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.

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

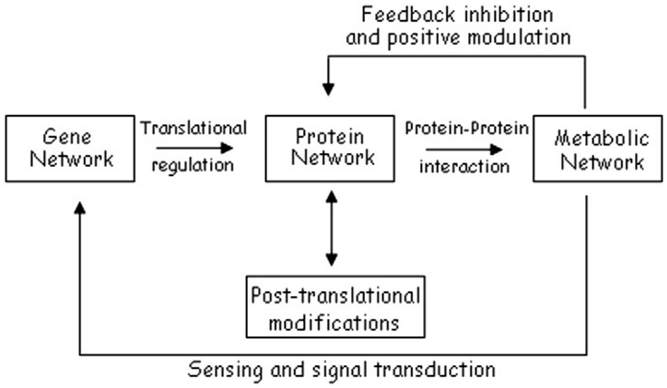

Feedback inhibition and positive modulation
Gene Network
Translational regulation
Protein Network
Protein-Protein interaction
Metabolic Network
Post-translational modifications
Sensing and signal transduction

Fig. 1. The metabolomics domain.

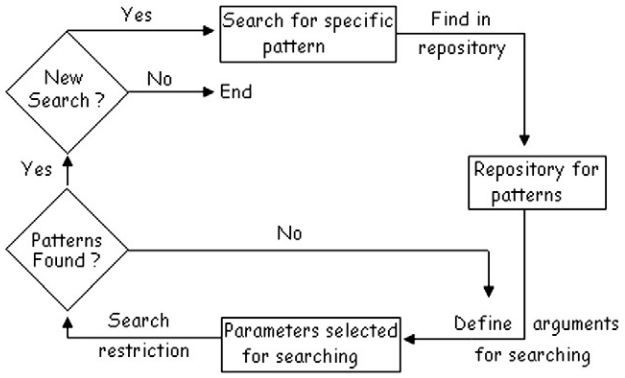


Fig. 3. Process mining: using concurrent processes to find sequential structures.

(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).

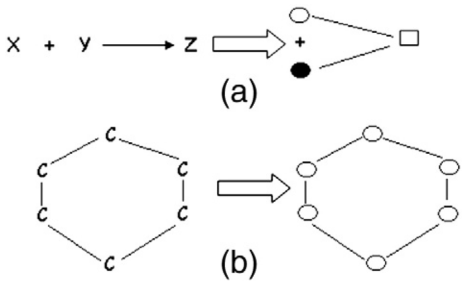
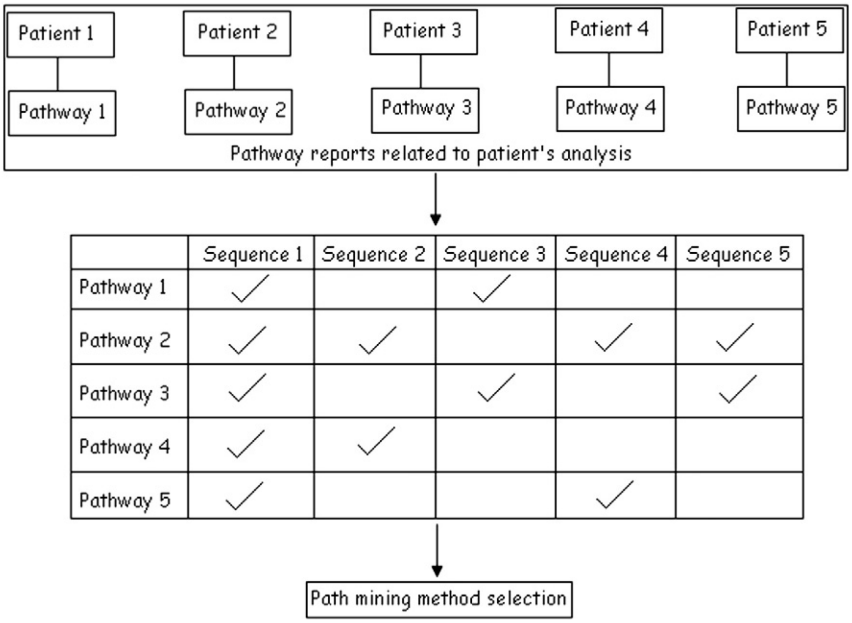


Fig. 2. Path mining: (a) in a reaction (b) in a molecule.

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



	Sequence 1	Sequence 2	Sequence 3	Sequence 4	Sequence 5
Pathway 1	✓		✓		
Pathway 2	✓	✓		✓	✓
Pathway 3	✓		✓		✓
Pathway 4	✓	✓			
Pathway 5	✓			✓	

Fig. 4. Sequential pattern mining: a way of determining the relationships between occurrences of sequential events, to find if there exists any specific order of the occurrences.

pathways are regarded as putative pathways as the sequence is prevalent in most of the patient dataset, otherwise they are discarded. Another useful algorithm in this regard is Sequential pattern mining with regular expression constraints (SPIRIT) designed to mine user-specified patterns (Garofalakis et al., 1999). It has been proved useful in case the searching is made for particular sequences of drug prescriptions, like known clinical pathways. Thus, by searching in clinical pathways, adherence of researchers to them can be evaluated in a much better manner.

Paths in biological data sets can be represented in various forms, such as graph-based grammars, structural keys, path patterns and distance-based indices. Graph-based grammars are one of the most classical concepts in path mining, which is known for its ability to convert complex features into simple patterns (Helms et al., 2009). In this case, certain rules for converting feature information into grammar need to be followed. For example, complex features can be used to generate simple or hyper-graph-based grammars using certain rules for generating simple or hyper-arcs joining pair-wise distinct vertices (Fig. 5a). In this case, the nodes represent metabolites whereas arcs represent reaction links. Arcs may also represent the tendency of one metabolite to convert to another in a metabolic reaction. Similarly, structural keys form a set of values that describe the structural and chemical composition in a molecule in the form of a boolean array. A boolean array is usually stored as an array of bits, where the bit is set to 1 if a particular structural feature is present and 0 if it is not (Xue et al., 2003). Furthermore, a boolean array can be studied in the form of a path consisting of a sequence of information where each step in the path consists of a bit. For example, for a biological moiety like $C_6H_{12}O_6$ having 3 elements C = 100001, H = 111000 and O = 100100 has a structural key 100001100001 100001100001 100001100001 111000111000 111000111000 111000111000 111000111000 111000111000 111000111000 100100100100 100100100100100100100100 (Fig. 5b). In case of path-patterns, biochemical reactions are studied in the form of node-to-node paths. Each metabolite in a biochemical reaction is represented as a node or a vertex whereas each reaction link stands for an edge (Flesca et al., 2006). Thus, the complete biochemical reaction can be studied in the form of a set of steps forming a path. For example, in reaction $A \rightarrow B$, A and B can be represented in the form of symbols ($A = \circ, B = \square$) whereas the reaction link can be represented in the form of conversion of one symbol to another (Fig. 5c). Finally, in distance-based indices

