BioSystems 144 (2016) 55–67

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Evolutionary algorithm for metabolic pathways synthesis

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

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

ARTICLE INFO

Article history:
Received 16 November 2015
Received in revised form 1 March 2016
Accepted 1 April 2016
Available online 11 April 2016

Keywords:
Evolutionary algorithms
Search strategies
De novo pathway building
Reactions network
Sets of compounds

ABSTRACT

Metabolic pathway building is an active field of research, necessary to understand and manipulate the metabolism of organisms. There are different approaches, mainly based on classical search methods, to find linear sequences of reactions linking two compounds. However, an important limitation of these methods is the exponential increase of search trees when a large number of compounds and reactions is considered. Besides, such models do not take into account all substrates for each reaction during the search, leading to solutions that lack biological feasibility in many cases. This work proposes a new evolutionary algorithm that allows searching not only linear, but also branched metabolic pathways, formed by feasible reactions that relate multiple compounds simultaneously. Tests performed using several sets of reactions show that this algorithm is able to find feasible linear and branched metabolic pathways.

1. Introduction

Systems biology has quickly progressed thanks to the technical advances made in recent years to obtain quantitative and qualitative information of biological systems at different scales. These developments, in addition to contributions made by bioinformatics in several areas such as sequence analysis, modeling of protein structures, and building of interaction networks, help to understand the functioning of living beings (Tenazinha and Vinga, 2011). However, the increasing volume of data produced in biological experiments has led to the need to develop new computational tools capable of manipulating and analyzing it to extract knowledge (Bordbar et al., 2014; Chen and Zhang, 2014).

In nature, metabolic processes do not occur in isolation, but rather through complex networks made up of metabolic pathways that branch and interconnect (Ravasz et al., 2002; Lacroix et al., 2008). They generate a large variety of compounds that are used, for example, for structural purposes or energy storage, or just as substrates for key reactions in other processes (Jeong et al., 2000). These networks are a natural way of organizing relations (biochemical reactions) between compounds. Each reaction acts as a rule that determines the compounds consumed (substrates) and produced (products) in the process. These intricate relations are frequently modeled employing different types of graphs (Arita, 2012). Determining the whole sequence of reactions to produce a compound from another one consists in searching for a path that links both compounds in the graph. This problem is of particular interest in systems biology nowadays. The effort is focused on developing tools that allow identification of metabolic pathways capable of being manipulated to produce compounds of interest (Lee et al., 2009; Yim et al., 2011).

There are different methods to automatically search for metabolic pathways between two compounds. They are mainly based on classical search algorithms, such as breadth-first and depth-first search, and the A* algorithm (Russell and Norvig, 2010). All of them start by transforming the data into a type of graph appropriate for the search (Pey et al., 2011). One problem with these representations are the abundant compounds such as water and Adenosine 5′-triphosphate (ATP), which have a high connectivity as they participate in a large number of reactions (Gerlee et al., 2009). Thus, frequently the solutions found by the search strategies do not make biological sense since they use abundant compounds as intermediate steps in the synthesis of the desired product, and the availability of the other required substrates is not verified.

Different approaches to solve the problem of abundant compounds have been proposed. Croes et al. (2005) propose a weighting scheme to search a pathway between two compounds. They assign to each node a weight equal to the number of reactions where it participates, and find the lightest pathways between both ends.

* Corresponding author.
E-mail addresses: mgerard@sinc.unl.edu.ar (M.F. Gerard), gstegmayer@sinc.unl.edu.ar (G. Stegmayer), dmilone@sinc.unl.edu.ar (D.H. Milone).

http://dx.doi.org/10.1016/j.biosystems.2016.04.002
0303-2647/© 2016 Elsevier Ireland Ltd. All rights reserved.

This approach was extended by Faust et al. (2009), who applied the weighting scheme to a graph where its edges indicates the transfer of atoms from one compound to another one. Employing structural information of the compounds, McShan et al. (2003) built vectors of characteristic for each compound and performed the search by selecting the successive nodes using heuristics based on the distance between vectors. Similarly, Rahman et al. (2005) generated a binary fingerprint for each compound and applied similarity measures to guide the search process. Heath et al. (2010) proposed an approach based on tracking the flow of atoms, from the starting to the ending compound, trying to preserve as many of these atoms as possible. This allowed finding linear and branched pathways between two compounds. Branched solutions contain several alternative mechanisms to transfer atoms from the start to the end of the pathway. 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. Recently, a method based on evolutionary algorithms to search metabolic pathways between two metabolites was developed (Gerard et al., 2013), which avoids the problems of working with growing search trees. These methods provide paths only between two compounds and take into account the last synthesized product to select a new reaction.

Despite their characteristics, all these methods cannot find branching metabolic pathways that relate more than two compounds. In an effort to solve this issue, Faust et al. (2010, 2011) extended their pathway search strategy to relate a set of compounds by means of a network of reactions. Thus, solutions found consist of networks built as a combination of linear pathways among all pairs of compounds specified. Even though these solutions have ramifications, the feasibility of solutions is not taken into account since the availability of all substrates is not guaranteed.

While all these proposals provide sequences of reactions that relate the indicated compounds, the solutions found are often not biologically feasible. This is due to the assumption that all substrates are available, thereby the solution consists in finding a sequence of reactions to establish the relation. Thus, the availability of the compounds is not taken into account to perform the search and no restrictions are imposed on the possible reactions used to generate the solutions. Furthermore, given that all the previously synthesized compounds in the reactions chain are not taken into account to select a new reaction, valuable information to guide the search is lost and not properly used. 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 substrates, 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 could not find any solution in this case.

This work proposes a new approach based on the expanded set of compounds concept (ESC), which allows to relate several compounds at the same time by means of a network of feasible reactions. Given a set of available compounds and a feasible reaction from them, it is possible to expand this set by adding the products of the reaction. In this way, it is possible for a higher number of reactions can take place from the new set of compounds. Following this idea, our method only needs an initial set of available compounds in order to search for a metabolic pathway that relates the compounds of interest. To efficiently explore the search space, an algorithm based on evolutionary computation is proposed. This family of algorithms are inspired in biology and employ the principle of natural selection to evolve a population of potential solutions (Pal et al., 2006; Affenzeller et al., 2009; Boussaïd et al., 2013). These methods have been successfully applied to solve a wide range of problems in bioinformatics (Lee and Hsiao, 2012; Kayaa and Şule Gündüz-Öğüdücü, 2013; de Magalhães et al., 2014; Garai and Chowdhury, 2015). The search is guided by the fitness of individual in the population, which is evaluated using functions without formal requirements. Each individual encodes a solution, evolved employing genetic operators that combine the information of different individuals and introduce small variations during the evolutionary process.

A web interface to the algorithm has appeared in (Gerard et al., 2015). That report simply described the software from a user point of view, without details of the model and its functioning, mainly with a focus on the usability of the tool and the visualizations provided. It has to be noticed that this present contribution, instead, develops the main ideas behind the tool, providing a detailed explanation of the evolutionary model, its internal parameters and a wide experimental validation, with artificial as well as several real data of increasing complexity. The analysis of sensibility to parameters and robustness when facing a real problem is also included in the results. Moreover, a real case study for a well-known metabolic pathway that relates four biologically relevant compounds is presented, and two alternative solutions found to the standard metabolic pathway are described.

The paper is organized as follows. Section 2 describes the model of sets of compounds employed, the encoding in chromosomes, and the elements of the evolutionary algorithm, analyzing in detail the proposed operators and the measures that make up the fitness function. Section 3 describes the data employed in the experiments, their processing, the measures used to evaluate the algorithm performance, and several aspects of the searched networks. Section 4 analyses the effect of the variation of different parameters of the algorithm, the ability of the algorithm to scale to larger spaces, and a real case study. Finally, Section 5 presents the conclusions and future work.

2. Evolutionary algorithm based on expanded sets of compounds

Metabolic networks are constituted by compounds and the biochemical reactions r relating them (Lacroix et al., 2008). These relations allow certain groups of substrates to be modified in order to produce new products. Formally, reactions can be represented by means of the relation $S(r)\xleftarrow{r}P(r)$, where $S(r)$ and $P(r)$ correspond to the substrates and products of the reaction. Clearly, these relations require all substrates to be present in order to take place. In some cases, substrates are available in the medium where the reaction occurs. In other cases, they must be provided externally or through a previous reaction. In any case, each reaction which takes place can increase the available compounds so that new reactions can take place. This idea can be employed to model a metabolic pathway by considering it as a set of reactions carried out with a given order, that starts from a specified set of available compounds. Additionally, it is also possible to evaluate the feasibility of each reaction in the pathway by analyzing the availability of its substrates.

