Computational and Structural Biotechnology Journal 13 (2015) 277–285

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Comparison of Metabolic Pathways in *Escherichia coli* by Using Genetic Algorithms

Patricia Ortegon [a,1], Augusto C. Poot-Hernández [a,b,1], Ernesto Perez-Rueda [b,c,*], Katya Rodriguez-Vazquez [a,**]

[a] Departamento de Ingeniería de Sistemas Computacionales y Automatización, IIMAS, Universidad Nacional Autónoma de México, Mexico
[b] Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico
[c] Unidad Multidisciplinaria de Docencia e Investigación, Sisal Facultad de Ciencias, Sisal, Yucatán, UNAM, Mexico

ARTICLE INFO

Article history:
Received 27 January 2015
Received in revised form 31 March 2015
Accepted 1 April 2015
Available online 9 April 2015

Keywords:
Metabolism
Genetic algorithms
KEGG database
k-medoids
Comparative genomics

ABSTRACT

In order to understand how cellular metabolism has taken its modern form, the conservation and variations between metabolic pathways were evaluated by using a genetic algorithm (GA). The GA approach considered information on the complete metabolism of the bacterium *Escherichia coli* K-12, as deposited in the KEGG database, and the enzymes belonging to a particular pathway were transformed into enzymatic step sequences by using the breadth-first search algorithm. These sequences represent contiguous enzymes linked to each other, based on their catalytic activities as they are encoded in the Enzyme Commission numbers. In a posterior step, these sequences were compared using a GA in an all-against-all (pairwise comparisons) approach. Individual reactions were chosen based on their measure of fitness to act as parents of offspring, which constitute the new generation. The sequences compared were used to construct a similarity matrix (of fitness values) that was then considered to be clustered by using a *k*-medoids algorithm. A total of 34 clusters of conserved reactions were obtained, and their sequences were finally aligned with a multiple-sequence alignment GA optimized to align all the reaction sequences included in each group or cluster. From these comparisons, maps associated with the metabolism of similar compounds also contained similar enzymatic step sequences, reinforcing the *Patchwork Model* for the evolution of metabolism in *E. coli* K-12, an observation that can be expanded to other organisms, for which there is metabolism information. Finally, our mapping of these reactions is discussed, with illustrations from a particular case.

1. Introduction

The study of the evolution of metabolism is fundamental to understanding the adaptive processes of cellular life, the emergence of high levels of organization (multicellularity), the diversity of cellular organization in the three major domains of life, *Archaea*, *Bacteria*, and *Eukarya*, and the complexity of the living world [4]. At present, the large scale of information derived from genomics and proteomics studies has allowed the construction of diverse databases devoted to organizing the metabolic processes, such as the KEGG [21], and MetaCyc [5]. Therefore, the information contained in these databases can be used to generate an integrative perspective on cellular functioning.

Metabolism can be considered one of the most ancient biological networks; in such a network, the nodes represent substrates or enzymes and the edges represent the relationships among them. From a global perspective, the comparative analysis of metabolic pathways aims to identify similarities and differences among them, providing insights for the identification of evolutionary events, such as enzyme recruitment and duplication events. In this regard, metabolic pathways exhibit high retention of duplicates within functional modules and a preferential biochemical coupling of reactions. This retention of duplicates may result from the biochemical rules governing substrate-enzyme–product relationships [1,8,14,19].

In this context, diverse studies have evaluated the variations among pathways, both intra- and interspecies [7,27], comparing the pathways based on their Enzyme Commission (EC) numbers and excluding information on the compounds. In addition, a method of path-and-graph matching has been proposed to query metabolic pathways based on a predefined graph, where a similarity measure based on EC numbers [30] and the distance between pathways as a combination of distances between compounds and between enzymes associated with amino acid biosynthesis networks are considered [11].

* Correspondence to: E. Perez-Rueda, Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico.
** Correspondence to: K. Rodriguez-Vazquez, Departamento de Ingeniería de Sistemas Computacionales y Automatización, IIMAS, Universidad Nacional Autónoma de México, Mexico.
E-mail addresses: erueda@ibt.unam.mx (E. Perez-Rueda), katya.rodriguez@iimas.unam.mx (K. Rodriguez-Vazquez).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.csbj.2015.04.001

In this work, we evaluated whether there are groups of similar reactions in different pathways, which might suggest a transfer of enzymatic activities, and whether these groups can be used to define common and variable regions of an organism's metabolism. This analysis was addressed using EC numbers, coded as a succession of reaction steps. To this end the metabolic maps of the bacterium *Escherichia coli* K-12, as deposited in the KEGG database, were transformed into linear Enzymatic Step Sequences (ESS), to be compared via a genetic algorithm (GA). The sequences compared were used to construct a similarity matrix to identify groups of conserved reactions based on a *k*-medoids clustering analysis, and then a multiple-sequence alignment (MSA) GA was optimized to align all the reaction sequences included in a group. Finally, we consider our comparisons in terms of the clues they provide in reinforcing the *Patchwork Model* in the evolution of metabolism for *E. coli* K-12 and probably for other organisms beyond this bacterium.