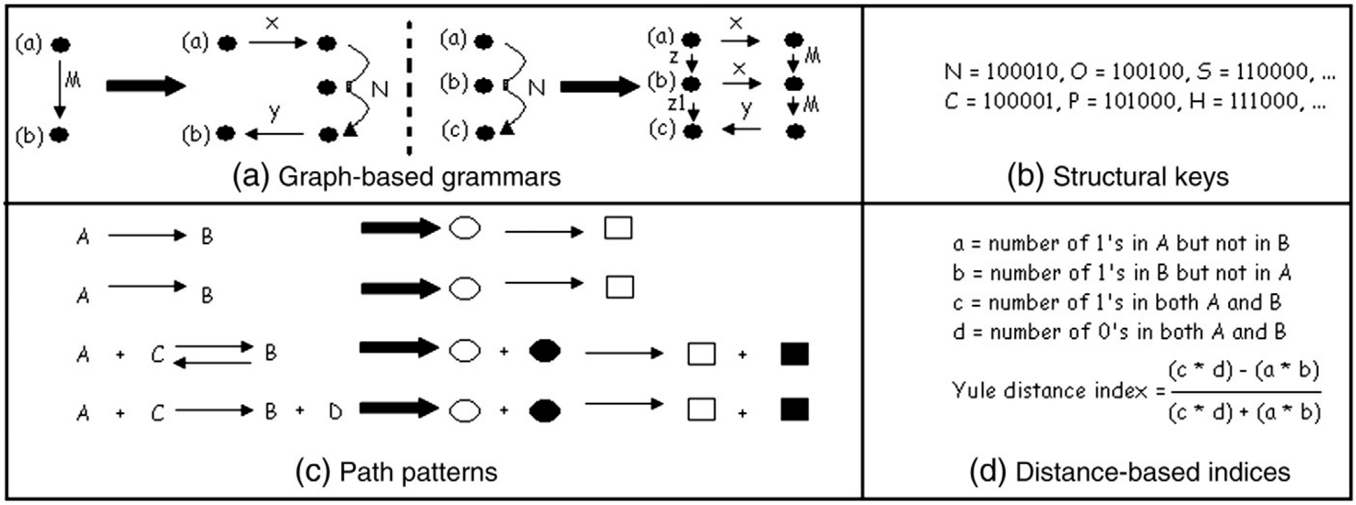

(a)
(b)
M
(a)
(b)
x
y
N
(a)
(b)
(c)
N
(a)
(b)
(c)
z
z1
x
x
y
M
M
(a) Graph-based grammars
N = 100010, O = 100100, S = 110000, ...
C = 100001, P = 101000, H = 111000, ...
(b) Structural keys
A → B
A → B
A + C ⇄ B
A + C → B + D
+
+
+
+
(c) Path patterns
a = number of 1's in A but not in B
b = number of 1's in B but not in A
c = number of 1's in both A and B
d = number of 0's in both A and B
Yule distance index = (c * d) - (a * b) / (c * d) + (a * b)
(d) Distance-based indices

Fig. 5. Paths in a metabolomics domain: (a) graph-based grammars (b) structural keys (c) path patterns (d) distance-based indices.

Table 1
List of some structure repositories.

Database name	Area	URL
BindingDB (Liu et al., 2006)	Experimental data on the non-covalent association of molecules in solution at University of Maryland Biotechnology Institute	http://www.bindingdb.org/
ChemIDplus (Tomasulo, 2002)	Allows searches by substance identification, toxicity, physical properties, molecular weight locator codes and chemical structure	chem.sis.nlm.nih.gov/chemidplus/
ChemFinder (Yoshii et al., 2002)	Provides chemical structures, physical properties, and hyperlinks	http://chemfinder.com
CSA (Thornton, 2004)	Catalytic sites and residues identified in enzymes using structural data	http://www.ebi.ac.uk/thornton-srv/databases/CSA/
CSLS (Cummings et al., 2007)	Compounds can be searched by InChI, SMILES, formula and other identifiers.	http://cactus.nci.nih.gov/cgi-bin/lookup/search
DSSP (Kabsch and Sander, 1983)	Database of secondary structure assignments for all of the entries in the Protein Data Bank (PDB)	http://swift.cmbi.ru.nl/gv/dssp/
EMDB (Tagari et al., 2002)	Public repository for electron microscopy density maps of macromolecular complexes and sub-cellular structures	http://www.ebi.ac.uk/pdbe/emdb/
FSSP (Holm et al., 1992)	Fold classification based on structure–structure assignments	srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+LibInfo+-lib+FSSP
HSSP (Sander and Schneider, 1994)	HSSP is a derived database merging structural (2-D and 3-D) and sequence information (1-D)	swift.cmbi.ru.nl/gv/hssp/
Molinspiration (Li et al., 2009)	Calculate the properties or predict bioactivity after drawing chemical structures	http://www.molinspiration.com/cgi-bin/properties
NDB (Berman et al., 2003)	A repository of three-dimensional structural information about nucleic acids	http://ndbserver.rutgers.edu/
PDB (Berman et al., 2000)	The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies	www.pdb.org
PDBe (Velankar et al., 2011)	The Protein Databank in Europe (PDBe)	http://www.ebi.ac.uk/pdbe/
PDBeChem (Dimitropoulos et al., 2006)	The ligand library, a complete chemical description of all the distinct chemical components found within the PDB	http://www.ebi.ac.uk/msd-srv/msdchem/cgi-bin/cgi.pl
PDBeFold (Berman et al., 2000)	Secondary Structure Matching (SSM) is an interactive service for comparing protein structures in 3D	http://www.ebi.ac.uk/msd-srv/ssm/
PDBeMotif (Berman et al., 2000)	PDBeMotif is an extremely fast and powerful search tool that facilitates exploration of the Protein Data Bank (PDB)	http://www.ebi.ac.uk/pdbe-site/pdbemotif/
PDBe NMR (Berman et al., 2000)	A respository of NMR structures in PDBe	http://www.ebi.ac.uk/pdbe-apps/nmr/main.html
PDBePisa (Berman et al., 2000)	A tool for the exploration of macromolecular interfaces, prediction of probable quaternary structures, database searches of structurally similar interfaces and assemblies, and searches on various assembly and PDB entry parameters	http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html
PDBsum (Laskowski et al., 1997)	PDBsum is a pictorial database providing an at-a-glance overview of every macromolecular structure deposited in PDB	http://www.ebi.ac.uk/pdbsum/
ProFunc (Laskowski et al., 2005)	ProFunc server helps to identify the likely biochemical function of protein from its 3D structure	http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/

similarities and dissimilarities between two or more metabolites can be found based upon the presence or absence of certain features in them (Ciaccia, 1998). This set of features can be found on the basis of presence or absence of a bit in a metabolite, in the form of a boolean array. For example, in a reaction *glucose*→*gluc*-6-*phosphate*, if glucose is represented by 100001100001 100001100001 100001100001 111000111000 111000111000 111000111000 111000111000 111000111000 111000111000 100100100100 100100100100 100100100100 and *gluc*-6-*phosphate* by 100001100001 1000011000011 000011000011 110001110001 110001110001 1100011100011 100011100011 100011100011 100011100011 100010010010 01001001001 001001001001 001001001001 00100100100101000 (using Fig. 5b), then Yule index = 0.995 (Nam, 2007) (Fig. 5d, Table 10 in Supplementary Information).

The present article reviews various methods of mining paths in metabolic networks from public repositories. This is followed by some applications of path mining methods in metabolomics. We also describe some strategies that have been regularly used by system biologists for interpreting the fundamental features of metabolome datasets followed by a case study discussion.

Table 2
List of some repositories used for graph-based path mining.

Database name	Area	URL
AllegroGraph (Alexander et al., 2009)	A scalable, high-performance graph database	http://www.franz.com/agraph/allegrograph/
DEX (Martinez-Bazan et al., 2007)	A high-performance graph database	http://www.sparsity-technologies.com/dex
Filament (Angles and Gutierrez, 2005)	Graph persistence framework and associated toolkits based on a navigational query style	http://filament.sourceforge.net/
FlockDB (Angles and Gutierrez, 2008)	An open source distributed, fault-tolerant graph database for managing data at webscale	https://github.com/twitter/flockdb/
Gremlin (Dries and Nijssen, 2007)	An open-source graph programming language that works over various graph database systems	http://gremlin.tinkerpop.com/
HyperGraphDB (Goertzel, 2006)	An open-source (LGPL) graph database supporting generalized hypergraphs where edges can point to other edges	http://www.hypergraphdb.org/index
InfiniteGraph (Martinez-Bazan et al., 2007)	A highly scalable, distributed and cloud-enabled commercial product with flexible licensing for startups	http://www.infinitegraph.com/
InfoGrid (Dries and Nijssen, 2007)	An open-source/commercial graph database with web front end and configurable storage engines	http://infogrid.org/
Neo4j (Alexander et al., 2009)	An open-source/commercial (AGPLv3) graph database	http://neo4j.org/
Sones (Martinez-Bazan et al., 2007)	An open-source/commercial graph database and universal access layer	http://www.sones.com/
OrientDB (Dries and Nijssen, 2007)	A high-performance open source document-graph database	http://www.orientechnologies.com/
VertexDB (Alexander et al., 2009)	High performance graph database server that	http://www.dekorte.com/projects/opensource/vertexdb/

