EvoMS: an evolutionary tool to find *de novo* metabolic pathways

Matias F. Gerard*, Georgina Stegmayer, Diego H. Milone

Research Institute for Signals, Systems and Computational Intelligence (sinc(i)), FICH–UNL/CONICET, Argentina.

Abstract

The Evolutionary Metabolic Synthesizer (EvoMS) is an evolutionary tool capable of finding novel metabolic pathways linking several compounds through feasible reactions. It allows system biologists to explore different alternatives for relating specific metabolites, offering the possibility of indicating the initial compound or allowing the algorithm to automatically select it. Searching process can be followed graphically through several plots of the evolutionary process. Metabolic pathways found are displayed in a web browser as directed graphs. In all cases, solutions are networks of reactions that produce linear or branched metabolic pathways which are feasible from the specified set of available compounds.

Source code of EvoMS is available at http://sourceforge.net/projects/sourcesinc/files/evoms/. Subsets of reactions are provided, as well as four examples for searching metabolic pathways among several compounds. Available as a web service at http://fich.unl.edu.ar/sinc/web-demo/evoms/.

Keywords: Evolutionary algorithms, Metabolic network representation,

*Corresponding author. Tel: +54 (0342) 457 5234 int 118.

Email addresses: mgerard@santafe-conicet.gov.ar (Matias F. Gerard), gstegmayer@santafe-conicet.gov.ar (Georgina Stegmayer), d.milone@ieee.org (Diego H. Milone)

1. Introduction

Nowadays, biological databases have turned into true atlas that store information about genes, proteins and metabolites for a wide range of organisms (Karp and Caspi, 2011). Traditionally, metabolic pathways are shown as static maps, built as sets of reactions and compounds that fulfill some biologically relevant purpose. There are several tools to create, combine and edit those maps, besides analyzing their topological properties (Droste *et al.*, 2011; Arakelyan and Nersisyan, 2013; Posma *et al.*, 2014). However, new pathways can be built linking specific compounds by searching a set of reactions providing the connections. Finding a way to produce some compounds starting from a set of given ones involves searching a network of reactions linking those compounds. This is a widely studied problem in bioinformatics (Planes and Beasley, 2008) since it allows to know if a given organism can produce specific compounds from a particular food source, or simply find new ways to synthesize them (Lee *et al.*, 2009; Yim *et al.*, 2011). However, finding those relations by hand from the available and well–known compound–to–compound links (reactions) can be a really hard task.

Several methods have been developed for finding new pathways automatically (Planes and Beasley, 2009), mainly based on classical search strategies (Faust *et al.*, 2011). These tools search for metabolic pathways between only two compounds taking into account information about connectivity of metabolites (Croes *et al.*, 2005), the transference of atoms (Heath *et al.*, 2010), or the molecular structure of compounds (Rahman *et al.*, 2005), whereas in Faust et al. (Faust *et al.*, 2010) several elements can be re-

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Gerard, G. Stegmayer & D. H. Milone; "EvoMS: an evolutionary tool to find de novo metabolic pathways."
BioSystems, No. 134, pp. 43-47, 2015.

lated. The main problem faced by those methods is the exponential growth of the search trees when a large number of highly connected reactions and compounds are involved. This problem can be easily addressed by evolutionary algorithms (Pal *et al.*, 2006), since they are able to explore multiple solutions simultaneously in a large searching space. Recently, an approach based on this kind of algorithms was developed to search metabolic pathways between two metabolites (Gerard *et al.*, 2013). However, all of these tools provide only linear paths, taking into account the last synthesized product to select a new reaction.

In this work we present EvoMS, an evolutionary algorithm for searching branched metabolic pathways among a set of compounds. This tool performs the search taking into account the availability of substrates for each reaction in the pathway, in order to obtain a completely feasible solution. Thus, all the compounds synthesized by the reactions in the network are considered as potential substrates for new reactions, allowing to synthesize not only linear but also branching pathways. EvoMS can be easily customized to perform the search under different initial conditions, by modifying a single text file. For example, different ways to synthesize one or more compounds from a given one can be found, by specifying the set of reactions of a particular organism. It is also possible to indicate which compounds should be used as start substrates for the search. Progress of the exploration can be followed through several plots that show information about the population of solutions over time. Furthermore, the pathways found can be visually analyzed by a biologist in a simple way through graphical representations displayed on a web browser.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Gerard, G. Stegmayer & D. H. Milone; "EvoMS: an evolutionary tool to find de novo metabolic pathways."
BioSystems, No. 134, pp. 43-47, 2015.