2. Methods

2.1. Construction of Enzymatic Step Sequences (ESS)

The KGML files (version 0.71) that describe the metabolic maps (pathways) of *E. coli* K-12 as of June 2011 were downloaded from the KEGG database (Fig. 1). Pathways were transformed into linear ESS by using the breadth-first search (BFS) algorithm [26], which infers the closer neighbor of each enzyme by considering a common compound, a substrate or product. In brief, a directed graphical representation of each metabolic map was created in which the nodes represented enzymes and the edges represented a shared substrate/product between two enzymes. This representation takes into account the reversibility of the reactions. Then, a group of BFS trees was generated for each metabolic map from a set of initialization nodes, which were used as roots. In this work, an initialization node was defined by two criteria: (i) a node whose substrate is not catalyzed by another enzyme in the metabolic map, and (ii) a node whose substrate comes from another metabolic map and has two or fewer neighbors in the graph. These criteria represent the metabolic input for each pathway; the first criterion considers the substrates not created in the same pathway, and the second one considers the connections with other pathways. Each initialization node was used as a root for the construction of a BFS tree. Thus, each tree was used as a guide for the construction of the corresponding ESS. In this way, a BFS tree creates as many ESS as the number of branches it contains. Finally, the first three levels of EC numbers are used to represent an enzyme as a string or sequence (Fig. 1). ESS constructed per metabolic map

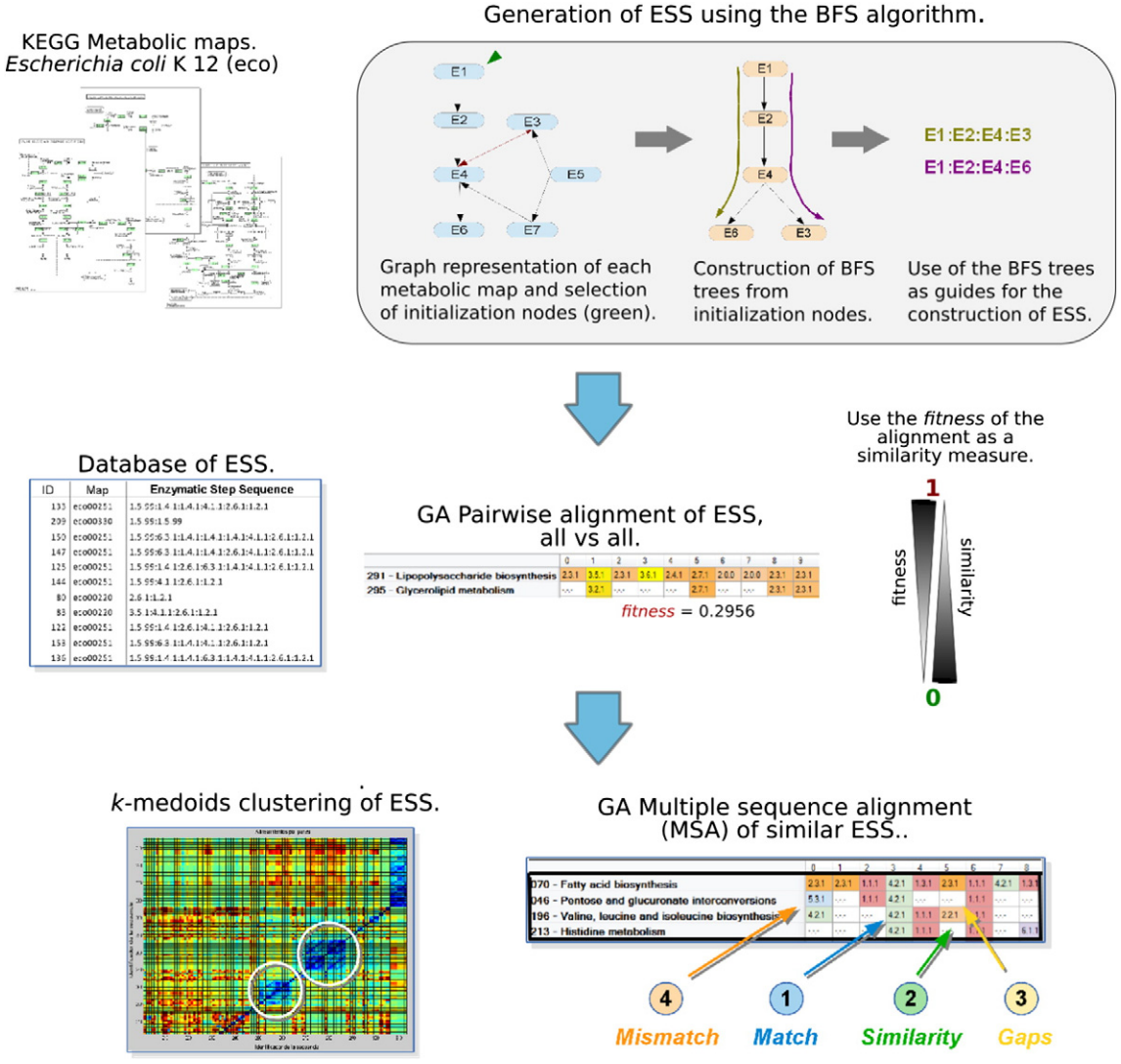

Fig. 1. General strategy for the comparative analysis of *E. coli* K-12 metabolism. The metabolic maps from KEGG were converted to ESS by using the breadth first search (BFS) algorithm. For each map a graphical representation was created, where nodes represent enzymes and edges are product-substrate relationships. Then, a set of initialization nodes was selected (green arrowhead) as roots for BFS trees. Those trees were used as guide for ESS construction. Afterwards all the ESS were compared against each other by GA pairwise alignments. The similarities among ESS were used to conduct a clustering analysis based on the *k*-medoids algorithm. Finally, clusters of similar sequences were aligned using an MSA approach.