In an evolutionary algorithm, the linear structure of genes into a chromosome $\mathbf{c}$ can be easily used to represent the sequence of reactions, considering its order as indicative of the order that they take place in the pathway. Besides, it is possible to evaluate the feasibility of the pathway by associating an initial set of available compounds C^0 to $\mathbf{c}$, and verifying whether each reaction is possible based on this set and the products of all feasible reactions that have been previously carried out. Additionally, the use of an ESC enables to model branched metabolic pathways, where two or more reactions must happen simultaneously in order to generate all the necessary substrates for a subsequent reaction. Therefore,

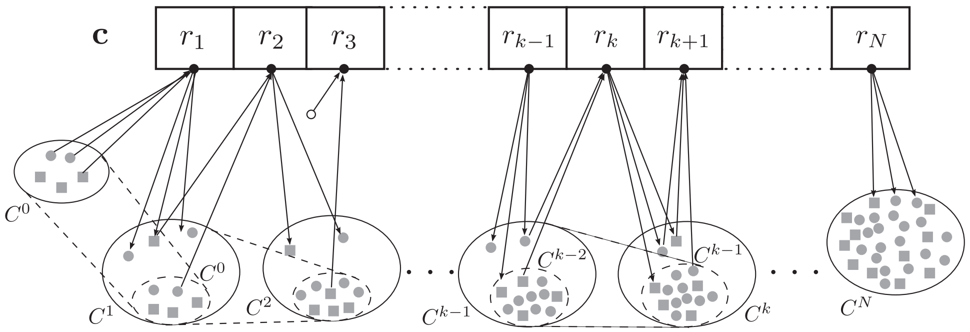

Fig. 1. Representation of the ESC model in a chromosome. *Top*: chromosome that encodes the reactions of a metabolic pathway. *Bottom*: the ESC for each reaction (solid lines) and previous sets (dash lines). Squares indicate available compounds. Filled circles correspond to new compounds generated in the metabolic pathway. The empty circle corresponds to a substrate required by the reaction r_3 that is not available in C^2.

each chromosome encodes a complete metabolic pathway, varying its size according to the number of reactions the pathway has.

Fig. 1 exemplifies a metabolic pathway encoded in a chromosome together with the ESC associated to each reaction. The substrates required for the reaction r_k must be available in the ESC C^{k-1}, otherwise the reaction will not be valid and the set of compounds will remain unmodified ($C^k = C^{k-1}$). Thus, if the substrates for the reaction r_k are available in the ESC $C^{k-1} = C^{k-2} \cup P(r_{k-1})$, this reaction produces the new set $C^k = C^{k-1} \cup P(r_k)$. Therefore, the ESC continues to be updated until the set C^N is reached.

2.1. Description of the algorithm

The proposed algorithm, named EvoMS (Evolutionary Metabolic Seeker), employs the sets of compounds model to search for feasible metabolic pathways that relate a group D of specified compounds. In order to facilitate further explanations, the term *initial substrate* is introduced to denote the compound belonging to D used to find the pathway, and *final products* to indicate the remaining compounds in D after selecting the initial one. The general structure of the algorithm and the selection operator are similar to the ones used in genetic algorithms (Bäck et al., 2000).

Briefly, the algorithm starts with the initialization and fitness evaluation of the population in the first generation, $f(\mathbb{P}^0)$, which is subjected to the evolution process until the stopping criterion is satisfied. This criterion consists of two elements: a maximum allowed number of generations G_M and a fitness value 1.0. The evolutionary process comprises six steps: extracting the best individual (chromosome $\mathbf{c}^*$), selecting the parents $\mathbb{X}^G$ for the new generation, creating the descendants $\mathbb{C}^G$ through crossover of the selected parents and mutation of their offspring, building the new population $\mathbb{P}^{G+1} \leftarrow \{\mathbf{c}^*\} \cup \mathbb{X}^G \cup \mathbb{C}^G$ and evaluating the fitness $f(\mathbb{P}^{G+1})$ of the new population. The solutions found by EvoMS correspond to networks of feasible reactions that use C^0 to relate compounds in D. The feasible reactions which are not part of these links are filtered later. The crossover operator employed consists of a combination of one-point and two-point crossover operators. Given two parents, this operator selects a portion of genetic material from one parent and inserts it in a random position of the other one, discarding the original genetic material in the second parent after the point of insertion. The mutation operator and the initialization strategy consider the use of sets of compounds. These will be explained in detail in Sections 2.2 and 2.4.

2.2. Initialization based on ESC

The initialization of EvoMS is carried out employing a strategy based on ESC and taking into account the validity of the reactions. The use of this strategy has two objectives. On the one hand, it avoids using random initialization, which could lead to very poor initial solutions. On the other hand, it introduces the use of subpopulations. Each one is made up by a set of individuals using the same initial substrate. It allows to overcome the problem of selecting the initial one when there is no information to make such decision. Thus, subpopulations will compete to determine the initial substrate for the metabolic pathway searched. The initialization process is carried out in two phases: identifying the number of subpopulations and initializing the individuals. Algorithm 1 describes the steps of this process. In order to initialize the population $\mathbb{P}$, it is necessary to define a set of abundant compounds A, such as water and ATP, which will be available for all reactions during the search. This set is automatically updated during the initialization, incorporating the external compounds E to generate the set $A' = A \cup E$. The set E is made up of all substrates that cannot be synthesized by any reaction provided.

The first phase of the initialization consists in determining the number of subpopulations to be generated (lines 7–12 of Algorithm 1). Each compound $d \in D$ is evaluated in order to identify those which are used as substrate of any reaction. Used compounds and substrates of those reactions are stored in two lists, I and R, respectively. The amount of compounds in list I define the number of subpopulations that should be created.

The second phase consists in the initialization of subpopulations, each one containing equal number of individuals (lines 13–26). This process is similar for all members. Firstly, the chromosome $\mathbf{c}$ is initialized as an empty list, and the number of genes N_I that it should contain is randomly selected. Secondly, a set of available compounds C^0 associated to the chromosome is built. It is made of the union of the abundant compounds (A) and the external ones (E), plus all the substrates (Q_j) required by reactions that use the initial substrate I_j. The initial reaction r_1 is randomly selected from those using I_j as initial substrate, and its products update the set of available compounds $C^1 = C^0 \cup P(r_1)$. Then, an iterative process is performed until the specified number of genes N_I is reached, or there is no more reactions to insert. In each step, a reaction r_k is selected at random, without repetition from all reactions than can take place from the compounds present in the set C^{k-1}. Afterwards, the set of accumulated compounds $C^k = C^{k-1} \cup P(r_k)$ is updated with products of the selected reaction. Finally, the individual is incorporated to the population $\mathbb{P}$ and the process is repeated.

If the final population has more than M individuals, some members are randomly removed until the specified size is reached (lines 27–28).

Algorithm 1. Initialization strategy based on sets of compounds.

```
1  $A' \leftarrow A \cup E$
2  $N_M \leftarrow$ maximum pathway size allowed
3  $M \leftarrow$ population size
4  $N \leftarrow 0$
5  $Q, I \leftarrow$ empty list
6  $U \leftarrow \emptyset$
7  foreach $d \in D$ do
8      $U \leftarrow \bigcup_{\forall r / d \in S(r)} S(r)$
9      if $U \neq \emptyset$ then
10         $N \leftarrow N+1$
11         $Q_N \leftarrow U$
12         $I_N \leftarrow d$
13 for $j \leftarrow 1$ to $N$ do
14     for $i \leftarrow 1$ to $\lceil \frac{M}{N} \rceil$ do
15         $k \leftarrow 1$
16         $N_I \leftarrow$ select a random integer in $[\frac{N_M}{2}, N_M]$
17         $\mathbf{c} \leftarrow$ empty list
18         $C^0 \leftarrow A' \cup (Q_j - D) \cup \{I_j\}$
19         $R \leftarrow \{r / |S(r) \cap C^{k-1}| = |S(r)| \wedge I_j \in S(r)\}$
20         while $k \leq N_I$ and $R \neq \emptyset$ do
21             $r_k \leftarrow$ select one reaction from $R$ not included in $\mathbf{c}$
22             $\mathbf{c} \leftarrow$ insert $r_k$
23             $C^k \leftarrow C^{k-1} \cup P(r_k)$
24             $k \leftarrow k+1$
25             $R \leftarrow \{r / |S(r) \cap C^{k-1}| = |S(r)|\}$
26         $\mathbb{P} \leftarrow$ insert $\mathbf{c}$
27 if $|\mathbb{P}| > M$ then
28     $\mathbb{P} \leftarrow$ randomly select $M$ individuals from $\mathbb{P}$
29 return $\mathbb{P}$
```

2.3. Fitness function

The fitness $f(\mathbf{c})$ of the individuals in the population is evaluated employing an additive function made up of four terms, each one focused on a specific property of the solution. The fitness function and its terms are normalized in [0, 1], and the maximum fitness is reached when a solution is found. A metabolic pathway is considered a solution when it meets two conditions: (i) each reaction has the necessary substrates, and (ii) there is a sequence of valid reactions that relate the initial substrate with each final product. Therefore, the fitness function is defined as

$$f(\mathbf{c}) = \frac{1}{4}[\mathcal{V}(\mathbf{c}) + \mathcal{L}(\mathbf{c}) + \mathcal{I}(\mathbf{c}) + \mathcal{C}(\mathbf{c})], \qquad (1)$$

and the way of calculating the four measures is described below.