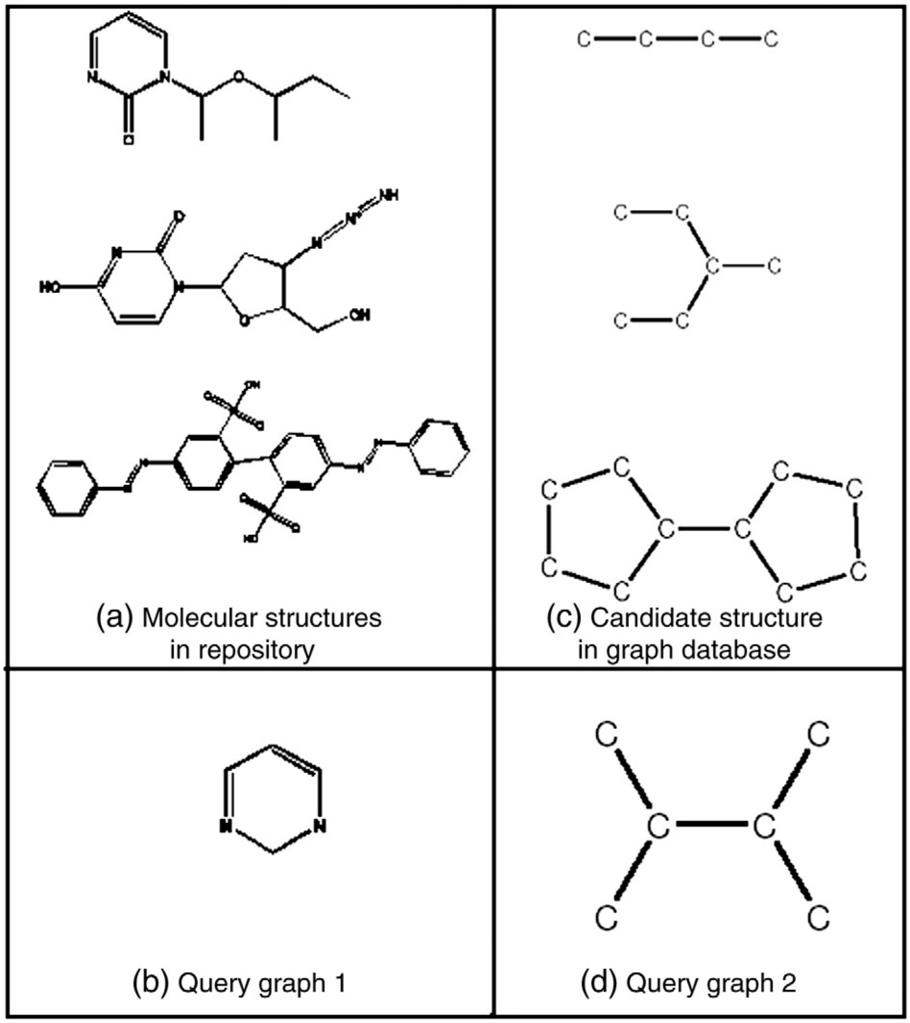

(a) Molecular structures in repository
(b) Query graph 1
(c) Candidate structure in graph database
(d) Query graph 2

Fig. 6. Structures present in repositories along with query structure and candidate structure.

2. Path mining methods for searching public repositories

Before we discuss the strategies of mining paths through graphs, we need to talk about the various public repositories which can act as useful resources in this regard. Considering the various path mining strategies, two categories of repositories are important, viz., structure and graph (Smalter et al., 2008). Due to the large scale sequencing projects, huge datasets are getting generated. In order to accommodate this data in a structured manner, the sizes of repositories have started to grow tremendously. Out of this large collection, structure repositories consist of information pertaining to 2D and 3D structure of bio-molecules (Wheeler et al., 2003). A list of some important structure repositories are given in Table 1. We have specifically listed only those which are essential with respect to path mining research. Similarly, graph repositories consist of information related to graph-based techniques, theories, applications and problems (Cook and Holder, 2000). Most of these repositories are open-source. Table 2 lists some notable repositories used for graph-based path mining techniques. Based upon the selection of any of the given repositories, the format of input data differs. Furthermore, the strategy for path mining also differs drastically. This is because mining a 2D structure repository is different from mining a 3D structure repository. For instance, in any path-based approach, there should be some structures for which some patterns or paths are identified (Hu and Wu, 2007). Input data can be selected by the users themselves or by taking the help of experimentalists who provide the raw data. If the candidate structure is provided by the user, some specific sub-graphs or patterns may be searched against these repositories.

The following list gives information about the various path mining approaches used for searching information in public repositories:

Pattern-based mining — Segmenting given inputs into patterns for analysis.
Process-based mining — Performing a series of sequential mining steps to achieve a solution.
Multi-method mining — Combining multiple mining strategies for getting a solution.

Here we discuss the three path mining approaches, namely, *pattern-based mining*, *process-based mining* and *multi-method mining* for handling datasets present in public repositories. A standard *pattern-based mining* could be to defragment the existing structures present in

Table 3
Grammars for $HH_2C{=}CH_2CL$.

Paths	Patterns		
0-Bond paths	C	H	Cl
1-Bond paths	HC	CC	CCl
2-Bond paths	HC=C	C = Cl	CC – Cl
3-Bond paths	HC – C – Cl	HC – C – H	

Table 4
Reaction grammars.

Paths	Patterns							
	Reactant				Product			
0-Bond paths	1 I	1 Na	3 C	1 Br	1 I	1 Na	3 C	1 Br
1-Bond paths	1 C=C	1 C – C	1 C – Br		1 C=C	1 C – C	1 C – I	
2-Bond paths	1 C=C – C	1 C – C – Br			1 C=C – C	1 C – C – I		
3-Bond paths	1 C=C – C – Br				1 C=C – C – I			

repositories into ‘patterns’ or ‘paths’ (Kuramochi and Karypis, 2001). Paths or patterns are defined as the bonds present between the atoms, i.e. path of 0-length (single atoms), path of 1-length (2 atoms connected by a bond), path of 2-length (3 atoms connected by 2 bonds) and so on. From Fig. 6a, the paths that can be found out from the given structures are, path of 0-length: C, O, N, …; path of 1-length: C – C, C – O, C – N, ..., N – N, ...; path of 2-length: C – C – C – C – O – C – C – N – C, …, and so on. Similarly, for the query graph given in Fig. 6b, the edges can be found out as, 0-edge: *C*, *N*; 1-edge: C – C, C – N; 2-edge: C – N – C, …, and so on. Furthermore, in a typical graph database, we have various candidate structures, as given in Fig. 6c. These structures may be a sub-structure of some other parent bio-molecule. For a query graph, such as in Fig. 6d, one can perform a path-based mining with the structures present in the repository (Xue et al., 2001). If one considers this query structure as a graph, it can be used to search in the repository directly. An alternate way is to break this graph into patterns and search them individually. After these steps, one can conclude that this query graph may be a sub-structure of a different molecule. This information may be further used for modeling and simulation. Moreover, we also see that the query graph in Fig. 6d is present only in the last structure of Fig. 6c.

Another category of path mining is *process-based mining* that defines the activities that need to be executed and the order in which this is to be done (Xue et al., 2001). In this case, specific conditions are usually given corresponding to causal dependencies between the activities. There are four basic types of relationships between the various activities, namely, sequential (i.e. several activities being executed in sequence), parallel (i.e. two or more activity instances being executed in parallel), conditional (i.e. one branch is selected and taken from multiple alternative work-flow branches) and iteration (a work-flow activity cycle involving the repetitive execution of one or more work-flow activities) (Rubin et al., 2007). During the execution of each work-flow process, certain logical expressions are evaluated by the work-flow manager to decide the sequence of activity execution within a process. These expressions are evaluated against a set of work-flow variables, specified in the work-flow definition and whose values can be altered during execution of the work-flow process.