2. Software description

2.1. Evolutionary model

EvoMS models each metabolic pathway as a network of reactions encoded into the chromosomes of an evolutionary algorithm. Thus, each gene is a reaction of the pathway. Chromosomes are initialized by selecting one reaction at a time from a set of available reactions. The products of each reaction are combined with the previous products to build an expanded set from which new reactions can be carried out. This model allows evolving branched metabolic pathways where two or more reactions could occur simultaneously to generate the substrates of a subsequent reaction. Once chromosomes are initialized, the search is performed by applying mutation and crossover, and evaluating the fitness of the encoded metabolic pathways for a particular set of available compounds.

Figure 1 presents an example of an EvoMS chromosome encoding a metabolic pathway. The network is composed by five reactions that relate three compounds, drawn as red and yellow circles. The red circle is the *initial substrate* of this network, while the yellow ones are the sought *final products*. Abundant compounds, such as water or ATP, are green circles. New compounds produced into the network are drawn as light blue circles.

To start the search, this tool only requires the definition of a set of compounds among which to find a metabolic network. Each one is specified in the COMPOUNDS.yaml text file by its code in KEGG notation (Kanehisa and Goto, 2000) and a label that indicates if it must be used to start the search. When there are two or more compounds with this label, EvoMS uses them all to start the search. At the end of the evolution, only the initial compound of the individual with the highest fitness will survive. Ad-

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Gerard, G. Stegmayer & D. H. Milone; "EvoMS: an evolutionary tool to find de novo metabolic pathways."
BioSystems, No. 134, pp. 43-47, 2015.

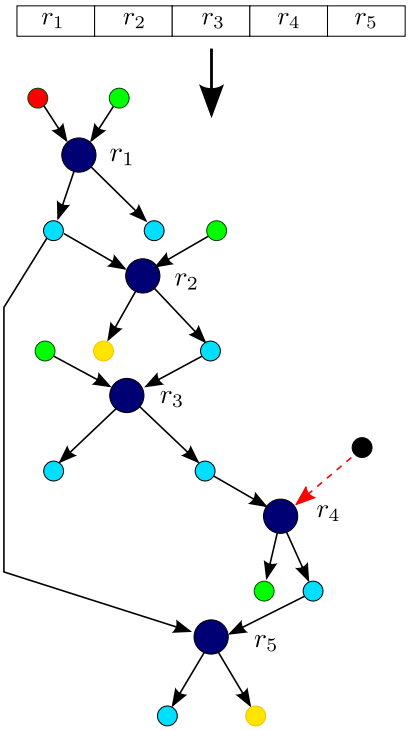

Figure 1: Example of a hypothetical metabolic pathway encoded into a chromosome. *Top*: Chromosome that encodes five reactions of a metabolic pathway. *Bottom*: Metabolic pathway linking three compounds through the five reactions (dark blue circles). Green circles indicate abundant compounds. Light blue circles correspond to new compounds generated by the reactions in the metabolic pathway. Initial substrate and final products are in red and yellow, respectively. Black circle corresponds to a non–available substrate, required by the reaction r_4.

ditionally, a set of abundant compounds, such as water and ATP, must be specified. Those will be freely available to be used for any reaction. A list with several abundant compounds, including several common cofactors, is provided by default in this file. They are all combined to build the initial set of available compounds.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Gerard, G. Stegmayer & D. H. Milone; "EvoMS: an evolutionary tool to find de novo metabolic pathways."
BioSystems, No. 134, pp. 43-47, 2015.