2.3.1. Validity

The term $\mathcal{V}(\cdot)$ evaluates the proportion of reactions in the metabolic pathway that have the required substrates. In this sense, the reaction r_k is valid if $S(r_k) \subseteq C^{k-1}$, which corresponds to the set of accumulated compounds until the reaction r_{k-1}. This measure is calculated as

$$\mathcal{V}(\mathbf{c}) = \frac{1}{|\mathbf{c}|} \sum_{k=1}^{|\mathbf{c}|} \mathbf{1}_{S(r_k) \subseteq C^{k-1}}, \qquad (2)$$

where $|\mathbf{c}|$ is the number of genes of $\mathbf{c}$, and $\mathbf{1}_{A \subseteq B}$ is the indicator function, which takes the value 1 when $A \subseteq B$ and 0 in another case. The validity of a metabolic pathway is maximum when each reaction has the substrates it needs.

2.3.2. Linking

The term $\mathcal{L}(\cdot)$ in (1) evaluates two aspects of the metabolic pathway: (i) if the initial substrate is used, at least, by one reaction, and (ii) the proportion of the final products that are synthesized. This measure is calculated as

$$\mathcal{L}(\mathbf{c}) = \frac{1}{2}\left(|S^*(\mathbf{c}) \cap \{d\}| + \frac{|P^*(\mathbf{c}) \cap (D - \{d\})|}{|D - \{d\}|}\right), \qquad (3)$$

where d denote the initial substrate of $\mathbf{c}$, $S^*(\mathbf{c}) = \bigcup_{\forall r \in \mathbf{c}} S(r)$ and $P^*(\mathbf{c}) = \bigcup_{\forall r \in \mathbf{c}} P(r)$ are the sets containing all substrates and products of the pathway, respectively. This measure reaches its maximum value when a reaction employs d as a substrate and all compounds $D - \{d\}$ are produced.

2.3.3. Innovation

The term $\mathcal{I}(\cdot)$ determines the proportion of reactions in the metabolic pathway that produce, at least, one compound that has not been previously generated in the sequence. Consequently, this term favors the incorporation of novel reactions that are not already present in the pathway and that produce new compounds. This measure is calculated as

$$\mathcal{I}(\mathbf{c}) = \frac{1}{|\mathbf{c}|} \sum_{k=1}^{|\mathbf{c}|} \mathbf{1}_{P(r_k) \nsubseteq C^{k-1}}. \qquad (4)$$

The maximum value is reached when each reaction produces, at least, a new compound.

2.3.4. Connectivity

The term $\mathcal{C}(\cdot)$ in (1) evaluates the proportion of the final products for which there is a sequence of reactions that relates them with the initial substrate d. This measure is calculated in two steps. The first step consists in building a set of accumulated compounds Z, which is then used in the second step to calculate the connectivity. The set Z employed in the first step is built using Algorithm 2. From the initial set $Z = \{d\}$, the algorithm evaluates each reaction in the chromosome and verifies whether the reaction employs any of the compounds in Z as a substrate, updating this set with its products if the reaction is a valid one. The algorithm returns the set of compounds that are employed to relate d with each member of $D - \{d\}$. Then, connectivity is calculated from the set Z obtained according to

Algorithm 2. Searching for compounds related to the initial substrate.

```
Z ← initial substrate of c
for $k$ ← 1 to $|\mathbf{c}|$ do
    if $|S(r_k) \cap Z| > 0$ then
        $Z \leftarrow Z \cup (P(r_k) - C^0)$
    if $|S(r_k) \cap C^{k-1}| = |S(r_k)|$ then
        $C^k \leftarrow P(r_k) \cup C^{k-1}$
    else
        $C^k \leftarrow C^{k-1}$
return Z
```

$$\mathcal{C}(\mathbf{c}) = \frac{|Z \cap D| - 1}{|D| - 1}. \qquad (5)$$

This measure takes its maximum value when there are sequences of reactions that relate the initial substrate with each final product.

2.4. Mutation based on ESC

The proposed mutation operator introduces changes based on the composition of the sets of accumulated compounds with a probability p_m. These changes can be the deletion or insertion of one gene into the chromosome, with probabilities p_e and $1 - p_e$, respectively. It introduces variations in the pathway size, because deletions remove randomly one gene from the sequence, and each insertion adds one reaction that was not already in the sequence.

Insertion starts by randomly selecting a position $k \in [1, N+1]$, where the gene will be inserted. Afterwards, two lists of reactions are built from C^{k-1}. When $k = 1$, C^0 corresponds to the initial set of available compounds associated to the chromosome **c**. The list of valid reactions contains all the possible reactions from the compounds in C^{k-1}, while the list of invalid reactions has all the remaining reactions of the search space. The list of valid or invalid reactions from which the reaction will be selected is chosen with probabilities p_v and $1 - p_v$, respectively. When the chosen position is in the interval $[1, N]$, the gene that is in that position and all genes coming after in the sequence are moved one place forward to allow the insertion.

Fig. 2 shows an example of the proposed mutation operator for a chromosome containing $N = 10$ genes. In the case of insertion, the chosen position is 4 and the set of accumulated compounds C^3 is built considering the products of all valid reactions until the gene that contains r_3. From this set, the list of valid reactions, whose substrates are available in C^3, is generated, as well as the list of invalid reactions, which do not have all necessary substrates in C^3. A reaction from these lists is randomly extracted, with probability p_v for the list of valid reactions. In the example of deletion, the gene containing the reaction r_8 is eliminated from the sequence, and adjacent reactions r_7 and r_9 are spliced. Clearly, in both cases the number of genes in the chromosome is modified.

3. Data and evaluation measures

3.1. Reactions information

Reactions employed in the experiments were extracted from the KEGG database. Actually, reactions from other repositories, such as MetaCyc (Altman et al., 2013), could be used as well. The direction for each reaction was assigned using the information contained in the KGML files associate to the reference maps (Ogata et al., 1998; Goto et al., 2002). Each reversible reaction was modeled as a pair of independent reactions with opposite direction. For example, the reaction $S(r) \longleftrightarrow P(r)$ was separated into the semi-reactions $S(r) \rightarrow P(r)$ and $P(r) \rightarrow S(r)$. The set of abundant compounds A employed in the experiments is shown in Table 1. The

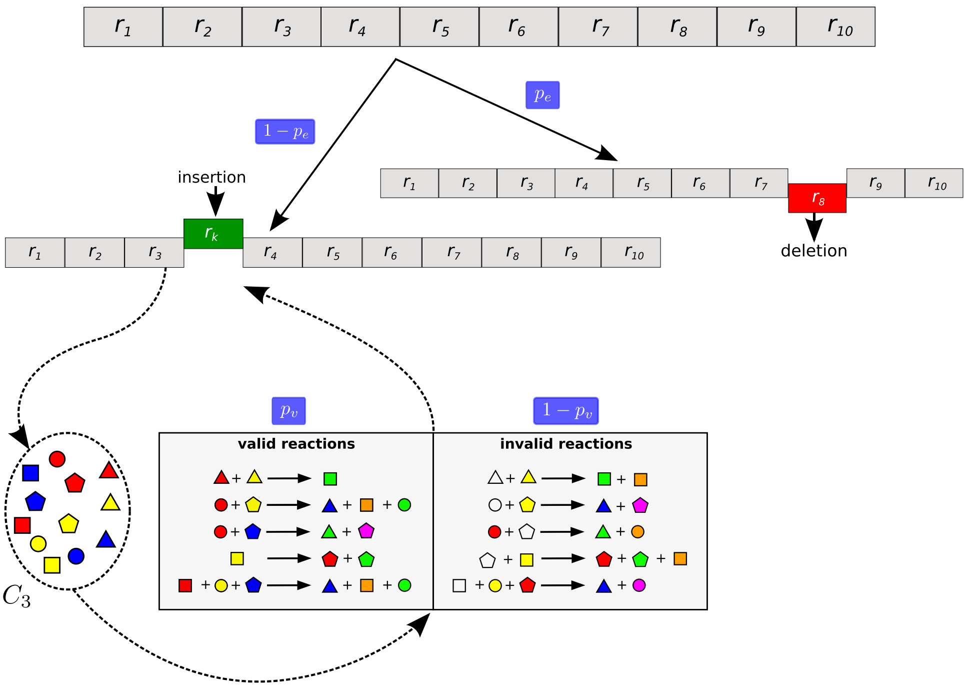

Fig. 2. Diagram of the proposed mutation operator for a chromosome containing $N = 10$ reactions. *Left*: example of gene insertion in position 4 of the chromosome. Available compounds in C^3 are indicated as filled polygons. *Right*: example of gene deletion.

