Computers in Biology and Medicine 43 (2013) 1704–1712

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

An evolutionary approach for searching metabolic pathways

Matias F. Gerard [a,b,*], Georgina Stegmayer [a,b], Diego H. Milone [b]

[a] Center for Research and Development of Information Systems (CIDISI), National Scientific and Technical Research Council (CONICET), Argentina
[b] Research Center for Signals, Systems and Computational Intelligence (Sinc(i)), FICH-UNL, National Scientific and Technical Research Council (CONICET), Argentina

ARTICLE INFO

Article history:
Received 10 September 2012
Accepted 21 August 2013

Keywords:
Search strategies
Evolutionary algorithms
Metabolic pathways

ABSTRACT

Searching metabolic pathways that relate two compounds is a common task in bioinformatics. This is of particular interest when trying, for example, to discover metabolic relations among compounds clustered with a data mining technique. Search strategies find sequences to relate two or more states (compounds) using an appropriate set of transitions (reactions). Evolutionary algorithms carry out the search guided by a fitness function and explore multiple candidate solutions using stochastic operators. In this work we propose an evolutionary algorithm for searching metabolic pathways between two compounds. The operators and fitness function employed are described and the effect of mutation rate is studied. Performance of this algorithm is compared with two classical search strategies. Source code and dataset are available at https://sourceforge.net/projects/sourcesinc/files/eamp/

1. Introduction

Using search strategies to solve different problems is common in many areas of knowledge. In many cases, employing classical strategies for sequential state space exploration allows to find solutions rapidly. When all possible solutions are exhaustively explored the strategies are called uninformed search strategies, and this is the case of breadth first search (BFS) and depth first search (DFS) algorithms [1]. It is a well-known fact that there are problems in which a very high number of solutions must be explored, making classical methods practically inapplicable. For example, in KEGG database [2] there are around 17,000 compounds with approximately 14,000 connections among them, and a high branching factor. There are different approaches to address these problems, among which evolutionary algorithms (EAs) have an important place. The main difference with DFS and BFS is that EAs do not explore the state space exhaustively, but rather use several heuristics to select the most promising regions to explore. These methods are grouped in four families: Genetic Algorithms [3], Evolutionary Strategies [4], Genetic Programming [5] and Evolutionary Programming [6]. Each one was originated by different motivations, and they differ mainly by their representation schemes, and operators of selection and reproduction [7]. Although the convergence for genetic algorithms is guaranteed by the schema theorem [8], real values codification is limited by the number of bits used. Instead, the evolutionary strategies directly use real values to encode the problem variables, but their convergence depends on the operators used. Evolutionary algorithms use stochastic search based on the evolution of a population of candidate solutions, applying a set of operators and a fitness function that evaluates the quality of the solutions generated. Some interesting aspects about these techniques are the simplicity of the operators used, the possibility of using fitness functions with very few formal requirements, and the ability to explore multiple points of the search space in each iteration [9]. These characteristics make them an attractive alternative to deal with several problems in biology [10–12].

Different search strategies to find metabolic pathways that relate compounds have been recently proposed. The algorithm described by Ogata et al. [13] is based on BFS and builds pathways between pairs of compounds by the combination of allowed relations (metabolic reactions). The method of Linked Metabolites [14] first builds an integrated graph and then performs the pathway search specifying a maximum number of reactions between source and target compounds. Metabolic PathFinding Tool [15] assigns to each operator a cost equal to the number of reactions where the compound participates. McShan et al. [16] use the A* search algorithm to explore the solutions space guided by a cost function based on the Manhattan distance and a heuristic function that uses structural information of compounds to generate characteristic descriptors. A more recent algorithm based on BFS is proposed by Heath et al. [17], where a metabolic pathway linking two compounds is found preserving a specified number of atoms

* Corresponding author at: Research Center for Signals, Systems and Computational Intelligence (Sinc(i)), FICH-UNL, CONICET, Argentina. Tel: +54 342 457 5234x119.
E-mail addresses: mgerard@santafe-conicet.gov.ar (M.F. Gerard), gstegmayer@santafe-conicet.gov.ar (G. Stegmayer), d.milone@ieee.org (D.H. Milone).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.08.017

(atom tracking) between the beginning and ending compounds. However, these last two algorithms require information about the molecular structure of the compounds to be used, and BFS-based methods require a significant amount of memory to store all the search tree. Furthermore, given that methods based on BFS look for a specified number of pathways, the order in which the nodes of the tree are visited can bias the search to particular solutions (unless the successors are selected using randomized traversals). Another alternative is based on elementary modes. These methods use stoichiometry of reactions and several restrictions to identify minimal sets of enzymes that can operate at steady state, with all irreversible reactions used in the appropriate direction [18,19]. Computing all elementary modes is expensive, even for small networks [20]. There also exist retrosynthesis-based methods that build new metabolic pathways to produce a compound of interest in an organism [21]. These methods begin with the desired compound and use reverse enzymatic reactions to synthesize a metabolic pathway from simpler compounds. Although the problem being addressed in those works is close to our own, their objective differs from the one proposed in this paper.