Process-based mining predicts the sequence of activity, within a process, that is executed at runtime by the work-flow manager (Fig. 7 in Supplementary Information) (Van der Aalst et al., 2009). In this case, during the execution of a specific path mining task, three different paths can be followed to complete the process: process path 1, process path 2 and process path 3. All the three possible process paths consist of several steps and their dependencies. Dependencies in this example are the sequence of steps required to complete the process path completely. For instance, in process path 1, step 1 is followed by step 5 and step 9. It is the duty of the work-flow or process manager to select the optimum process path that best suits for this task (Lu and Sadiq, 2006). For example, in process path 1, an initial query can be a metabolic pathway dataset from KEGG such as *glycolysis*. Step 1 can be conversion of this dataset to KGML format for graph-based representation. This can be followed by application of any algorithm for identifying interactions among the metabolites and finding structural similarities among metabolites (i.e. step 2). On the basis of similarity, the metabolites are further grouped into classes. For instance, *glucose* and *glucose*-6-*phosphate* are very similar in their structural features as their Yule index is 0.995 (using Fig. 5d). Thus, the tendency for conversion of *glucose* to *glucose*-6-*phosphate* is very high. This result is validated using literature references (i.e. step 3) giving rise to the final conclusion.

Table 5
Difference grammars.

Paths	Grammars			
0-Bond paths	0 I	0 Na	0 C	0 Br
1-Bond paths	0 C=C	0 C – C	1 C – Br	1 C – I
2-Bond paths	0 C=C – C	1 C – C – Br	1 C – C – I	
3-Bond paths	1 C=C – C – Br	1 C=C – C – I		

The next useful strategy of path mining that is used for searching public repositories is *multi-method mining*. This strategy dynamically combines different methods with no predefined order to the same problem (Kriegel et al., 2007). The methods are applied in sequence as well as at various sub-steps of the entire process of the internal knowledge representation. The logic behind multi-method path mining is to incrementally evolve and improve knowledge by using various methods. Thus, one gets multiple solutions for the same problem, where each solution is obtained using the application of different methods with different sets of parameters. Moreover, with this approach the solution space and expressive power is dramatically increased. The only drawback of the approach is that the increased search space requires more computing power to find an optimal solution. In Fig. 8 of Supplementary Information, we start with a problem, for which we have multiple frameworks and multiple methods to solve it. Each operator signifies the sub-steps of each method (Kriegel et al., 2007). Thus, in this problem we have two frameworks (1 and 2) for solving the problem using two methods (1) and (2). Framework 3 has a list of search spaces (3(1) to 3(n)) from which searching can be done with the given problem query. For each search space, we have a specific strategy (4(1) to 4(2)) for solving the problem. Thus, the multithread approach uses multiple strategies, search spaces and frameworks for obtaining a solution. In the given problem, a single unique solution is found, which is very rare (Hirsh, 2008). In general, multiple solutions for the same problem are found. For example, a simple problem in metabolism can be analyzing a metabolic pathway structurally by studying its components as well as analyzing the interaction among the various metabolites present. For this purpose, a typical analysis system consists of 2 frameworks where framework 1 has various operators like kegg2sbml (operator 1) that is used to convert the given metabolic pathway dataset to sbml representation (method 1), cytoscape (operator 2) that can be used to upload the sbml file and to view pathway (method 2), subdue (operator 3) for studying interaction among metabolites. Similarly, in framework 2 ASN.1 format (operator 1) of the metabolic pathway that can be collected from KEGG using sampling (method 1), cell designer (operator 2), can be used for uploading and viewing the file (method 2), and jarnac (operator 3) to study interaction among metabolites (method 3). The combined frameworks 1 and 2 can be used for studying this metabolic pathway structurally as well as analyzing the interactions among its various metabolites (Hirsh, 2008).

3. Current methods in path mining

Here we describe various standard path mining methods for analyzing metabolic pathways. These methods have been categorized on the basis of the input data format they require and the manner in which they work. We discuss the following path mining methods in the categories:

Grammar-based mining — Using inputs in the form of grammars/ patterns for mining.

Regular expression-based mining — Using regular expressions for searching of inputs.

Key-based mining — Converting inputs to keys for mining.

Pattern-based mining — Detecting frequently occurring patterns in datasets.

Neighborhood-based mining — Analyzing interactions among metabolites based on neighborhood connectivities.

Index mining — Using comparison indices for analyzing similarity among metabolites.

3.1. Grammar-based mining

Grammar-based mining follows the strategy of converting the input datasets into certain patterns or grammars. Grammars are bitmaps or boolean arrays generated from a metabolite having no pre-assigned meaning to each bit representing the characteristics of a feature (Witten et al., 2011). A general grammar-based mining approach examines a metabolite and generates a pattern for each atom; a pattern representing bonds between an atom and its neighbors; a pattern representing atoms and bonds connected by paths up to 2, 3, … $(n-1)$ bonds long, where 'n' stands for the path length of the metabolite (Hirsh, 2005). Thus, an exhaustive list of patterns is generated. For instance, the metabolite chloroethane (HH_2C-CH_2Cl) generates 11 patterns as given in Table 3.

Furthermore, we can extend this procedure for calculating the overall change in bonds in a chemical or bio-chemical reaction. For example, we consider a reaction $[I-]$: $[Na+]{:}C{=}CCBr \gg [Na+].[Br-]{:}C{=}CCI$ and identify the overall change in bonds (Table 4). After we generate the grammars for both reactant and product, we also find the difference between the grammars of reactant and product for all the paths. We only consider the counterparts that have non-zero differences (Table 5) (Bille, 2006). We consider only these paths as we get the idea of overall change in bonds in the given reaction.

A major problem with this method is that the number of patterns generated becomes large with increase in the number of atoms present in the input metabolite. Due to this reason, assignment of a bit to its pattern is not possible as there is no pre-defined set of patterns (Hirsh, 2005). For example, even a bio-molecule like *UDP-glucose* ($C1_5H2_4N_2O_{17}P_2$) has more than 200 patterns, some of which are 0-bond paths (C, C, C, C, …, O, O, O, O, …, N, …, H, H, …, P, …), 1-bond paths (C−N, …, C═O, …, C−NH, …, C−C, C−C, …, C═C, C−OH, C-OH, …, C−H, C−H, …, CH_2-C, $O-CH_2$, P−O, P−O, …, O−O, P═O, CH_2OH-C), and so on. Thus, each pattern is hashed and combined to produce the corresponding grammar of the metabolite. For instance, the final grammar of *UDP-glucose* becomes sixty 0-bond paths (15 *C*, 17 *O*, 2 *N*, 24 *H*, 2 *P*), forty-four 1-bond paths (3 C−N, 7 C═O, 2 C−NH, 8 C−C, 1 C═C, 5 C−OH, 8 C−H, 1 CH_2-C, 1 $O-CH_2$, 5 P−O, 1 O−O, 1 P═O, 1 CH_2OH-C), and so on. Also, if a pattern is a substructure of a molecule, each bit that is set to the pattern's grammar, is also there in the metabolite's grammar. Thus, *UDP-glucose* has many repeating groups whose patterns are combined together to form the final grammar of the molecule. Moreover, one can even use boolean operations on grammars for screening metabolites in a chemical repository. Thus, each pattern generates its set of bits, out of which at least one is unique and not shared with any other pattern present in the metabolite. For instance, *glucose* ($C_6H_{12}O_6$) and *glucose*-6-*phosphate* ($C_6H_{13}O_9P$) have ninety-seven patterns each only differing in fifteen patterns, namely, $C-CH_2OH$, $C-C-CH_2OH$, $H-C-CH_2OH$, $OH-C-CH_2OH$, $C-C-C-CH_2OH$, $H-C-C-CH_2OH$, $OH-C-C-CH_2OH$, $C-C-C-C-CH_2OH$, $OH-C-C-C-CH_2OH$, $H-C-C-C-CH_2OH$, $C-C-C-C-C-CH_2OH$, $H-C-C-C-C-CH_2OH$, $OH-C-C-C-C-CH_2OH$, $H-C-C-C-C-C-CH_2OH$, $O=C-C-C-C-C-H_2OH$ in *glucose* whereas $C-CH_2OPO_3^{-2}$, $C-C-CH_2OPO_3^{-2}$, $H-C-CH_2OPO_3^{-2}$, $OH-C-CH_2OPO_3^{-2}$, $C-C-C-CH_2OPO_3^{-2}$, $H-C-C-CH_2OPO_3^{-2}$, $OH-C-C-CH_2OPO_3^{-2}$, $C-C-C-C-CH_2OPO_3^{-2}$, $OH-C-C-C-CH_2OPO_3^{-2}$, $H-C-C-C-CH_2OPO_3^{-2}$, $C-C-C-C-C-CH_2OPO_3^{-2}$, $H-C-C-C-C-CH_2OPO_3^{-2}$, $OH-C-C-C-C-CH_2OPO_3^{-2}$, $H-C-C-C-C-C-CH_2OPO_3^{-2}$, $O=C-C-C-C-C-CH_2OPO_3{}^{-2}$ in *glucose*-6-*phosphate*. The difference between the patterns among these metabolites is on the basis of their grammar-based linkages.

