J., B. G. Olivier, and B. Teusink (2012). "FAME, the Flux Analysis and Modeling Environment." In: *BMC systems biology* 6.1, p. 8.

Bornstein, B. J., S. M. Keating, A. Jouraku, and M. Hucka (2008). "LibSBML: an API library for SBML." In: *Bioinformatics* 24.6, pp. 880–881.

Bruggeman, F. J. and H. V. Westerhoff (2007). "The nature of systems biology." In: *Trends in microbiology* 15.1, pp. 45–50.

Duarte, N. C., M. J. Herrgård, and B. Ø. Palsson (2004). "Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model." In: *Genome research* 14.7, pp. 1298–309.

Gale, D. and L. Shapley (1962). "College Admissions and the Stability of Marriage". In: *American Mathematical Monthly* 69, pp. 9–14.

Gay, S., S. Soliman, and F. Fages (2010). "A graphical method for reducing and relating models in systems biology". In: *Bioinformatics* 26.18, pp. i575–i581.

Ideker, T, V Thorsson, J. A. Ranish, R Christmas, J Buhler, J. K. Eng, R Bumgarner, D. R. Goodlett, R Aebersold, and L Hood (2001). "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network." In: *Science* 292.5518, pp. 929–34.

Krause, F., J. Uhlendorf, T. Lubitz, M. Schulz, E. Klipp, and W. Liebermeister (2010). "Annotation and merging of SBML models with semanticSBML." In: *Bioinformatics (Oxford, England)* 26.3, pp. 421–2.

Kuepfer, L., U. Sauer, and L. M. Blank (2005). "Metabolic functions of duplicate genes in Saccharomyces cerevisiae." In: *Genome research* 15.10, pp. 1421–30.

Kumar, A., P. F. Suthers, and C. D. Maranas (2012). "MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases." In: *BMC bioinformatics* 13.1, p. 6.

Le Novère, N., B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, J. L. Snoep, and M. Hucka (2006). "BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems". In: *Nucleic Acids Research* 34.Database issue, pp. D689–D691.

Lee, K. Y., J. M. Park, T. Y. Kim, H. Yun, and S. Y. Lee (2010). "The genome-scale metabolic network analysis of Zymomonas mobilis ZM4 explains physiological features and suggests ethanol and succinic acid production strategies". In: *Microbial Cell Factories* 9.1, p. 94.

Levenshtein, V. I. (1966). "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet Physics-Doklady* 10.8, pp. 707–710.

McAnulty, M. J., J. Y. Yen, B. G. Freedman, and R. S. Senger (2012). "Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism in silico." In: *BMC systems biology* 6, p. 42.

Mednis, M. and M. K. Aurich (2012). "Application of string similarity ratio and edit distance in automatic metabolite reconciliation comparing reconstructions and models". In: *Biosystems and Information technology* 1.1, pp. 14–18.

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

Oberhardt, M. A., J. Puchaka, V. A. P. Martins Dos Santos, and J. A. Papin (2011). “Reconciliation of Genome-Scale Metabolic Reconstructions for Comparative Systems Analysis”. In: *PLoS Computational Biology* 7.3. Ed. by P. E. Bourne, p. 18.

Pabinger, S., R. Rader, R. Agren, J. Nielsen, and Z. Trajanoski (2011). “MEMOSys: Bioinformatics platform for genome-scale metabolic models.” In: *BMC systems biology* 5, p. 20.

Palsson, B. Ø. (2006). *Systems Biology: Properties of reconstructed networks*. Cambridge University Press.

Qi, X. and G. Ozsoyoglu (2013). “Locating basic bio-entities in genome-scale reconstructed metabolic networks”. In: *2013 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, pp. 434–439.

Qi, X., Z. M. Ozsoyoglu, and G. Ozsoyoglu (2014). “Matching metabolites and reactions in different metabolic networks.” In: *Methods (San Diego, Calif.)* 69.3, pp. 282–97.