Table 1
Abundant compounds employed to search for branched metabolic pathways. The table indicates the name of the compound and the corresponding KEGG code.

KEGG code	Name	KEGG code	Name	KEGG code	Name	KEGG code	Name
C00001	H_2O	C00005	NADPH	C00009	Phosphate	C00020	AMP
C00002	ATP	C00006	$NADP^+$	C00010	CoA	C00028	Hydrogen acceptor
C00003	NAD^+	C00007	O_2	C00011	CO_2	C00030	Hydrogen donor
C00004	NADH	C00008	ADP	C00014	Ammonia	C00080	H^+

external compounds E were extracted automatically for each set of reactions.

3.2. Measures to evaluate the output network and the algorithm performance

The algorithm performance was analyzed on the basis of 30 runs for each combination of parameters. When results presented an asymmetric distribution, the median and the median absolute deviation were employed as robust estimators to characterize the distribution. The statistical analysis was performed using the Wilcoxon signed-rank test, because it does not assumes normal distribution on the data and the outliers have less effect than on a classical t-test (Derrac et al., 2011).

Two measure groups were used to carry out the evaluations. The first one includes measures that evaluate the algorithm performance such as: N_G, the number of generations employed to find a solution; N_g, number of generations required for all the population to be initialized with the same compound, and F_S, the number of runs where compounds in D are linked for a metabolic pathway. The second group corresponds to measures that evaluate characteristics of the metabolic pathways found. The measures employed to characterize the metabolic pathways are:

- Reactions (N_R): it provides information about the number of steps required to relate the compounds, counting the number of reactions the metabolic pathway has.
- Branching (ρ): it evaluates the relation between the pathway reactions by measuring the mean number of reactions that employ a non-abundant substrate. Compounds belonging to A' (abundant compounds) are not considered to calculate this measure, because the main interest is in the relationships among new compounds produced in the network. The pathway branching is calculated according to

$$\rho(\mathbf{c}) = \frac{1}{|S_f^*|}\sum_{i=1}^{|S_f^*|}\sum_{j=1}^{|\mathbf{c}|}\mathbf{1}_{s_i \subseteq S(r_j)}, \tag{6}$$

  where S_f^* are the substrates of all reactions in $\mathbf{c}$ after filtering the abundant compounds, $|\mathbf{c}|$ is the pathway size, $\mathbf{1}$ is the indicator function, and s_i is the ith compound of the set $S(r_j)$ of substrates for reaction r_j.
- Leaves (λ): it counts the number of compounds produced by the metabolic pathway that are not employed as substrates by any reaction. This measure gives an idea of the degree of specificity the pathway has. A pathway with a high number of leaves indicates that it participates as an intermediary of a great variety of processes; a pathway with a low number of leaves indicates a high specificity for the synthesis of the indicated compounds. Let $S^*(\mathbf{c})$ and $P^*(\mathbf{c})$ be the sets of substrates and products of all reactions encoded in $\mathbf{c}$, respectively, the number of leaves λ is calculated as

$$\lambda(\mathbf{c}) = |P^*(\mathbf{c}) - \left(S^*(\mathbf{c}) \cup A'\right)|. \tag{7}$$

- Difference between metabolic pathways (σ): this measure compares the sequence of compounds used to relate the elements in D and determines the proportion of compounds shared between two pathways. Let d_i and d_j be the initial substrates of the chromosomes $\mathbf{c}_i$ and $\mathbf{c}_j$, respectively, and let $Q_i = (P^*(\mathbf{c}_i) \cap S^*(\mathbf{c}_i) \cup \{d_i\}) - A'$ and $Q_j = \left(P^*(\mathbf{c}_j) \cap S^*(\mathbf{c}_j) \cup \{d_j\}\right) - A'$ be the subsets of compounds belonging each pathway. The difference between both metabolic pathways is calculated as

$$\sigma(\mathbf{c}_i, \mathbf{c}_j) = 1 - \left[\frac{|Q_i \cap Q_j|}{\min(|Q_i|, |Q_j|)}\right]. \tag{8}$$

  Two pathways have a difference $\sigma(\mathbf{c}_i, \mathbf{c}_j) = 0$ when they employ exactly the same compounds to relate the elements in D. This not implies that both are the same pathway, but rather one can be included in the other.

4. Results and discussion

In this section, the proposed algorithm performance is studied in three phases. The first one, presented in Section 4.1, studies the behavior of the algorithm for different parameters and operators. Section 4.2 analyzes the algorithm performance when the set of reactions previously employed is extended. Finally, Section 4.3 presents two case studies, where biological aspects of the solutions found are analyzed and discussed.

Experiments were conducted setting as finalization criteria a fitness equal to 1.0 and a maximum of $G_M = 1000$ generations per search. Populations were initialized with $M = 200$ individuals and a maximum size of chromosome $N_M = 100$ genes, to appropriately explore the solutions space. In every case, the tournament selection strategy was employed with 5 individuals and a crossover probability $p_x = 0.8$, since that value produced the best results in preliminary experiments.

4.1. Sensitivity to parameters and operators

This section presents the performance measures for EvoMS. The effect of the crossover type is analyzed and the influence of the different probabilities that control the mutation operator is evaluated. In the experiments, metabolic pathways relating the compounds L-Glutamate (C00025), Fumarate (C00122), and L-Proline (C00763) were searched for. These particular compounds were selected given their importance in the metabolism, and because only one (C00025) can be used to built a metabolic pathway that links the three compounds. Thus, this situation allows to test the method to determine the initial substrate. The search was carried out using the set of reactions belonging to the arginine and proline reference metabolic pathway (*apdata*).[1] A total of 139 reactions were extracted, 24 of which are reversible (broken down in 48 reactions) and 91 irreversible.

4.1.1. Influence of the crossover type

The EvoMS performance was compared using the standard one-point crossover and the proposed crossover operator. The performance analysis was evaluated in terms of the number of runs

[1] http://www.genome.jp/kegg/pathway/map/map00330.html.

Table 2
Effect of the crossover type on the evolutionary algorithm performance. The median and the median absolute deviation values are provided for N_G and N_g.

	One-point	Modified
F_S	0.83	0.97
N_G	59 ± 27	57 ± 18
N_g	3 ± 0	3 ± 0

that produce a solution F_S, the number of generations required to find a solution N_G, and the number of generations required to obtain a unique subpopulation N_g. Table 2 shows the results obtained with each operator.

The most interesting fact observed in the table is the increase from 0.83 to 0.97 in the fraction of runs that lead to a solution when the modified operator is employed; there are not significant differences in the other measures. This increase can be explained by the way in which metabolic pathways are modeled. Since reactions are stored in the chromosome from left to right, the ones located on the far right are more sensitive to the changes introduced to the sequence, as they depend, to a greater extent, on the products of previous reactions. On the other hand, since the algorithm requires all reactions in the chromosome to be valid, incorporating a higher number of reactions than the one needed to relate the compounds in D translates into an additional effort the algorithm must make to meet this requirement. Therefore, the insertion of only one portion of the genetic material from the second parent decreases the number of reactions that do not probably meet the validity requirement. At the same time, a lower number of generations is required to find a solution.

4.1.2. Variation of mutation probabilities

The proposed mutation operator plays an important role by introducing specific modifications that can change the branching of the metabolic pathway, and favors the exploration of new regions in the search space. Inserting new reactions can lead to the production of compounds necessary for other reactions to occur. Deletion allows to eliminate reactions that can be invalid or redundant. An appropriate balance of these operations can reduce the number of generations required to find the solution. To find the combination of probabilities leading to the best results, the values $p_m = \{0.02, 0.05, 0.08\}$, $p_e = \{0.20, 0.50, 0.80\}$ and $p_v = \{0.20, 0.50, 0.80\}$ were analyzed. Table 3 shows the median and deviation values for the number of generations employed in the runs for a specific set of parameters. The table has three blocks, each one corresponding to one mutation probability. For each block, valid insertion and deletion probabilities are shown in rows and columns, respectively. The combinations of probabilities with which solutions were obtained before 1000 generations and in more than 90% of the runs are indicated with a mark (†).

In general terms, the combinations between deletion and valid insertion probabilities lead to the same tendencies for the three mutation probabilities evaluated. The increase in p_e is accompanied by the reduction in the number of generations required to find a solution, as it is clearly seen when $p_m = 0.05$ and $p_v = 0.5$, where there is a decrease from 102 to 35 generations when the value of p_e is increased. This is to be expected since, during the initialization, a wide variety of reactions are incorporated, most of which should be discarded during the evolution. Thus, the application of mutations favoring the elimination of reactions will improve the algorithm performance. Moreover, although no clear tendency is observed regarding the effect produced by the valid insertion probability, in some cases, it is seen that the increase in p_v is accompanied by a decrease in the number of generations ($p_m = 0.02$ and $p_e = 0.5$).