Similarly, searching *glucose*-6-*phosphate* in a repository tends to be easier as the searching is done on the basis of number of patterns and the type of patterns, such as 0-bond paths, 1-bond paths, to name a few. This searching criterion is unique for each molecule present in the repository and the presence of this unique pattern can differentiate the given metabolite from the others. Furthermore, the patterns that go into a grammar can also be overlapped, except for lone atoms that may lead to problems in correct identification of the metabolites in the repository. For example, in *UDP-glucose*, some of the patterns that are overlapped are $CH_2OH-C-O-C$, *C−O−C−C*, $CH_2OH-C-C-C$ and C−C−C−C, whereas some of them are repeated such as C−OH (8 times), C−N (3 times), O−P−O (2 times), to name a few, which may lead to redundancy and improper matching in the repository (Bille and Farach-Colton, 2005). This leads to the application of a different concept in path mining that is better than grammar-based method, i.e., regular expression-based mining. It is easier to represent and simpler to implement. Also, there is minimal amount of redundancy associated with this approach. The next subsection gives detailed information about the regular expression-based mining.

3.2. Regular expression-based mining

In regular expression-based mining, expressions are matched on the basis of the presence or absence of a specific feature in the query data. This specific feature can be matched using certain special characters or sequence of characters (Table 6 in Supplementary Information). Along with these features, certain parameter values, called 'dependency' can also be specified. These dependency values may be based on a matching criterion between the query and the data sets already present in the repository (Myers and Miller, 1989). Thus, given a pattern 'p' and a target 't' present in the repository, one can define a set of production rules along with their associated dependency/cost. Its main objective is to find an optimal ordered list of transformations 'p' into 't'. For example, searching the occurrence of two amino acids serine ('S') and methionine ('M') can be in the form of a pattern, 'S–M' ('p'), whereas the target ('t') can be the list of all patterns present in the repository. A careful selection of production rules will allow the search engine to evaluate regular expressions by simply changing the scoring parameters. This kind of path mining is more often used for metabolic pathway alignment containing three key dependency rules, viz., insert, delete and match/mismatch (Bairoch and Apweiler, 2000). In this case, each rule has an associated score which is evaluated using a scoring matrix. Moreover, it may be desirable to score match/mismatch, insert and delete differently based on pattern specifications. A simple score calculation strategy can be assigning a value for match (+1), mismatch (−0.5), insert (+0.5) and delete (−0.25), and calculating summation of all the values based upon positional match/mismatch among both the queries. For example, aligning two reactions *phosphoenol pyruvate* → *pyruvate* ('seq1') and *phosphoenol pyruvate* → *oxaloacetate* ('seq2'), based upon their enzyme contributions ('pyruvate kinase' and 'phosphoenolpyruvate carboxykinase'), patterns for enzymes in 'seq1' and 'seq2' are $[LIVAC]-x-[LIVM](2)-[SAPCV]-K-[LIV]-E-[NKRST]-x-[DEQHS]-[GSTA]-[LIVM]$ and $L-I-G-D-D-EH-x-W-x-[DEPKVNA]-x-[GVS]-[IV]-x-N$, respectively. For aligning these two patterns using insertion, the modified patterns are $[LIVAC]-x-[LIVM](2)-[SAPCV]-K-[LIV]-E-[NKRST]-x-[DEQHS]-[GSTA]-[LIVM]andL-I-G-D-D-E-H-x-W-x-[DEPKVNA]-x--E-[GVS]-[IV]-x-N$ (insertion in 20th position with 'E'). On the other hand, whereas using deletion $[LIVAC]-x-[LIVM](2)-[SAPCV]-K-[xxx]-E-[NKRST]-x-[DEQHS]-[GSTA]-[LIVM]$ (deleting *L*, *I*, *V* at 17th, 18th, 19th positions, and inserting 'x', 'x', 'x'). Thus, using both 1^{st} and 2^{nd} alignment strategies give a score of −15.5, as there are only three matches in both ‘seq1’ and ‘seq2’.

This category of path mining is extremely useful for searching repeated patterns, rarely repeated patterns in the dataset. One can even take into account some repeated units in the metabolic networks that can be detected using this strategy. For example, a specific biochemical reaction may be repeatedly occurring in different metabolic networks (Frenz, 2007). In order to identify the frequency of occurrences of this reaction in all the metabolic networks, one can select the regular expression-based mining strategy. Furthermore, the same strategy can be employed while searching for any metabolite in a given metabolic pathway (Kramer et al., 2001). For example, a text search using coenzyme A in a repository as an input can be searched as ‘co-A’, ‘coenzyme-A’ or ‘coA’, or using ‘*’ (wildcard character) for which the regular expression is ‘co.*a’ corresponding to any metabolite name that starts with ‘co’, followed by any number of additional characters and ends with ‘a’. Moreover, a problem with this approach can be using an incorrect regular expression-based pattern mining. Some examples of regular expression-based patterns of bio-molecules taken from Prosite repository are presented in Table 7 (in Supplementary Information) (Bairoch and Apweiler, 2000). This kind of a regular expression determines the sequence of components in the biomolecule. For instance, signal peptidases I serine active site consist of a series of amino acids like *glycine* (*G*) or a *serine* (*S*) followed by any arbitrary amino-acid (‘x’), a *serine* (*S*), a *methionine* (*M*), another arbitrary amino acid (y), a *proline* (*P*), an *alanine* (*A*) or *threonine* (*T*), a *leucine* (*L*) or *phenylalanine* (*F*). Thus, the pattern ‘S–M’ is found only in signal peptidases I serine active site. Furthermore, in case a pattern is absent from a specific query, the corresponding regular expression-based strategy cannot be applied. Thus, the success of this approach is strictly based on the pattern present in a repository for any given query.

3.3. Key-based mining

Key-based mining is one of the most primitive path mining strategies followed for searching a pattern in any structural or graph repositories. They are also very important from the point-of-view of screening large scale public repositories. A key can be represented in the form of a boolean array, where each element is either ‘1’ or ‘0’ (Calders and Goethals, 2002). Thus, any boolean array is a set of bits or bitmaps where each bit represents one position. For constructing a key, we need to decide based on any distinctive pattern(s). This is followed by assigning each pattern to its corresponding bit thereby generating a set of bits or the key for the complete molecule (Agarwal et al., 2001). This process is repeated for all the molecules present in the repository. Some examples of features for which keys can be generated are given in Table 8 (in Supplementary Information). For example, for a molecule like benzene (C_6H_6), an important key can be ‘at least 6 C’. If the key is taken as an input query and is matched with all the molecules present in the repository, only those molecules would be matched having at least six carbon atoms in their structures, i.e., having this key as a feature. Moreover, when a repository is searched for a specific pattern, its key is generated. As the mining task proceeds, the pattern's key is compared to that of each molecule in the repository. The matching criterion is such that in each position if a true or false bit is present in the query, the same should be present in the repository. If this case happens, the pattern is present in the repository whereas, if the matching is negative, pattern is not found (Zaki and Hsiao, 2002). This method of searching is especially fast in case of sub-structure matching in any repository. For instance, if we consider 3 query keys ‘1 C’, ‘2 C’ and ‘3 C’ corresponding to one carbon, two carbon and three carbon for matching structures present in repositories, we see that ‘1 C’ key is matched with all the structures they have at least one carbon atom, while ‘2 C’ is not matched for CH_3 as it does not have two carbon atoms, and ‘3 C’ is not matched for CH_3 and C_2H_6 respectively as both do not have three carbon atoms (Table 9 in Supplementary Information).