Finding relationships between two given compounds is not an easy task, in particular when data come from different sources such as metabolic and transcriptional profiles.[1] In this case, one possible approach is to create clusters from the combined sources using a data mining tool [22]. This way, applying the "guilt-by-association" principle [23,24] genes and metabolites that vary coordinately can be found. Since the relationship among metabolites and transcripts is mainly given by metabolic pathways,[2] the following step would be searching such pathways with the available data. Traditionally, metabolic pathways search was manually performed, but the current increase in the volume of data demands for computational tools to perform the search automatically [17]. Many efforts have been made to automate the process, but obtained results are not biologically feasible. For example, in [25] a search for metabolic pathways with up to 9 reactions among glucose and pyruvate was performed and approximately 5×10^5 metabolic pathways were found, many of them biologically not possible.

The main contribution of this work is the proposal of an evolutionary algorithm to find metabolic pathways, capable of relating two compounds in a common and valid reaction chain. To achieve this, a data mining tool was used to generate clusters from a real biological dataset, and pairs of compounds within the clusters were used for the genetic search of metabolic pathways. Afterwards, objective measures were defined to quantify the performance of the algorithms, and the effect of the mutation rate on the evolution was studied. Finally, the proposed algorithm was compared with two methods based on classical search algorithms.

The paper is organized as follows: Section 2 describes the proposed algorithm for the evolutionary search of metabolic pathways between two compounds. The data used, the objective measures, and the results obtained are briefly described in Section 3. Finally, Section 4 presents the conclusions of this work.

2. Proposed algorithm

This section presents the proposed algorithm, that we will call evolutionary algorithm for the search of metabolic pathways (EAMPs)[3]. First, the state space and search operators employed are defined. Then, the structure of the chromosomes and the way the information is coded is presented. Afterwards, the genetic operators used and their functioning are described. Finally, the fitness function employed is presented, the terms that compose it are analyzed and the effect that each of them produces on the search is described.

There are different approaches to explore the space of all the possible metabolic pathways linking two specific compounds. One proposal consists of generating a list of compounds that must be excluded from the search [26]. However, incorrect definitions can exclude compounds necessary to produce results of biological interest. A different approach was proposed in [27] where sets of "substrate-product" binary relations were used to represent the reactions and each relation was labeled according to its function inside the reaction. The main stream of the pathways was built using only the relations containing information about the transformation of the substrates.

Following that idea, the state space is defined as the set C of all metabolic compounds in the KEGG database. This database contains information of genes, proteins and metabolic compounds of hundreds of different organisms and the allowed binary relations between compounds are describe by transformations r. The compound on which the transformation is applied will be called substrate s, and p will be the product or new resulting state. Transformations will be represented as ordered pairs $r_i = (s_i, p_i)$, with $s_i, p_i \in C$ and $s_i \neq p_i$. In addition, the substrate and product of r_i will be identified using the notation s_i and p_i respectively, being $\hat{s}$ the initial compound and $\hat{p}$ the final compound of the metabolic pathway. In this way, a metabolic pathway is built as a sequence of transformations that produce $\hat{p}$ starting from $\hat{s}$. Finally, the sequence of possible states $\mathbf{q} = [\hat{s}, p_1, p_2, \ldots, \hat{p}]$ is defined as the sequence of compounds that take part in the transformation.

2.1. Structure of the chromosomes

The sequence of transformations r leading to the production of $\hat{p}$ from $\hat{s}$ is coded in the chromosome as $\mathbf{c} = [r_1, r_2, \ldots, r_i, \ldots, r_N]$, where N indicates the number of genes and the sequence is read from left to right. In this context, the term *chromosome* indicates a data structure such as a vector, and should not be interpreted in a biological way. This value can vary in the range $[1, N_M]$, where N_M is the maximum number of reactions the metabolic pathway can contain. When the number of reactions exceeds this level, the chromosome truncates to contain only the first N_M reactions.

2.2. Genetic operators

This section describes the genetic operators[4] designed for the EAMP. Due to the requirements of this application in particular, it has been necessary to make various changes to classical genetic operators, which, if directly applied, would limit the convergence of the algorithm. In order to facilitate their explanation, four sets of transformations are defined. R^* contains the complete set of allowed transformations, $R^1 = \{r_i \backslash r_i = (\hat{s}, p_i)\} \wedge R^1 \subset R^*$ contains only those transformations that use $\hat{s}$, $R^N = \{r_i \backslash r_i = (s_i, \hat{p})\} \wedge R^N \subset R^*$ contains all transformations that produce $\hat{p}$, and $R^+ = R^1 \cup R^N$ contains the union of the two previous sets. The algorithm finds a solution when it reaches a predefined maximum number of generations or when the fitness of an individual takes the value 1, indicating that it encodes a metabolic pathway that relates the indicated compounds.

[1] Metabolic profile: measurement of concentration levels of small molecules. Transcriptional profile: measurement of activity levels of a set of genes.

[2] A metabolic pathway is a sequence of chemical reactions that transform a substrate into one or various products through a series of intermediary compounds.

[3] Source code and dataset are available at http://sourcesinc.sourceforge.net/eamp/

[4] This general term indicates operations applied over chromosomes. For example, the crossover operator combines genetic information of two chromosomes to produce a new one.