Radulescu, O., A. N. Gorban, A. Zinovyev, and A. Lilienbaum (2008). “Robust simplifications of multiscale biochemical networks”. In: *BMC systems biology* 2.1, p. 86.

Ratcliff, J. W. and D. Metzener (1988). “Pattern Matching: The Gestalt Approach”. In: *Dr Dobbs Journal* 13.7, pp. 46–72.

Rubina, T. and E. Stalidzans (2010). “Topological features and parameters of Biochemical Network Structures”. In: *International Industrial Simulation Conference*. Budapest: EUROSIS, pp. 228–236.

— (2012). “Evolution of alternative control loops of biological systems”. In: *5th International Scientific Conference on Applied Information and Communication Technologies*. Jelgava, Latvia., pp. 317–324.

— (2013). “BINESA a software tool for evolution modelling of biochemical networks structure”. In: *14th IEEE International Symposium on Computational Intelligence and Informatics, CINTI 2013 - Proceedings*, pp. 345–350.

Rubina, T., M. Mednis, and E. Stalidzans (2013). “Agreement assessment of biochemical pathway models by structural analysis of their intersection”. In: *Proceedings of 14th IEEE International Symposium on Computational Intelligence and Informatics*. Budapest, Hungary.

Salimi, F and R Mandal (2010). “Understanding Clostridium acetobutylicum ATCC 824 metabolism using genome-scale thermodynamics and metabolomics-based modeling”. In: *Computer Applications . . .* Cab, pp. 126–131.

Schulz, M., J. Uhlendorf, E. Klipp, and W. Liebermeister (2006). “SBMLmerge, a system for combining biochemical network models.” In: *Genome informatics International Conference on Genome Informatics* 17.1, pp. 62–71.

Snoep, J. L. and B. G. Olivier (2003). “JWS online cellular systems modelling and microbiology.” In: *Microbiology* 149.Pt 11, pp. 3045–3047.

Thavappiragasam, M., C. M. Lushbough, and E. Z. Gnimpieba (2014). "Automatic biosystems comparison using semantic and name similarity". In: *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '14*. New York, New York, USA: ACM Press, pp. 790–796.

Thiele, I., N. Swainston, R. M. T. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, S. G. Thorleifsson, R. Agren, C. Bölling, S. Bordel, A. K. Chavali, P. Dobson, W. B. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J. J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le Novère, N. Malys, A. Mazein, J. a. Papin, N. D. Price, E. Selkov, M. I. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J. H. G. M. van Beek, D. Weichart, I. Goryanin, J. Nielsen, H. V. Westerhoff, D. B. Kell, P. Mendes, and B. Ø. Palsson (2013). "A community-driven global reconstruction of human metabolism". In: *Nature biotechnology* 31.5, pp. 419–25.

Van Gend, C., R. Conradie, F. B. Du Preez, and J. L. Snoep (2007). "Data and model integration using JWS Online." In: *In Silico Biology* 7.2 Suppl, S27–S35.

Waltemath, D., R. Henkel, R. Hälke, M. Scharm, and O. Wolkenhauer (2013). "Improving the Reuse of Computational Models Through Version Control". In: *Bioinformatics* 29, pp. 742–748.

Wang, Y.-T., Y.-H. Huang, Y.-C. Chen, C.-L. Hsu, and U.-C. Yang (2010). "PINT: Pathways INtegration Tool." In: *Nucleic acids research* 38, W124–W131.

Widiastuti, H., J. Y. Kim, S. Selvarasu, I. A. Karimi, H. Kim, J.-S. Seo, and D.-Y. Lee (2011). "Genome-scale modeling and in silico analysis of ethanologenic bacteria Zymomonas mobilis." In: *Biotechnology and Bioengineering* 108.3, pp. 655–665.

Xavier, D., S. Vázquez, C. Higuera, F. Morán, and F. Montero (2011). "Tools-4-Metatool (T4M): Online suite of web-tools to process stoichiometric network analysis data from METATOOL". In: *Biosystems*, pp. 1–4.

Received October 14, 2016 , accepted October 19, 2016