3.4. Pattern-based mining

Pattern-based mining involves discovering patterns or association rules present in a dataset. These patterns may also include anomalous data that might be associated with noise (Han et al., 2007). Pattern-based mining also help in analyzing and comprehending existing metabolome systems, customizing systems to fit user needs and construction of synthetic systems. Patterns provide a greater framework for the solution process, and are based on equally broad characteristics of the problem. The pattern-based mining approach is loosely divided into two steps, *first*, choosing a pattern from a list based on some problem characteristics, and *second*, using various path mining algorithms for the sub-tasks (Adam et al., 2004). In pattern-based approaches, patterns are applied to discover specific relationships between terms from the general first-order co-occurrences. One of the important subcategories of patterns are called ‘frequent item-sets’, used to find interesting patterns from repositories. The item-sets can be in terms of association rules, correlations, sequences, episodes, classes and clusters (Abulaish and Dey, 2005). Fig. 9 in Supplementary Information discusses the steps and components of an itemset-based pattern mining approach. In this case, for a set of items ‘S’, a set ‘W’, called an itemset, contains ‘k’ items present in transaction repository ‘R’, where a transaction over ‘W’ is called couple ‘C’. The cover (*Cor*) of an itemset ‘W’ in ‘R’ consists of the set of transaction identifiers of transactions in ‘R’ supporting ‘W’. The support (*Sup*) of an itemset ‘W’ in ‘R’ is the number of transactions in the cover of ‘W’ in ‘R’. Finally, the frequency (*Frq*) of an itemset ‘W’ in ‘R’ is theprobability of ‘W’ occurring in a transaction ‘C’ in ‘R’ (Krallinger and Valencia, 2005). For example, let us consider 3 metabolic pathways in *H. sapiens*, viz., *glycolysis*, *TCA cycle* and *pentose-phosphate pathway*. We have a set of metabolites or items present in these 3 pathways, i.e., $W =$ *glucose* (‘g’), *glucose*-6-*phosphate* (‘g_1’), *pyruvate* (‘p’), *phospho-enol pyruvate* (‘p_1’), *oxaloacetate* (‘o’), *acetyl coA* (‘a’), *α-D-gluc*-1-*P* (‘a_1’), *D-ribose*-5-*P* (‘r’), *D-ribulose*-5-*P* (‘r_1’), *D-erythrose*-5-*P* (‘e’). In this case, the total number of items present in ‘I’ is ‘k’ = 10. Given this input, our aim is to identify those metabolites that are present in at least two metabolic pathways, i.e., $Sup = 2$. Now, we create 3 pathway sets, ‘MP_1’ = {‘g’, ‘g_1’, ‘p’, ‘p_1’, ‘o’, ‘a’, ‘a_1’}, ‘MP_2’ = {‘a_1’, ‘g’, ‘p_1’, ‘o’, ‘g_1’, ‘a’} and ‘MP_3’ = {‘r’, ‘r_1’, ‘e’} corresponding to the three metabolic pathways and their metabolites. Based upon the support, we find the itemset frequency, i.e., $Frq =$ {‘g’}, {‘g_1’}, {‘p’}, {‘p_1’} and {‘o’} respectively. This kind of path mining is capable of detecting frequent patterns, but is capable of detecting patterns occurring in neighborhood of the selected metabolites as well as metabolic pathway. This lacuna can be removed by shifting to another path mining strategy called ‘neighborhood-based mining’.

3.5. Neighborhood-based mining

Neighborhood-based mining are mainly used for analyzing interactions among biological networks. These strategies find out dependencies among metabolites as well as the processes in which they participate. Thus, by analyzing the interaction patterns of these networks, implicit information embedded in their topologies can be discovered (Bichindaritz and Akkineni, 2005). An important feature of metabolic networks, which can be found using neighborhood-based mining, is topological proximity, representing the closeness among metabolites in terms of their interaction strength. Two categories of topological distances can be found out using neighborhood mining, namely, *shortest-path-based* and *direct neighbor-based distances* (Seth et al., 2009). Shortest-path-based mining is the detection of the least number of sub-steps that are present between the initial and terminating metabolites in a metabolic network. Direct-neighbor-based distance is the number of direct links that are present between two metabolites in a network (Zhang et al., 2006). In the hypothetical metabolic network given in Fig. 10 in Supplementary Information, if we assume that ‘a’ is the initial starting metabolite (*) and ‘h’ is the terminating metabolite

(#), then the path 'a-g-h' is the shortest path while the path 'g–h' is one of the direct neighbors present in it. Furthermore, one can use the concept of neighborhood-based mining for various tasks, such as comparison of various metabolites and finding interaction sub-pattern in interaction networks. Neighborhood-based mining approaches also use the notion of overlap measure to define the neighborhood of a metabolite determining similarity between a pair of metabolites (Saha et al., 2010). This overlap can be on the basis of the neighborhood density around a metabolite, i.e., the number of links around a metabolite (Fig. 11 in Supplementary Information).

Similarly, it can be further calculated based upon this neighborhood density, that which metabolite is the most frequently occurring in the entire metabolic network. Thus, it is interacting with many more metabolites in the metabolic network. For example, metabolite 'a' is interacting with metabolites 'b', 'c', 'e' and 'g'. This can further be extended by saying that these maximally occurring metabolites have the highest tendency to get annotation and converted to another metabolite because of its participation in a large number of reactions. Furthermore, a distance measure, called 'Czekanowski–Dice distance', can be calculated based upon the interaction between two different metabolites within a particular topological proximity. It is defined as the distance between metabolites M_1 and M_2 that considers shared neighborhoods. Thus, Czekanowski–Dice $(M_1, M_2) = |N(M_1)\Delta N(M_2)|/[|N(M_1) \cup N(M_2)| + |N(M_1) \cap N(M_2)|]$, where $N(M_1)$ and $N(M_2)$ are sets of the interactors of M_1 and M_2 respectively. The terms Δ is the symmetrical difference between the two sets (Lin, 1998). In Fig. 12 in Supplementary Information, $M_1 =$ 'a', $M_2 =$ 1'c', N('a') = 4, N('c') = 3, $|N(`a')\Delta N(`c')| = 4$, $|N(`a') \cup N(`c')| = 6$, $|N(`a') \cap N(`c')| = 1$, Czekanowski–Dice ('a', 'c') = 0.5.

3.6. Index mining

Index mining is based on similarity measures that are used for comparing two or more sites, with respect to their feature overlap. For example, a feature can be identified on the basis of the presence or absence of a particular atom in the overall metabolic structure. All the features represent variations over certain parameters, such as presence of structural features (Wolda, 1981). We can look for compounds that are most similar to the query compound which is followed by rank determination of each of these compounds in the database. A similarity measure is defined to quantify the degree of similarity (Brun et al., 2004). Let us consider two objects 'A' and 'B', where f_1 is the number of features present in 'A' but absent in 'B', f_2 is the number of features absent in 'A' but present in 'B', f_3 is the number of features common to both 'A' and 'B', and f_4 is the number of features absent from both 'A' and 'B'. Thus, f_3 and f_4 measure the present and the absent matches, respectively, i.e., similarity; while f_1 and f_2 measure the corresponding mismatches, i.e., dissimilarity. Table 10 (in Supplementary Information) represents a list of measures used for index mining. For example, let us consider three metabolites α-D-*glucose* ($C_6H_{12}O_6$), α-D-*glucose*-1P ($C_6H_{13}O_9P$) and pyruvate ($C_3H_4O_3$). Assuming that the features present in the molecules are in terms of their bits, we select the bit representation of the atoms present in the molecules as C = 1000, O = 1010, H = 1100, P = 1110. Thus, the corresponding feature-based representation for α-D-*glucose* is 100010 001000 100010 001000 110011 001100 110011 001100 110011 001100 110011 001100 101010 101010 101010 101010, that for α-D-*glucose*-1P is 100010 001000 100010 001000 110011 001100 110011 001100 110011 001100 110011 001100 110010 101010 101010 101010 101010 101010 10101110 and for *pyruvate* 100010 001000 110011 001100 110010 101010 1010. Then, comparing α-D-*glucose* and α-D-*glucose*-1P, we have $f_1 = 1$; $f_2 = 21$; $f\ _3 = 94$; $f_4 = 22$ and Russell–Rao index = 0.68 (Rao, 1948). For comparison of α-D-*glucose* and *pyruvate*, we have $f_1 = 31$; $f_2 = 6$; $f_3 = 11$; $f_4 = 30$ and Russell–Rao index = 0.14. Thus, the similarity measure is higher for the former suggesting that α-D-*glucose* is more similar to α-D-*glucose*-1P than *pyruvate*. This also suggests that tendency of forming α-D-*glucose*-1P from α-D-*glucose* is more than that of *pyruvate*.