In order to build the reactions file, reversible reactions were split into two independent semireactions with opposite direction. Each one is specified by its code, substrates $S(r)$ and products $P(r)$, using the standard KEGG notation. Thus, reaction $S(r) \longleftrightarrow P(r)$ was decomposed as the direct and reverse semireactions, $S(r) \longrightarrow P(r)$ and $P(r) \longrightarrow S(r)$, respectively. The direction information for irreversible reactions was extracted from the KGML files of KEGG.

Tool settings are stored in a single text file, and includes the algorithm parameters and the names of the files that store the compounds to relate and the available reactions. The number of cores to perform a parallelized search can be specified in the desktop version of EvoMS. The behaviour of the tool can be easily modified by editing this file. For example, it is possible to easily change the search space just by indicating a different subset of reactions.

2.2. Fitness function

The evolutionary searching process is guided by an additive fitness function based on four terms, that evaluates the quality of the metabolic pathways found. This function and its terms are normalized in the interval $[0-1]$, and a maximum fitness is reached when a metabolic pathway meets two conditions: i) each reaction has all necessary substrates, and ii) there is a network of feasible reactions that relates all the compounds required. The four terms of the fitness function are: *validity*, *related compounds*, *rate of useful products* and *connectivity*

The *validity* term evaluates the proportion of reactions in the metabolic pathway for which substrates are available. For example, for the chromosome in Figure 1, reaction r_4 is invalid because it requires a substrate that is not available. Since the reaction r_5 uses a product of an invalid reaction

(r_4), it is also invalid. As a consequence, only 3 of the 5 reactions are valid, and the validity is 3/5.

The *related compounds* term evaluates if at least one reaction uses the initial substrate, as well as the proportion of the final products produced in the network. For the example in Figure 1, this term is 3/3 since all compounds to relate are in the pathway.

The *rate of useful products* determines the proportion of reactions in the metabolic pathway that produce, at least, one compound that has not been previously produced in the network. Assuming that all the light blue circles in the Figure 1 were different, this term would be 1.0.

Finally, the *connectivity* term evaluates the proportion of final products for which there is a network of reactions that relates them with the initial substrate. Analyzing the metabolic pathway in Figure 1, it can be seen that there is a network of reactions that produces both final products (yellow circles) starting from the initial substrate (red circle). In consequence, the connectivity term is 1.0.

2.3. Visualizations

An important feature of EvoMS is the visualization of the search process and the solutions found. Figure 2 presents an example of all graphics generated by EvoMS along the search, where each one shows the evolution of different aspects of the solutions. While the web version generates a static figure with all the information at the end of the search, the desktop version updates the search progress with each generation in the evolution. In both cases, all this information is automatically saved in a text file for further analysis.

Figure 2a shows the evolution of the average (red line) and maximum

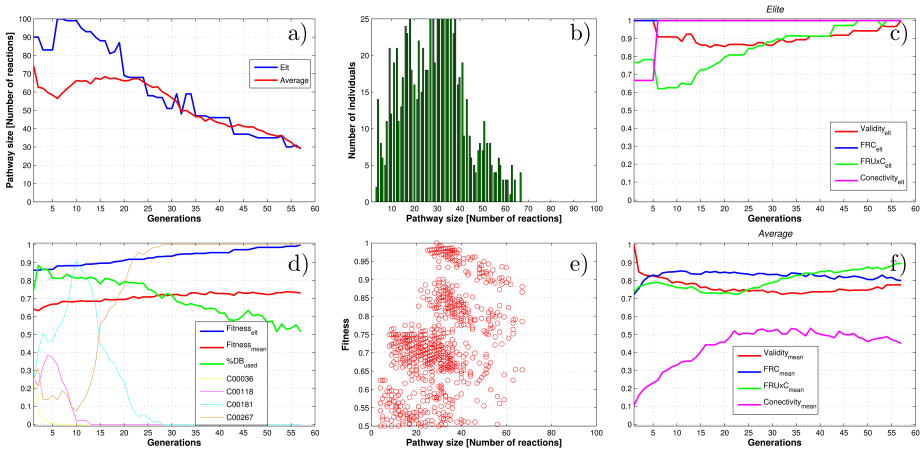

Figure 2: Views generated by EvoMS. a) evolution of pathways size; b) histogram of pathway sizes; c) evolution of the fitness terms for the best individual (*validity* in red, *related compounds* in blue, *rate of useful products* in green and *connectivity* in magenta); d) evolution of the average (red) and the maximum (blue) fitness for the population, proportion of available reactions used (green), and proportion of pathways initialized with each initial substrate; e) solutions displayed in terms of pathway size and fitness for the current generation; f) evolution of average values of the fitness terms.

