Gene 534 (2014) 125–138

Contents lists available at ScienceDirect

Gene

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

Review

Analyzing methods for path mining with applications in metabolomics

Somnath Tagore [a], Nirmalya Chowdhury [b], Rajat K. De [c,*,1]

[a] Department of Biotechnology and Bioinformatics, Padmashree Dr. D. Y. Patil University, Navi Mumbai, India
[b] Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
[c] Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

ARTICLE INFO

Article history:
Accepted 25 October 2013
Available online 12 November 2013

Keywords:
Grammars
Metabolomics
Path mining
Patterns
Quantitative analysis

ABSTRACT

Metabolomics is one of the key approaches of systems biology that consists of studying biochemical networks having a set of metabolites, enzymes, reactions and their interactions. As biological networks are very complex in nature, proper techniques and models need to be chosen for their better understanding and interpretation. One of the useful strategies in this regard is using path mining strategies and graph-theoretical approaches that help in building hypothetical models and perform quantitative analysis. Furthermore, they also contribute to analyzing topological parameters in metabolome networks. Path mining techniques can be based on grammars, keys, patterns and indexing. Moreover, they can also be used for modeling metabolome networks, finding structural similarities between metabolites, in-silico metabolic engineering, shortest path estimation and for various graph-based analysis. In this manuscript, we have highlighted some core and applied areas of path-mining for modeling and analysis of metabolic networks.

© 2013 Elsevier B.V. All rights reserved.

Contents

1. Introduction

Systems biology deals with analyzing and modeling biological networks, visualizing complex pathways, identifying sub-steps of pathways, measuring gene expression levels, predicting outcome of various alterations made to the cells, and identifying intracellular targets for drugs and genetic engineering. One of the most applied areas of systems biology is *metabolomics*, dealing with analyzing metabolic reactions and studying the interactions among them

Abbreviations: SPM, sequential pattern mining; SPIRIT, sequential pattern mining with regular expression constraints; KEGG, Kyoto Encyclopedia of Genes and Genomes; KGML, KEGG Markup Language; ASN, Abstract Syntax Notation; SMILES, simplified molecular-input line-entry system; BRENDA, BRaunschweig ENzyme Database; T1D, Type 1 Diabetes mellitus; GC, giant component; DC, divide-and-conquer; HD, hamming distance; TC, Tanimoto coefficient; IM, incidence matrix.

[*] Corresponding author.
E-mail address: rajat@isical.ac.in (R.K. De).

[1] Tel.: +91 33 25753105; fax: +91 33 25783357.

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.10.056

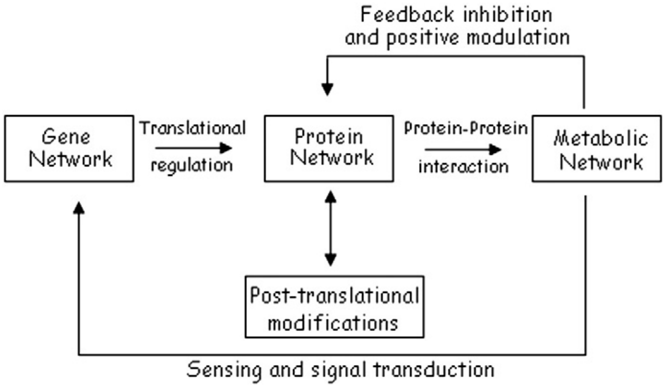
Feedback inhibition and positive modulation
Gene Network
Translational regulation
Protein Network
Protein-Protein interaction
Metabolic Network
Post-translational modifications
Sensing and signal transduction

Fig. 1. The metabolomics domain.

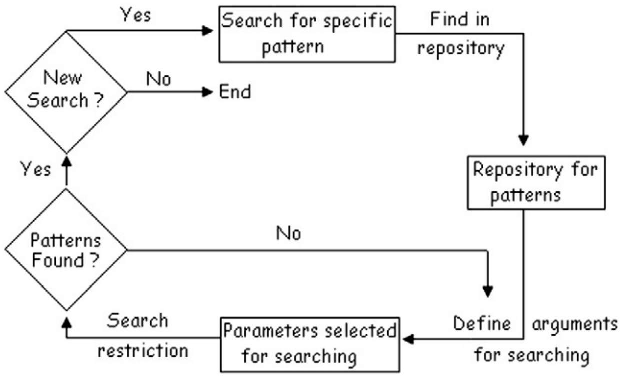

Fig. 3. Process mining: using concurrent processes to find sequential structures.