As regards the mutation probability, it is possible to observe an increasing tendency on cases in which a solution is obtained in more than 90% of the runs with the raise of p_m. This trend might be explained considering two effects produced by the mutation operator: increasing the genetic diversity and contributing to the consolidation of the validity of the sequence of reactions. For that reason, there is an optimum number of insertions that contributes to perform the search in the lowest number of generations. Consequently, a low number of insertions makes the search slower, probably because of the lack of genetic diversity; whereas an excess in the number of insertions leads to the disproportionate increase of the pathways size and makes it difficult to preserve the sequences validity. When the mutation probability is low (few changes in the chromosome), the insertion of new reactions has a more important contribution than deletion (low values for p_e), probably collaborating to the generation of a sequence of valid reactions and introducing genetic diversity. Nevertheless, when the mutation probability increases (a higher number of changes in the sequence), it is necessary to increase the deletion of reactions in order to keep the balance between insertions and the size of the pathways (containing unnecessary reactions). In addition, it should be remembered that these results correspond to runs in which the maximum number of generations is limited. Finally, it is observed that the lowest number of generations employed with each mutation probability (highlighted in bold) is obtained with $p_e = 0.80$ and $p_v = 0.50$, being minimum for $p_m = 0.08$. Besides, this combination of probabilities also provides solutions in 90% of the runs.

4.2. Scalability of the algorithm

In order to study the ability of the algorithm to perform similar searches in spaces that scale in size, the search made in the previous section was performed expanding the set of *apdata* reactions. The new dataset (*sdata*) was built adding the reactions belonging to five reference metabolic pathways.[2] Thus, *sdata* has 443 one-way reactions, 132 of which are reversible (broken down in 264 reactions) and 179 irreversible. Runs were carried out employing the best parameters obtained in Section 4.1.

4.2.1. Algorithm performance and characteristics of the pathways

Table 4 shows the evaluation measures for the searches performed with the two datasets. Blocks separate the performance measures (upper block) from the solutions quality measures (lower block). In general terms, no practical differences are observed in the algorithm performance. In both cases, a solution is obtained in more than 90% of the runs ($F_S > 0.9$). Moreover, although the number of generations is lower when *sdata* is employed, this behavior is only at a tendency level since the confidence intervals are overlapped. Although the value of N_g is increased in one generation, from a practical point of view this difference is not important, as in both cases the winning subpopulation is quickly selected during the first generations.

Regarding the measures associated to the structure of the metabolic pathways, a significant reduction is observed ($p < 0.0001$) in the size of the pathways (N_R) found using *sdata*. This is to be expected since the number of possible connections between compounds is higher and makes possible the existence of smaller alternative paths that connect the compounds in D. The branching ρ calculated for the solutions found with *sdata* supports this explanation, as it experiences a significant increase ($p < 0.005$) from 1.2 to 1.3. The number of leaves λ, indicating that the pathways found with *sdata* include reactions that generate a lower number of

[2] Glycolysis/Gluconeogenesis (map00010), Citrate cycle (map00020), Pentose phosphate pathway (map00030), Pentose and glucuronate interconversions (map00040) and Alanine, aspartate and glutamate metabolism (map00250) in KEGG.

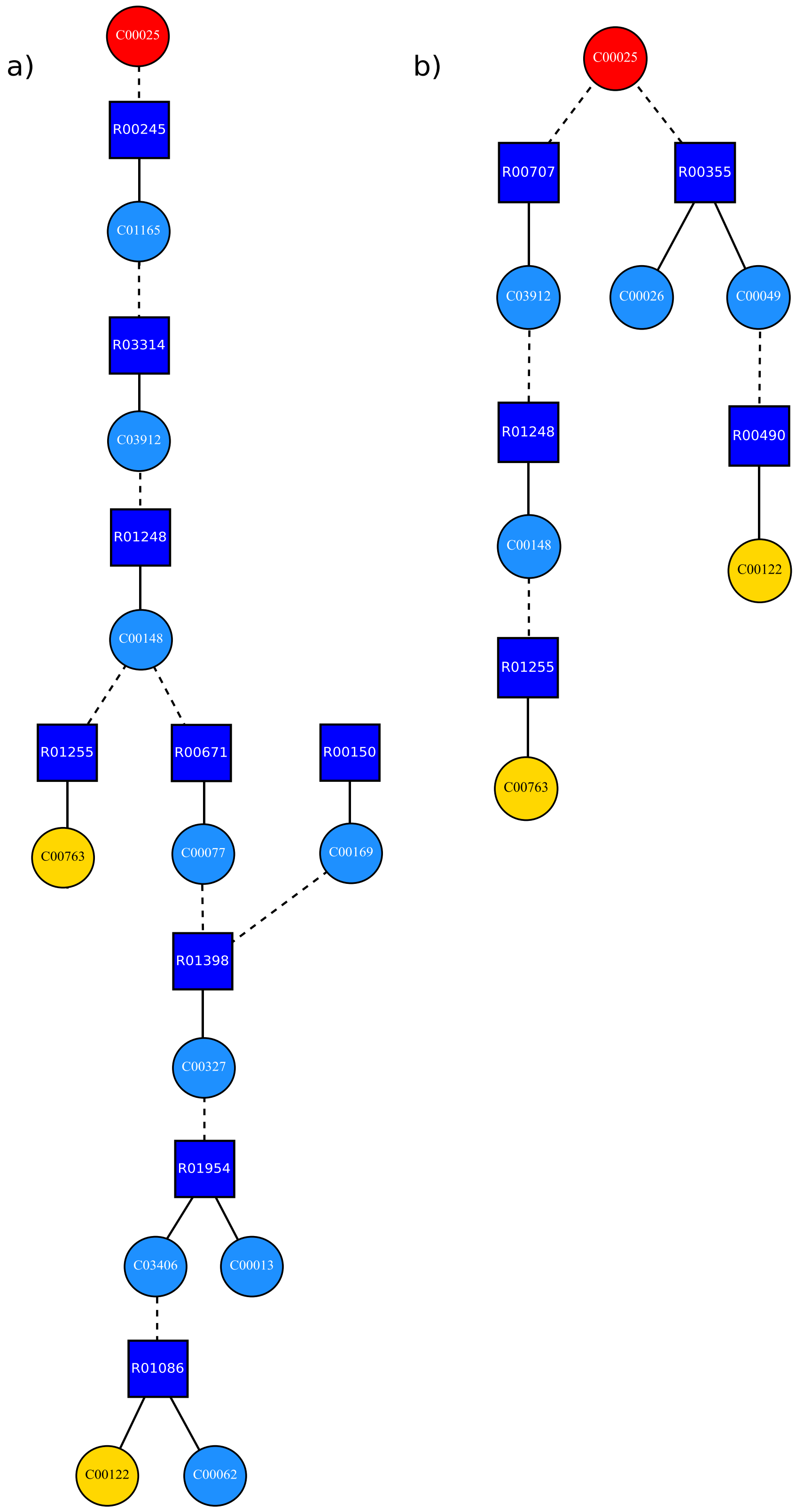

Fig. 3. Pathways belonging to two groups of solutions found with *apdata* and *sdata*, respectively. (a) Examples for: `IVa`, (b) `Vb`. The initial compound is indicated in red (C00025), the compounds to be produced are indicated in yellow (C00122, C00763), and the compounds produced by the metabolic pathway are indicated in light blue. Reactions are indicated as blue squares. Available compounds are not included in the metabolic pathway. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Generations required by EvoMS to find a metabolic pathway employing the initialization with a variable chromosome size. Results correspond to the median values and its deviations.

N_G	$p_m = 0.02$			$p_m = 0.05$			$p_m = 0.08$		
	$p_e = 0.20$	$p_e = 0.50$	$p_e = 0.80$	$p_e = 0.20$	$p_e = 0.50$	$p_e = 0.80$	$p_e = 0.20$	$p_e = 0.50$	$p_e = 0.80$
$p_v = 0.20$	87 ± 37†	72 ± 25	41 ± 10	164 ± 65	60 ± 19†	36 ± 9†	256 ± 130	69 ± 18	39 ± 13†
$p_v = 0.50$	57 ± 18	56 ± 13†	**40** ± **11**	102 ± 42	49 ± 10†	**35** ± **11**	159 ± 80	70 ± 29†	**33** ± 8†
$p_v = 0.80$	72 ± 35†	45 ± 6†	41 ± 13	97 ± 48†	47 ± 17	37 ± 12†	162 ± 95†	62 ± 28†	36 ± 8†

† Experiments where a solution is found before 1000 generations and in more than 90% of the runs. The best results obtained with each mutation probability are highlighted in bold.

unnecessary products ($p < 0.0001$), is possibly due to the use of more specific process reactions. It should be highlighted that, regardless of the branching differences, both sets of reactions lead to solutions with values of ρ higher than the unit, since some compounds in the networks found act as a substrate in more than one reaction.