3.7. An ensemble approach for path mining

In the previous sub-sections, we discussed certain methods in path mining having their own advantages, disadvantages and application areas. Here, we discuss an ensemble of the above methods for illustrating their combined role in solving some critical biological problem. The problem that we are considering here is 'automated link prediction' of metabolic pathways based upon certain path mining based parameters. Automated link prediction is an ab-initio approach where initial input for modeling a metabolic pathway is a set of metabolites which are incomplete and randomly chosen (especially in case of diseased pathways). Only a certain amount of progress has been done for in this regard, as biological moieties are extremely complex and may have more than one relation. Furthermore, given a set of metabolites it is quite cumbersome to predict their relations. Thus, one of the preliminary steps for automated link prediction is identifying certain criterion for relating two or more biological moieties (in this case, metabolites) so that they can be linked together. We assume that given a set of metabolites, the probability that one is converted to another (thus giving rise to a reaction link), is higher if they are structurally similar (Wishart, 2010). Moreover, there should be some standard nomenclature for predicting the structural similarity among certain metabolites. Here, we use the concept of *SMILES* strings-based representation for the metabolites. We would also like to foretell that only considering the structural features of two or more metabolites for constructing a reaction link may not be completely feasible, as the biological significance of such a relation needs to be taken into account. But, to restrict ourselves to the path mining part, we are only discussing a possible method of using an ensemble of path mining approaches to analyze links among given metabolites, which can be further used to predict the biological significance for constructing the complete metabolic pathway.

We consider an ensemble of five path mining approaches, namely, key-based, pattern-based, neighborhood-based, index-based and regular expression-based mining. We use the fusion of first four approaches for performing the analysis and the fifth one for validating our results (Fig. 13 in Supplementary Information). We initiate our analysis by performing a key-based search in pathway repositories, namely, Kyoto Encyclopedia for Genes and Genomes (KEGG) (Kanehisa et al., 2012), HumanCyc (Romero et al., 2004) and BRENDA (Schomburg et al., 2013) respectively. Our intention is to select a metabolic pathway, for which we can implement our method. We have selected a known metabolic pathway with the simple purpose to validate whether the links created by our strategy actually exists in the original pathway or not. This is essential for standardization of our strategy, so that it could be used for predicting links in case of unknown or incomplete datasets (Barba et al., 2013). Thus, the initial query or 'key' used for searching in the three repositories was 'at + least + 18 + metabolites AND homo + sapiens'. This key would search for all pathway entries having *at least* 18 *metabolites* belonging to *Homo sapiens*. We linked both these keywords using a logical operator 'AND'. The search revealed many metabolic pathways like *purine* metabolism, *pyrimidine* metabolism, *pentosephosphate* pathway, to name a few. We selected *pentosephosphate* pathway as our dataset. It has 20 metabolites (as per KEGG), namely, D-*glucono - 1, 5 - lactose*(*a*), D-*gluconate*(*b*), D-*gluconate - 6 - P*(*c*), D-*glucono - 1, 5 - lactone - 6 - P*(*d*), β-D-*glucose - 6 - P*(*e*), α-D-*glucose - 6 - P*(*f*), β-D-*fructose — 6 — P*(*g*), D-*xylulose - 5 - P*(*h*), D-*ribulose*5 - *P*(*i*), D-*ribose*5 — *P*(*j*), D-*erythrose*4 - *P*(*k*), β-D-*fructose - 1, 6 - bis - P*(*l*), D-*ribose*(*m*), D-*ribose*1 - *P*(*n*), *PRPP*(*o*), D-*sedoheptulose*7 - *P*(*p*), D-*glyceraldehyde* - 3 -*P*(*q*), 2 - *deoxy* -D-*ribose* - 5 - *P*(*r*), 2 - *deoxy* -D-*ribose*(*s*), and 2 - *deoxy* -D-*ribose* - 1 - *P*(*t*) respectively. Their occurrence frequencies are 1, 1, 3, 2, 3, 2, 6, 3, 3, 5, 2, 3, 1, 1, 1, 2, 5, 3, 1, and 1 respectively. Next, we create four pathway sets on the basis of their frequencies as $MP_1 = \{q,r,s,t,l,k,h\}$, $MP_2 = \{g,f,l,h,k,q\}$, $MP_3 = \{j,p,i,m,n,o\}$, and $MP_4 = \{c,b,a,d,e,i\}$. Thus, $Frq = \{g,j,q\}$. These steps correspond to *pattern-based* mining approach.

Next, we identify the paths arising from metabolites g, j, and q to all others (Fig. 14 in Supplementary Information). We observe that the number of paths arising from g, j, and q are 17 each whereas the number of least separated neighbors are 5, 6 and 6 respectively. For g, the direct links are $g > f$, $g > h$, $g > k$, $g > l$, and $g > q$; for j the direct links are $j > i$, $j > m$, $j > n$, $j > o$, $j > p$, and $j > q$ and for q the direct links are $q > h$, $q > j$, $q > k$, $q > l$, $q > p$, and $q > r$ respectively. Next, we identify the topological proximity of the least separated neighbors using the Czekanowski–Dice distance for identifying the strength of their interactions (Sorensen, 1957), i.e., $(g, f) = 0.54$, $(g, h) = 0.5$, $(g, k) = 0.62$, $(g, l) = 0.71$, $(g, q) = 0.6$, $(j, i) = 0.8$, $(j, m) = 1.0$, $(j, n) = 1.0$, $(j, o) = 1.0$, $(j, p) = 0.8$, $(j, q) = 0.83$, $(q, h) = 0.6$, $(q, k) = 0.57$, $(q, l) = 0.57$, $(q, p) = 0.62$, and $(q, r) = 1.0$ respectively. Thus, we observe that the interaction strength is highest among (g, k), (g, l), (j, i), (j, m), (j, n), (j, o), and (q, j) respectively. Thus, the possibility of conversions among these metabolite pairs is more than others. These steps correspond to *neighborhood-based* mining approach (Zhang et al., 2006). We extended our approach to *index-based* mining approach, where we identified the similarity scores among these metabolite pairs (as identified from *neighborhood-based* mining) to conclude our analysis. For this purpose, we represented the metabolites into their SMILES strings as $g = \mathrm{C(C1C(C(C(O1)(CO)O)O)O)OP(=O)(O)O}$, $k = \mathrm{C(C(C(C = O)O)O)OP(=O)(O)O}$, $l = \mathrm{C(C1C(C(C(O1)(COP(=O)(O)O)O)O)O)OP(=O)(O)O}$, $m = \mathrm{C(C(C(C(C = O)O)O)O)O}$, $j = \mathrm{C(C1C(C(C(O1)O)O)O)OP(=O)(O)O}$, $q = \mathrm{C(C(C = O)O)OP(=O)([O —])[O —]}$, and $r = \mathrm{C1C(C(OC1O)COP(=O)(O)O)O}$ respectively. Next, we convert these SMILES strings into their corresponding feature-based representations, i.e., in strings of 1 and 0. Now, we use an index measure 'Russell–Rao' for identifying similarities among the feature-based representations of metabolites (Rao, 1948). The Russell–Rao values for the pairs $(g, k) = 0.23$, $(g, l) = 0.01$, $(g, m) = 0.23$, $(g, j) = 0.11$, $(g, q) = 0.37$, $(g, r) = 0.13$, $(k, l) = 0.08$, $(k, m) = 0.03$, $(k, j) = 0.05$, $(k, q) = 0.21$, $(j, q) = 0.83$, $(k, r) = 0.05$, $(l, m) = 0.31$, $(l, j) = 0.19$, $(l, q) = 0.43$, $(l, r) = 0.04$, $(m, j) = 0.01$, $(m, q) = 0.2$, $(m, r) = 0.01$, $(j, q) = 0.3$, $(j, r) = 0.03$, and $(q, r) = 0.12$ respectively. Here, we have identified the similarity scores among all pairs and not restricted to those found using neighborhood-based mining approach. The reason being, we made sure that none of the other metabolite pairs having higher similarity but lower interaction strength is ignored. We observed that highest similarity values are identified among (g, k), (g, q), (k, q), (l, q), (m, q), (j, q), and (q, r) respectively.