(blue line) size of pathways, in terms of number of reactions. This can be useful to follow the incorporation and elimination of reactions. Figure 2b shows the histogram of pathway sizes in the current generation. Figure 2c shows the evolution of the four terms of the fitness function for the best solution, while Figure 2f shows their average value for the whole population. Figure 2d shows the average fitness for the population and the best solution. This subplot also displays the proportion of reactions of the search space used on each generation, together with the evolution of the proportion of individuals initialized with each initial substrate (a color line for each one). They allow to follow the competition among compounds to be used as the beginning of the metabolic network. Figure 2e shows the solutions in terms of the number of reactions and the fitness. It provides a quick overview of the evolution of pathway sizes associated to the best solutions.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Gerard, G. Stegmayer & D. H. Milone; "EvoMS: an evolutionary tool to find de novo metabolic pathways."
BioSystems, No. 134, pp. 43-47, 2015.

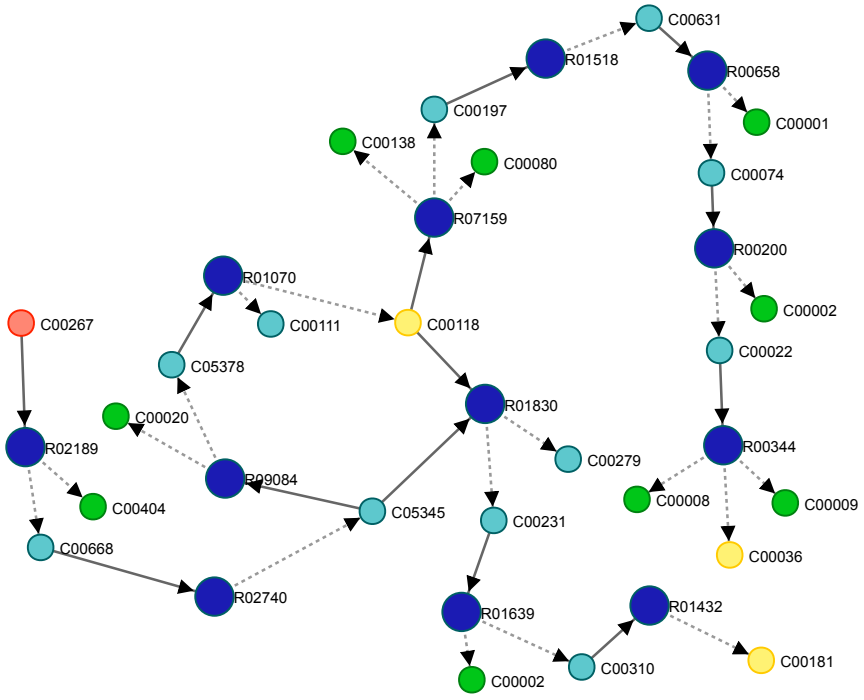

Figure 3: Graphical representation of a metabolic pathway found by EvoMS. Reactions are in dark blue, initial substrate is in red, final products are in yellow, and abundant compounds are in green. New products of the network are in light blue.

Information on this figure is a visual way of evaluating if the characteristics of the desired solution are actually improving during evolution which can be followed through the terms of the fitness function (validity, connectivity, etc). Thus, for example, an increasing value in the red line on Figure 2c indicates that the number of reactions that have the required substrates is rising. Similarly, an increase in the magenta line suggests that metabolic pathway is relating an increasing proportion of the desired compounds.

Resulting metabolic pathways are shown as an interactive web page. Figure 3 displays an example of this visualization. The solution is drawn as a network where reactions and compounds are represented by circles linked with two kinds of lines. Solid lines indicate substrates, while dashed lines indicate products. Furthermore, compounds are painted with several

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Gerard, G. Stegmayer & D. H. Milone; "EvoMS: an evolutionary tool to find de novo metabolic pathways."
BioSystems, No. 134, pp. 43-47, 2015.

colors: the initial compound C00267 is red, while the final products C00036, C00118 and C00181 are yellow. Compounds like C00668, produced inside the network, are light blue while abundant compounds are green. This representation can be manipulated interactively to rearrange the elements of the metabolic pathway.