4.2.2. Difference between solutions

In order to measure if the proposed evolutionary algorithm is capable of reproducing the searches in a solutions space extended by the incorporation of additional reactions, the solutions found with both datasets were studied and compared

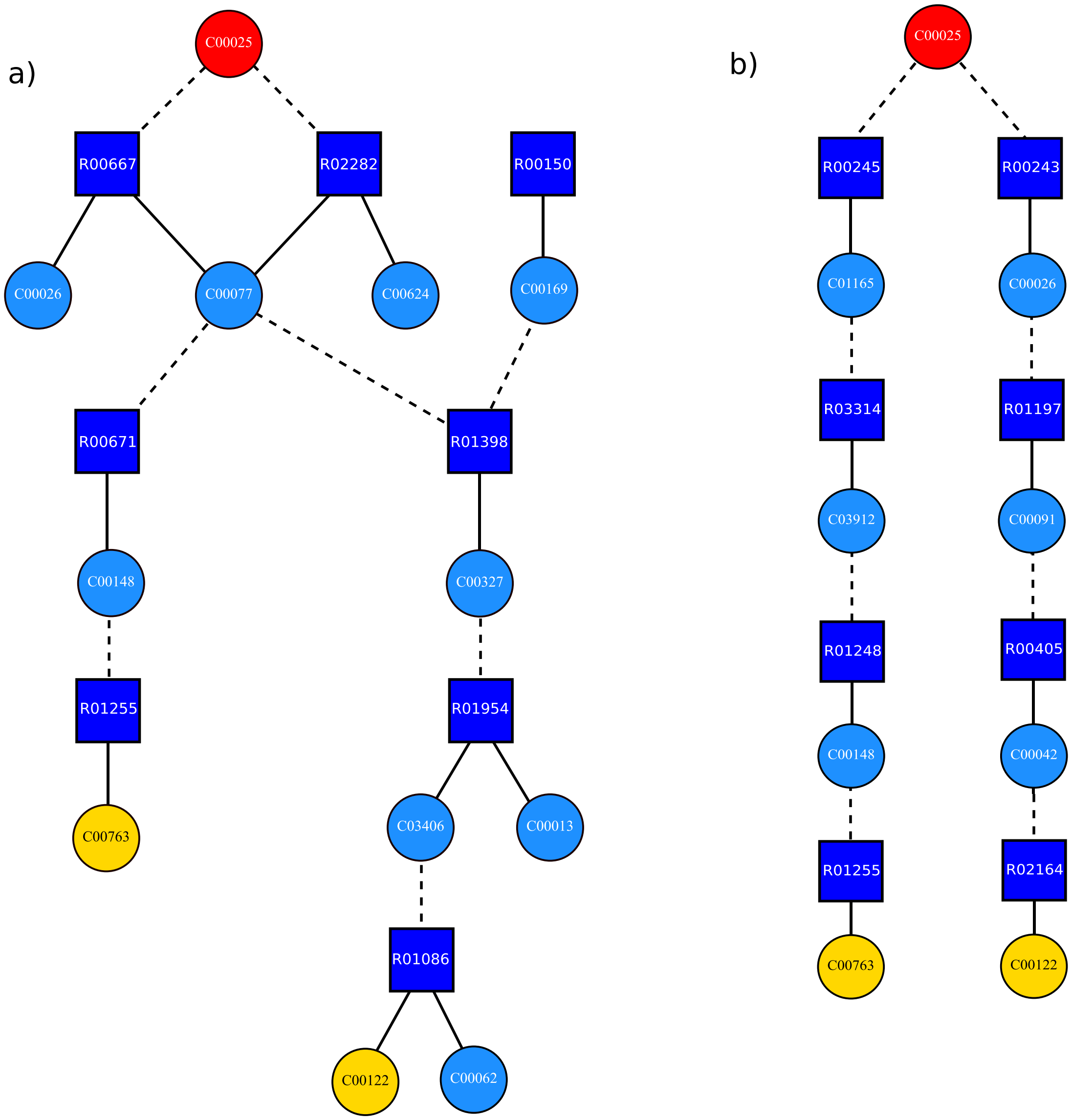

Fig. 4. Pathways belonging to two groups of solutions found with *apdata* and *sdata*, respectively. (a) Examples for: (a) `Ia`, (b) `Ib`. The initial compound is indicated in red (C00025), the compounds to be produced are indicated in yellow (C00122, C00763), and the compounds produced by the metabolic pathway are indicated in light blue. Reactions are indicated as blue squares. Available compounds are not included in the metabolic pathway. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Comparison of the algorithm performance employing the arginine and proline dataset (*apdata*) and extended dataset (*sdata*).

	apdata	*sdata*
F_S	1.00	0.97
N_G	33 ± 8	29 ± 7
N_g	3 ± 0	4 ± 1
N_R	8 ± 1	6 ± 1
ρ	1.2 ± 0.1	1.3 ± 0.1
λ	5 ± 1	4 ± 1

to determine the number of novel metabolic pathways in common.

Typically, synthesizing a compound implies a number of steps until the desired product is reached. Thus, a sequence of several intermediate compounds linking the initial substrate and the final product is generated. However, in many cases, those intermediate compounds can be produced by more than one reaction. This leads to metabolic pathways which are different in terms of reactions, but equivalent in terms of the sequence of compounds needed to perform the synthesis. According to (8), two metabolic pathways $\mathbf{c}_1$ and $\mathbf{c}_2$ will be equivalent when they have a difference value $\sigma(\mathbf{c}_1, \mathbf{c}_2) = 0.0$. Furthermore, this measure will increase when the number of shared compounds decreases.

In a preliminary analysis, five *groups of equivalent solutions* were found for *apdata* and seven for *sdata*. Table 5 shows the difference values between the groups found with both sets of reactions. Rows and columns indicate the group of equivalent solutions for *apdata* and *sdata*, respectively. The intersection between a row and a column indicates the difference between the groups considered. It can be seen that some groups of solutions are equivalent, as it could be expected, since *apdata* and *sdata* share the mechanisms to synthesize the three specified compounds. For instance, group VIIb does not show differences with any of the solutions found with *apdata*. This is because the five groups of equal solutions found with *apdata* employ the same sequence of compounds that the group VIIb, together with other additional compounds. The group of solutions IVb also shows a similar behavior, presenting a difference of 0.29 with all *apdata* solutions, which indicates that it shares a portion of the sequence of compounds.

In order to analyze the differences found in more detail, two groups of solutions with extreme difference values were selected and one metabolic pathway representing each one was plotted. Fig. 3 shows the metabolic pathways corresponding to VIa and Vb groups, while Fig. 4 contains the pathways of Ia and Ib groups. In every case, the pathways must be interpreted in a descending manner, starting by the initial substrate C00025 (in red), and descending through the sequence of reactions and until each one of the final products (in yellow). Representations are simplified, not showing abundant compounds.

Pathways representing solutions of groups IVa and Vb in Fig. 3 do not show any difference according to (8). Clearly, pathway

Table 5
Values of the difference between groups of equivalent solutions found with *apdata* and *sdata*. Difference values lower than 0.15 are highlighted in bold.

		sdata						
		Ib	IIb	IIIb	IVb	Vb	VIb	VIIb
apdata	Ia	0.50	0.43	0.38	0.29	0.20	**0.13**	**0.00**
	IIa	0.50	0.43	0.25	0.29	0.20	0.22	**0.00**
	IIIa	0.38	0.43	0.38	0.29	0.20	0.22	**0.00**
	IVa	0.25	0.29	0.38	0.29	**0.00**	0.22	**0.00**
	Va	0.25	0.29	0.38	0.29	**0.00**	**0.11**	**0.00**