(Weckwerth, 2010). Due to the rapid growth in computational techniques, it is now possible to uncover the various vital characteristics and properties of biological systems, and to explore applications backed up by understanding their behaviors. Metabolomics also involves the systematic study of metabolites and reactions in a biological network by analyzing their response to genetic and physiological modifications (Wishart, 2010). Furthermore, analyzing the various components of biological networks can be used in pharmacology and drug designing. Together with the other omics techniques including transcriptomics, cytomics, fluxomics, metabolomics also contributes to understanding the function of metabolic, gene regulatory and signaling pathways as well as designing and simulating a whole cell using a system-based approach (Fig. 1). Another important application of metabolomics is detection of the differences between diseased and healthy metabolic pathways for precise diagnosis of diseases (Netzer et al., 2012).

Various strategies for metabolomic data acquisition have emerged for studying the nature, organization and control of metabolic networks. Also, several quantitative models, i.e., models based on graph-theoretical strategies, allow the true representation of complex biochemical systems. Moreover, strategies based upon graph-theoretical and path mining measures are regularly used in order to investigate the structure of biological systems, their dynamics, control and designing systems for understanding desired properties (Ferro et al., 2008). In this regard, a very useful technique in analyzing metabolomic data is by concentrating on the various reaction links and paths in biological pathways. This is particularly useful in case of finding structural as well as functional features in a metabolic pathway along with finding specific nodal points that can act as novel drug targets. Path mining is the application of graph theory and strategies based on quantitative models for analyzing datasets. One of the most important applications of path mining is in the area of systems biology, where many graph-based analyses are being done for studying complex biological networks (Van Helden et al., 2002).

One of the fundamental path mining strategies for analyzing complex metabolic networks is by segmenting these into simpler components. Moreover, each component can then be analyzed by studying them in the form of paths and their interconnections (Tan et al., 2002). For example, a path in a biochemical reaction can be the flow of information from one substrate to another (Fig. 2a), i.e., information flowing from reactants (*X*, *Y*) to product (*Z*) in the form of functional and structural details of the product formed. Similarly, a path in a biochemical moiety can be the interconnections among various elements and/or atoms present in them (Fig. 2b), i.e., the path in this case can be C – C – C – C – C – C. Extraction of frequently occurring patterns in public repositories, transaction and time-series repositories are the most popular areas in path mining. For instance, frequent-pattern-based path mining can be used in mining associations, correlations, sequential patterns, multi-dimensional patterns, partial periodicity and emerging patterns, to name a few (Inokuchi et al., 2003). In some instances, these path-based approaches can also use generic mining tools to extract implicit rules governing the path of tasks followed during the execution of a process. This realization of a process can be carried out by executing a subset of tasks. Furthermore, path mining fundamentally refers to identifying the subset of tasks that are triggered during the realization of a process (Rosemann and Zur Muehlen, 2000). An important extension of path mining is ‘process mining’, i.e., using concurrent processes to find sequential structures (Fig. 3). For instance, a process mining task may involve searching for a specific pattern (e.g. C – C – C) in biological repository. The repository may have a large set of molecules having a large number of patterns like, C – C, C═C, C═C – C – C, C≡C, to name a few. For detecting such patterns, one needs to use certain conditions or parameters as essential arguments. In the given example, these parameters or arguments can be the length of the pattern (i.e., at least 3), elements of the pattern (i.e., *C* and –) and occurrence of elements in the pattern (i.e., *C* followed by – followed by *C* followed by – followed by *C*). These arguments are essential for restricting the search to limited domains. Process mining techniques try to extract non-trivial and useful information from unformatted and experimental datasets. An important element of process mining is ‘control-flow discovery’, i.e., automatically constructing a process model that discusses the causal dependencies among various on-going processes (Weijters et al., 2003).

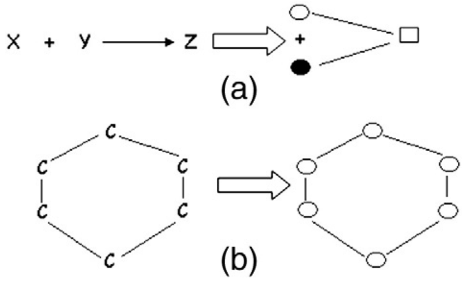

Fig. 2. Path mining: (a) in a reaction (b) in a molecule.

Another application of path mining is ‘sequential pattern mining (SPM)’, a method of determining the relationships between occurrences of sequential events, to find if there exists any specific order of the occurrences (Ayres et al., 2002). One of the earliest and possibly the simplest algorithms developed for SPM is AprioriAll that finds single frequently occurring items in the dataset and then attempts to find sequences of them. For instance, if a patient's reports with information related to their metabolome analysis (i.e. pathway based studies) is taken as sequence or input from a repository, patterns within that input are not identified, but the sequence is detected as a candidate. Furthermore, if this sequence is frequent across all patients, only then is it identified as a pattern (Fig. 4). For example, in Fig. 4 pathway-related patient information is fed as input. The aim is to identify a sequence that is frequent in all the 5 pathways. If this sequence is present, then the

Download English Version:

https://daneshyari.com/en/article/5905984

Download Persian Version:

https://daneshyari.com/article/5905984

Daneshyari.com