3. Applications

EvoMS has important advantages over previous tools. Its main feature is the ability of finding networks that relate several compounds, at the same time. This is really important because most of metabolic pathways in nature are rather complex networks of interacting reactions among several compounds. EvoMS takes it into account to find solutions that resemble those found in nature.

To measure the network branching of a metabolic pathway it must be taken into account that a compound may lead to (that is, be substrate of) one or more reactions. Thus, the number of reactions that can be performed from each compound provides a simple way to calculate it. Note that abundant compounds must not be considered. Then, the branching factor was calculated here as the ratio between the sum of the number of reactions that employ each substrates and the total number of substrates. Accordingly, a branched pathway should have a ratio higher than 1.0. For example, in the pathway of Figure 1 only one reaction arises from 5 of the 6 non–abundant substrates (one red and five light blue circles), while 2 reactions (r_2 and r_5) arise from the remaining compound. Therefore the branching factor is $\rho = 7/6 = 1.167$.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Gerard, G. Stegmayer & D. H. Milone; "EvoMS: an evolutionary tool to find de novo metabolic pathways."
BioSystems, No. 134, pp. 43-47, 2015.

Table 1 shows results of 20 runs* for EvoMS versus methods based on classical search algorithms such as breadth-first search (BFS) and depth-first search (DFS) (Croes *et al.*, 2005; Rahman *et al.*, 2005; Heath *et al.*, 2010), and also with an evolutionary algorithm for searching linear metabolic pathways (EAMP) (Gerard *et al.*, 2013)†. Comparisons were performed by searching metabolic pathways between L-Histidine (code C00135 in KEGG) and L-Serine (code C00065 in KEGG). Both are essential for humans, and produce intermediate compounds for the citric acid cycle. For fair comparison with existing methods this is a simple linear case relating two compounds‡

As it could be expected, BFS found the shortest paths (5 reactions) while DFS, the longest ones (100 reactions, the maximum allowed in these runs). In both cases, only linear pathways were found, as reflected by the 1.00 value in the branching factor. EAMP found solutions with more reactions, being each one a linear pathway. Regarding EvoMS, it could also find pathways with a variable number of reactions, offering alternative mechanisms for relating compounds. It should be noted that the minimum number of reactions to relate several compounds is not known in advance. Intuitively, it could be expected that pathways requiring a few reactions to link compounds of interest would be more specific than those containing a lot of them. However, a bit larger solutions can provide additional information to understand the biological process, and therefore be more interesting from

*Runs were performed on a single computer with an Intel i7 CPU and 8 parallel threads.

†This previous work provides further comparisons between three methods for linear pathways.

‡Details of runs on Table 1 and samples of pathways obtained can be found in supplementary material.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Gerard, G. Stegmayer & D. H. Milone; "EvoMS: an evolutionary tool to find de novo metabolic pathways."
BioSystems, No. 134, pp. 43-47, 2015.

Table 1: Comparisons with other algorithms for searching a pathway between L-Histidine and L-Serine.

	Generations		Pathway size		Branching	
	med	max	med	max	ave	max
BFS	–	–	5	5	1.00	1.00
DFS	–	–	100	100	1.00	1.00
EAMP	23	305	9	19	1.00	1.00
EvoMS	23	518	6	9	1.06	1.22

the application point of view. EvoMS achieved an average 1.06 branching because this case could be solved with a linear pathway, and a more complex case could not be fairly used for a quantitative comparison with the other simpler methods. In spite of this simplification, it should be noted that EvoMS was able to find a pathway with branching factor of 1.22.

It is important to highlight that there are cases where a pathway between two compounds needs a branching to be possible. For example, in the case where a reaction needs two substrate, and each one of them should be provided by independent reactions that must be carried out in parallel. Supposing that only feasible solutions should be found, algorithms searching lineal pathways cannot find any solution in this case. Instead, EvoMS will be certainly capable of providing a solution to such problems because of its ability to model branched pathways.