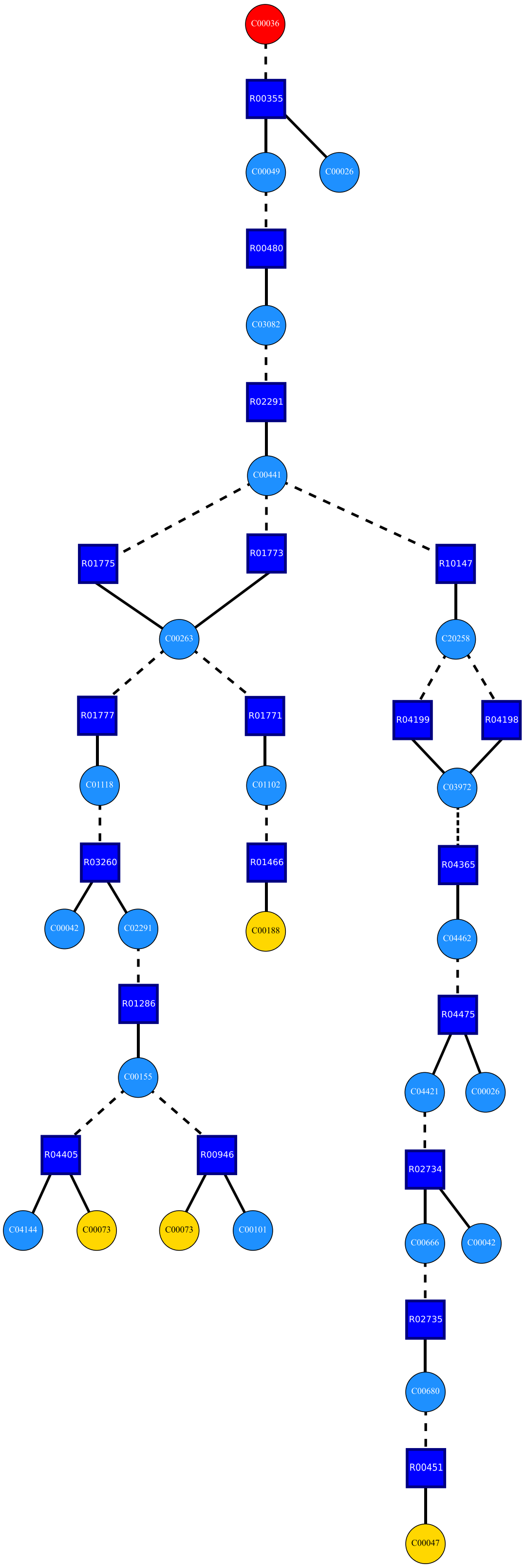

Fig. 5. Reference metabolic pathway involving lysine, threonine and methionine biosynthesis. Note that reaction R00946 and R04405 produce the same compound C00073 in two different ways. Initial substrate is in red and the compounds to be produced are indicated in yellow. Reactions are indicated in blue, and their substrates are products are in dashed and solid lines, respectively. To provide a clearest view, only the more relevant compounds are shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

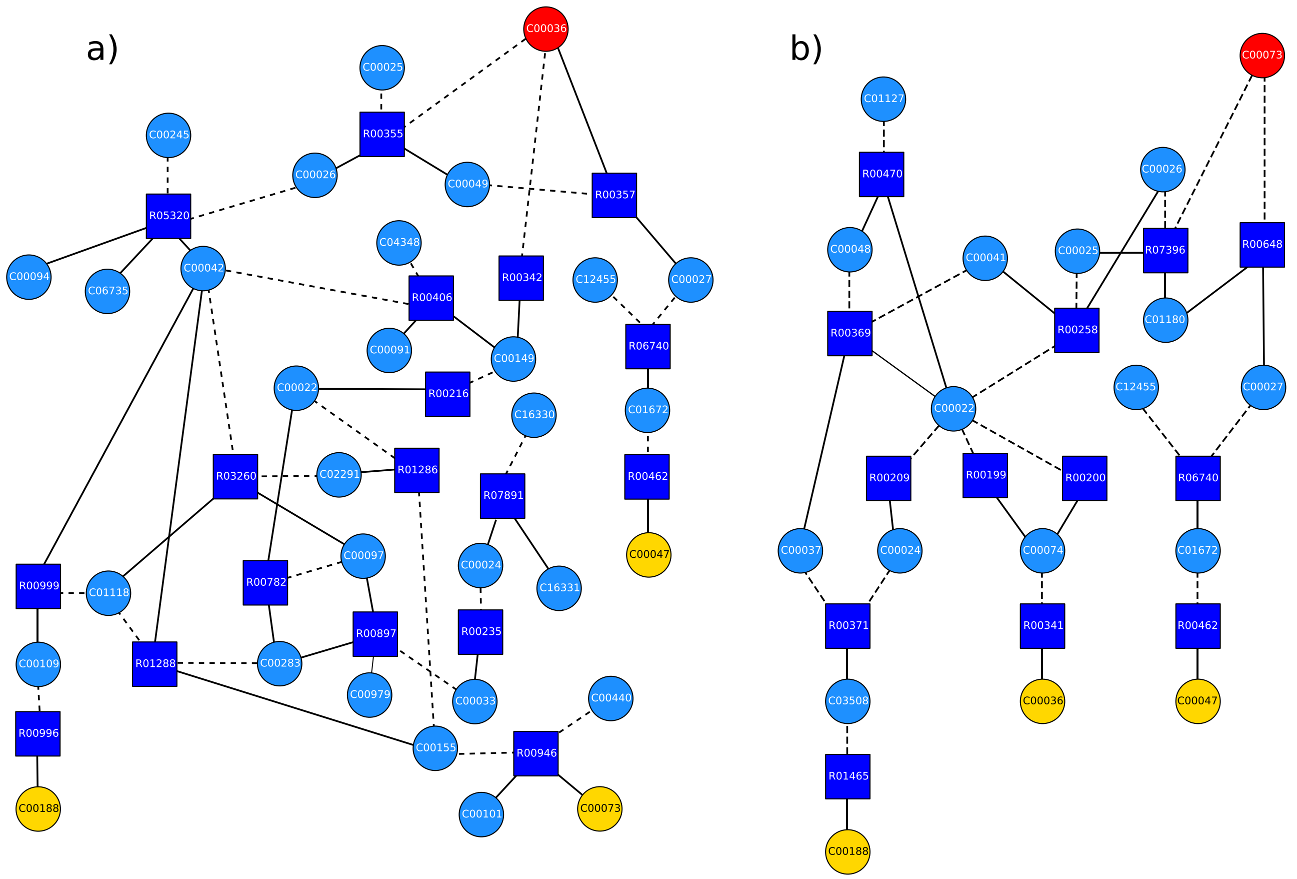

Fig. 6. New metabolic pathways linking compounds C00036, C00047, C00073 and C00188. (a) Metabolic pathway found by EvoMS with C00036 as initial substrate. (b) Metabolic pathway found by EvoMS with C00073 as initial substrate. In every case, the initial substrate is indicated in red and the compounds to be produced are indicated in yellow. Reactions are indicated in blue. Substrates are connected with dashed lines and products with solid ones. To provide a clearest view, only the more relevant compounds are shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

from IVa employs almost twice the reactions as Vb to relate the same compounds. However, analyzing in detail the sequences of compounds used by both pathways, it is observed that the compounds used by Vb (Q_{Vb} = {C00025, C03912, C00148, C00763, *C00049*, C00122}) are also employed by IVa (Q_{VIa} = {C00025, C01165, C03912, C00148, C00763, C00169, C00077, C00327, C03406, C00122}). Although the compound C00049 (which in the sequence is indicated in italics) is not shared by the pathways, it should not be considered to calculate the differences since it is part of the set of abundant compounds. As a consequence, both solutions relate members of *D* employing the same compounds.

When analyzing the solutions from Ia and Ib (Fig. 4) it can be seen that both contain the same number of reactions. However, the sequence of compounds used by Ia (Q_{Ia} = {C00025, C00077, C00148, C00763, C00169, C00327, C03406, C00122}) presents a large difference compared to the one employed by Ib (Q_{Ib} = {C00025, C01165, C03912, C00148, C00763, C00026, C00091, C00042, C00122}). On the one hand, the sequences of compounds used to synthesize C00763 from C00025 only share the intermediary compound C00148, which is produced through different reactions in each solution. On the other hand, the C00122 synthesis is carried out using sequences of completely different compounds in both metabolic pathways, sharing only the compounds in the extremes. Thus, only 4 compounds (C00148 and the three compounds to be related) are shared between Ia and Ib, leading to a difference $\sigma = 0.5$ according to (8).

4.3. Case study: searching for relations between 4 compounds

The EvoMS performance in a more complex real problem was evaluated and compared with a state-of-the-art algorithm (Faust et al., 2011) for searching a metabolic pathway relating 4 specific compounds. The search involved the complete set of reactions stored in KEGG for the *Escherichia coli* bacterium metabolism. After the pre-processing, the search space was made up of 2354 reactions, 1061 of which were reversible (broken down in 2122 one way reactions) and 232 irreversible. The reference pathway for lysine, threonine, and methionine biosynthesis (Fig. 5) was taken as a case study of a branched metabolic pathway. It synthesizes compounds C00047 (L-Lysine), C00073 (L-Methionine), and C00188 (L-Threonine) from C00036 (Oxaloacetate).

The algorithm of Faust et al. (2011) combines several linear paths to build a network of relationships among compounds. It performs the search of the shortest path between each pair of compounds and combines all of them into a network. With this approach, the authors were able to find a pathway for the compounds using a high proportion (85%) of the reactions belonging to the reference metabolic pathway. In comparison, EvoMS was able to find a pathway with all the reactions (100%) of the reference pathway. Furthermore, another important advantage is that feasibility of the solutions found by EvoMS is guaranteed. EvoMS builds the pathway by verifying that all reactions use available substrates, while the other algorithm does not even take into account that information during the search.