4. Conclusions

EvoMS provides a simple tool for searching *de novo* metabolic pathways. Solutions are networks of feasible reactions from a set of available

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Gerard, G. Stegmayer & D. H. Milone; "EvoMS: an evolutionary tool to find de novo metabolic pathways."
BioSystems, No. 134, pp. 43-47, 2015.

compounds, that relate the specified metabolites. Configuration from text files provides a great flexibility to adapt this tool to different datasets of reactions and initial conditions. The displaying of measures assessed over solutions gives a simple way to follow the search process. Moreover, the graphical representation of the metabolic pathways on a web browser allows to easily appreciate in a glimpse its structure and main connections.

Acknowledgements

This work was supported by CONICET (PIP 2013-2015 #117) and UNL (CAI+D 2011 #548).

References

Arakelyan, A. and Nersisyan, L. (2013). KEGGParser: parsing and editing KEGG pathway maps in Matlab. *Bioinformatics*, **29**, 518–519.

Croes, D., Couche, F., Wodak, S., and van Helden, J. (2005). Metabolic Pathfinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Research*, **33**, W326–W330.

Droste, P., Miebach, S., Niedenführ, S., Wiechert, W., and Nöh, K. (2011). Visualizing multi-omics data in metabolic networks with the software Omix–A case study. *BioSystems*, **105**, 154–161.

Faust, K., Dupont, P., Callut, J., and van Helden, J. (2010). Pathway discovery in metabolic networks by subgraph extraction. *Bioinformatics*, **26**, 1211–1218.

Faust, K., Croes, D., and van Helden, J. (2011). Prediction of metabolic pathways from genome-scale metabolic networks. *BioSystems*, **105**, 109–121.

Gerard, M. F., Stegmayer, G., and Milone, D. H. (2013). An evolutionary approach for searching metabolic pathways. *Computers in Biology and Medicine*, **43**, 1704–1712.

Heath, A., Bennett, G., and Kavraki, L. (2010). Finding metabolic pathways using atom tracking. *Bioinformatics*, **26**, 1548–1555.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **28**(1), 27–30.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Gerard, G. Stegmayer & D. H. Milone; "EvoMS: an evolutionary tool to find de novo metabolic pathways."
BioSystems, No. 134, pp. 43-47, 2015.

Karp, P. and Caspi, R. (2011). A survey of metabolic databases emphasizing the MetaCyc family. *Archives of Toxicology*, **85**, 1015–1053.

Lee, S. Y., Kim, H. U., Park, J. H., Park, J. M., and Kim, T. Y. (2009). Metabolic engineering of microorganisms: general strategies and drug production. *Drug Discovery Today*, **14**, 78–88.

Pal, S., Bandyopadhyay, S., and Ray, S. (2006). Evolutionary Computation in Bioinformatics: A Review. *IEEE Transactions on Systems Man and Cybernetics*, **36**, 601–615.

Planes, F. and Beasley, J. (2008). A critical examination of stoichiometric and path-finding approaches to metabolic pathways. *Briefings in Bioinformatics*, **9**, 422–436.

Planes, F. and Beasley, J. (2009). Path finding approaches and metabolic pathways. *Discrete Applied Mathematics*, **157**, 2244–2256.

Posma, J., Robinette, S., Holmes, E., and Nicholson, J. (2014). MetaboNetworks, an interactive Matlab-based toolbox for creating, customizing and exploring sub-networks from KEGG. *Bioinformatics*, **30**, 893–895.

Rahman, S., Advani, P., Schunk, R., Schrader, R., and Schomburg, D. (2005). Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics*, **21**, 1189–1193.

Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J. D., Osterhout, R. E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H. B., Andrae, S., Yang, T. H., Lee, S. Y., Burk, M. J., and Dien, S. V. (2011). Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nature Chemical Biology*, **7**, 445–452.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Gerard, G. Stegmayer & D. H. Milone; "EvoMS: an evolutionary tool to find de novo metabolic pathways."
BioSystems, No. 134, pp. 43-47, 2015.