Besides the reference pathway, Fig. 6 shows two other examples of metabolic pathways synthesized by EvoMS, for the same search. In both cases, solutions were fully feasible and allowed to relate the same 4 compounds. Fig. 6a shows the metabolic pathway found with C00036 as initial substrate, containing four reactions also present in the reference pathway (R00355, R03260, R01286 and R00946). It must be noted that reaction R03260 plays a central function in the new pathway, producing two key compounds (C01118 and C00097) needed to synthesize C00073 and C00188. Also, it can be appreciated that the initial substrate has a key role in this pathway, being a precursor to synthesize C00027 (needed for C00047), and C00042 (needed for C00073 and C00188). Furthermore, the large number of interconnections among reactions in this pathway shows an important collaborative work to synthesize the final products.

Fig. 6b presents another alternative metabolic pathway that is also fully feasible and relates the same compounds. At a glance, it can be observed that this novel pathway could be more efficient to link the 4 compounds of interest than the previous one, because it requires a lower number of reactions to relate them. This solution uses C00073 as initial substrate, not sharing any reaction with the reference pathway. The novel pathway is built by two main branches starting from C00073, one of which produces C00047 and the other produces the remaining two products. As it can be seen, C00022 plays a key role as precursor in the synthesis of C00036 and C00188. Similarly to the pathway in Fig. 6a, C00022 in this novel pathway allows to infer a relation between the glycolysis (a reference pathway for many life forms) and the synthesis of both products. These examples evidence the natural interconnections present among metabolic pathways in living organisms. This also highlights the importance of developing new algorithms for searching on large sets of reactions, providing branched metabolic pathways of feasible reactions that relate multiple compounds simultaneously.

5. Conclusions and future work

This work approached the problem of searching for metabolic pathways that relate a set of compounds through networks of feasible reactions. A model to build the pathways based on a set of compounds was proposed and a new evolutionary algorithm, called EvoMS was developed to search for the reactions required to build pathways between specific compounds. Also, new operators and an initialization strategy that employ the set of compounds model were developed. The fitness function was designed to evaluate essential characteristics required in the metabolic pathways searched, in order to find feasible metabolic pathways. The tests carried out for a real problem showed that EvoMS was capable of reproducing known metabolic pathways and also finding alternative connections to synthesize the same final compounds. In all searches, the algorithm found branched metabolic pathways made up of feasible reactions from the initial compounds indicated. Besides, in cases where reactions require compounds that do not belong to the abundant ones, the algorithm was capable of previously incorporating reactions to generate them. In summary, the possibility of generating a wide range of connections between compounds, together with the ability to provide feasible solutions makes EvoMS a simple and powerful method to find feasible networks connecting metabolic compounds. Moreover, flexibility of the evaluation function allows to easily extend it to incorporate new objectives to optimize in the solution.

Future work will aim to improve the search process by adding information to the evaluation function, for example, regarding the stoichiometry and thermodynamics of the reactions, the degree of connectivity of compounds, and/or the availability of enzymes. In addition, the crossover operator will be modified to employ information of the compounds used by the metabolic pathway, and mechanisms to automatically adjust the parameters of the algorithm during the evolution will be studied.

The full source code for EvoMS algorithm is available for free academic use at http://sourceforge.net/projects/sourcesinc/files/evoms/. A web-interface to run the evolutionary algorithm proposed in this work is available online at http://fich.unl.edu.ar/sinc/web-demo/evoms/, whose main inputs, outputs, features and ways of use are explained in Gerard et al. (2015).

Acknowledgements

This work was supported by National Scientific and Technical Research Council (CONICET) [PIP 2013-2015 117], Universidad Nacional del Litoral (UNL) [CAI+D 2011 548] and Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT) [PICT 2014 2627].

References

Affenzeller, M., Winkler, S., Wagner, S., Beham, A., 2009. Genetic Algorithms and Genetic Programming: Modern Concepts and Practical Applications. CRC Press.

Altman, T., Travers, M., Kothari, A., Caspi, R., Karp, P., 2013. A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinform. 14, 112.

Arita, M., 2012. Metabolic reactions to networks and pathways. In: Bacterial Molecular Networks: Methods and Protocols, Volume 804 of Methods in Molecular Biology. Springer, pp. 93–106.

Bäck, T., Fogel, D., Michalewicz, Z., 2000. Evolutionary Computation I: Basic Algorithms and Operators. Institute of Physics Publishing.

Bordbar, A., Monk, J.M., King, Z.A., Palsson, B.O., 2014. Constraint-based models predict metabolic and associated cellular functions. Nat. Rev. Genet. 15, 107–120.

Boussaïd, I., Lepagnot, J., Siarry, P., 2013. A survey on optimization metaheuristics. Inf. Sci. 237, 82–117.

Chen, C.P., Zhang, C.-Y., 2014. Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. Inf. Sci. 275, 314–347.

Croes, D., Couche, F., Wodak, S., van Helden, J., 2005. Metabolic pathfinding: inferring relevant pathways in biochemical networks. Nucl. Acids Res. 33, W326–W330.

de Magalhães, C.S., Almeida, D.M., Barbosa, H.J.C., Dardenne, L.E., 2014. A dynamic niching genetic algorithm strategy for docking highly flexible ligands. Inf. Sci. 289, 206–224.

Derrac, J., García, S., Molina, D., Herrera, F., 2011. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. Swarm Evol. Comput. 1, 3–18.

Faust, K., Croes, D., van Helden, J., 2009. Metabolic pathfinding using RPAIR annotation? J. Mol. Biol. 388 (2), 390–414.

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

Faust, K., Croes, D., van Helden, J., 2011. Prediction of metabolic pathways from genome-scale metabolic networks. Biosystems 105, 109–121.

Garai, G., Chowdhury, B., 2015. A cascaded pairwise biomolecular sequence alignment technique using evolutionary algorithm. Inf. Sci. 297, 118–139.

Gerard, M., Stegmayer, G., Milone, D., 2013. An evolutionary approach for searching metabolic pathways. Comput. Biol. Med. 43, 1704–1712.

Gerard, M., Stegmayer, G., Milone, D., 2015. EvoMS: an evolutionary tool to find de novo metabolic pathways. Biosystems 134, 43–47.

Gerlee, P., Lizana, L., Sneppen, K., 2009. Pathway identification by network pruning in the metabolic network of *Escherichia coli*. Bioinformatics 25, 3282–3288.

Goto, S., Okuno, Y., Hattori, M., Nishioka, T., Kanehisa, M., 2002. LIGAND: database of chemical compounds and reactions in biological pathways. Nucl. Acids Res. 30, 402–404.

Heath, A., Bennett, G., Kavraki, L., 2010. Finding metabolic pathways using atom tracking. Syst. Biol. 26, 1548–1555.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z., Barabási, A., 2000. The large-scale organization of metabolic networks. Nature 407, 651–654.

Kayaa, H., Şule Gündüz-Öğüdücü, 2013. SAGA: a novel signal alignment method based on genetic algorithm. Inf. Sci. 228, 113–130.

Lacroix, V., Cottret, L., Thebault, P., Sagot, M.-F., 2008. An introduction to metabolic networks and their structural analysis. IEEE/ACM Trans. Comput. Biol. Bioinform. 5 (4), 594–617.

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

Lee, W.-P., Hsiao, Y.-T., 2012. Inferring gene regulatory networks using a hybrid GA–PSO approach with numerical constraints and network decomposition. Inf. Sci. 188, 80–99.

McShan, D., Rao, S., Shah, I., 2003. PathMiner: predicting metabolic pathways by heuristic search. Bioinformatics 19, 1692–1698.

Ogata, H., Goto, S., Fujibuchi, W., Kanehisa, M., 1998. Computation with the KEGG pathway database. Biosystems 47, 119–128.

Pal, S., Bandyopadhyay, S., Ray, S., 2006. Evolutionary computation in bioinformatics: a review. IEEE Trans. Syst. Man Cybern. 36, 601–615.

Pey, J., Prada, J., Beasley, J., Planes, F., 2011. Path finding methods accounting for stoichiometry in metabolic networks. Genome Biol. 12, R49.

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

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.-L., 2002. Hierarchical organization of modularity in metabolic networks. Science 297, 1551–1555.

Russell, S., Norvig, P., 2010. Artificial Intelligence: A Modern Approach, 3rd ed. Prentice Hall.

Tenazinha, N., Vinga, S., 2011. A survey on methods for modeling and analyzing integrated biological networks. IEEE/ACM Trans. Comput. Biol. Bioinform. 8, 943–958.

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., Dien, S.V., 2011. Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. Nat. Chem. Biol. 7, 445–452.