To validate the above predicted results we selected the *regular expression-based* mining approach. On the basis of similarity among metabolite pairs, we identified the enzymes from BRENDA and ENZYME repositories, that actually catalyzed the reactions involving these metabolites (if the reaction is actually existing in the repository). The enzymes found were ($g, k =$ *transketolase*), ($g, l =$ *6-phosphofructokinase* 1), ($g, q =$ *transaldolase*), ($k, q =$ *transaldolase*), ($l, q =$ *fructose-bisphosphate aldolase*), ($m, j =$ *ribokinase*), ($j, q =$ *transketolase*) and ($q, r =$ *deoxyribose-phosphate aldolase*) respectively. For the other metabolite pairs, no probable enzyme was found. Next, we found the patterns associated with these enzymes, such as *transketolase* $= [R - x(3) - [LIVMTA] - [DENQSTHKF] - x(5,6) - [GSN] - G - H - [PLIVMF] - [GSTA] - x(2) - [LIMC] - [GS]]$, 6-*phosphofructokinase* 1 $= [AG] - G - x(0,1) - [GAP] - x - N - AGLS - [STA] - x(2) - A - x - G - GNKA - [GS] - x(9) - G$, *transaldolase* $= [DGH] - [IVSAC] - T - [ST] - N - P - [STA] - [LIVMF](2)$, *fructose-bisphosphate aldolase* $= [LIVM] - x - [LIVMFYW] - E - G - x - [LSI] - L - K - [PA] - [SN]$, *ribokinase* $= [SAV] - [IVW] - [LVA] - [LIV] - G - [PNS] - G - L - [GP] - x - [DENQT]$, and *deoxyribose-phosphate aldolase* $= G - [LIVM] - x(3) - E - [LIV] - T - [LF] - R$ respectively (Bairoch and Apweiler, 2000). Now, we compare these enzymatic patterns for identifying possible matches among them using a scoring scheme $SScore = matchscore * num_1 + mismatchscore * num_2 + insert * num_3 + delete * num_4$, where num_i stands for the number of times that event has happened. We assume that $matchscore = +1.0$, $mismatchscore = -0.5$, $insert = +0.5$ and $delete = -0.25$ respectively. Here, we have performed an 'insert' mutation in the normal pattern for increasing the *SScore*. The *SScore* values for all compared pairs are shown in Table 11 (in Supplementary Information). We observe that *SScore* for the pairs *transketolase*, *fructose-bisphosphate aldolase*; *transketolase*, *deoxyribose-phosphate aldolase*; 6-*phosphofructokinase* 1, *deoxyribose-phosphate aldolase*; 6-*phosphofructokinase* 1, *fructose-bisphosphate aldolase*; *transaldolase*, *fructose-bisphosphate aldolase*; *fructose-bisphosphate aldolase*, *deoxyribose-phosphate aldolase* and *ribokinase*, *deoxyribose-phosphate aldolase* with values +2.0, +8.0, −1.5, −3.0, +4.0, +6.5, and +2.0 are highest among all. Furthermore, the metabolite pairs for these enzymes are g, k, j, q, l, q, g, k, j, k, q, r, g, l, g, q, and m, j respectively. Finally, we compared the result of *regular expression-based* validation and the already found results of the other four methods and find observe that the metabolite pairs whose relations could be validated were g, k, g, q, l, q, j, q, and q, r, whereas those which were found to be related but couldn't be validated were k, q and m, q respectively.

4. Interpreting metabolomics datasets using path mining approaches

The strategies discussed in the previous sections help the systems biologists to extract meaningful information from metabolomics datasets. The next task lies in interpreting these various highlighted features for understanding the internal structure and function of metabolic pathways. This section deals with elucidating some useful interpretation techniques and their role in analyzing the data generated from various path mining approaches. With the advancement of high-end metabolomic techniques, it is possible to investigate metabolism at the scale of global metabolic network, consisting of chemical compounds, biochemical reactions, enzymes and genes to name a few. Metabolic network datasets can be modeled in the form of bipartite graphs consisting of two types of nodes namely, reactions and metabolites (Gifford et al., 2001). An edge exists from a metabolite node to a reaction node if the metabolite is a substrate of the reaction, and there is an edge from a reaction node to a metabolite node if the reaction produces the metabolite. These also depict the complete linkage pattern between the set of substrates and the set of products of a reaction (Fig. 11 in Supplementary Information) (Becker and Rojas, 2001; Van Helden et al., 2002). We can also identify and extract sub-pathways from bigger metabolic networks. These metabolic pathways consist of various components like a set of metabolites ('a'–'f') and enzymes (E_1–E_3) in Fig. 12 (in Supplementary Information) and they can be mined on the basis of their inter-connectivities. In Fig. 12(a) (in Supplementary Information), there are 3 reactions having 1 enzyme each. Thus, the complete pathway consists of 3 reactions and formation of the end product ('f') depends on the precursor reactions to complete properly. Also, mining information from this simple pathway leads to calculating the 'simple path (P)' and 'scope (S)' of metabolites (Fig. 12b in Supplementary Information). A simple path is defined as the set of metabolites and their corresponding reaction links as well as enzymes that lead to completion of the reaction in a successful manner (Gifford et al., 2001). For example, in Fig. 12(a) (in Supplementary Information), a simple path can be from metabolites 'a' and 'b' to metabolite 'c', catalyzed by enzyme E_1. Similarly, scope of a metabolite is defined as the effect it has on the functionality of other metabolites (You et al., 2006). For example, in Fig. 12(b) (in Supplementary Information), unless e is fired the reaction from 'd' to 'f' won't be catalyzed by 'c'. This is defined as the scope of 'e' on 'd'.

We can also employ various computational strategies like 'divide-and-conquer' to predict and join missing links in certain disease pathways (Genc and Dogrusoz, 2004). Each metabolite of the pathway is considered in turn as a 'basis node', and metabolites that are connected to this basis node are identified against the metabolites that are not connected to it. This local model can then be applied to predict new reaction links between the basis and other metabolites. This process is iterated again and again to obtain reaction links throughout the pathway (Fig. 15 in Supplementary Information). In Fig. 15 (in Supplementary Information), taking 'd' as a basis node, other metabolites in the pathway are labeled as $+1$ or -1 depending on whether they are known to be connected or not to the basis node. The goal is then to predict the label of metabolites whose information is not known completely. The process is then repeated by labeling each metabolite in turn and predicting their connectivity (Osterman and Overbeek, 2003). For example, a sample interpretation strategy using path mining approach can be initiating with input metabolome datasets like metabolic networks, representing them quantitatively as bipartite graphs, that are composed of reaction nodes and metabolite nodes. Various structural properties can be interpreted from this representation, such as, identification of Linkage patterns, extracting sub-pathway information, identifying inter-connectivities among nodes and finding 'path' and 'scope'. We can now implement a Divide-and-Conquer strategy to identify missing links, extract 'degree centrality' information, 'closeness centrality', 'betweenness centrality' and global network connectivity. Furthermore, some important properties related to global connectivity can be identified like 'strongly connected component', 'weakly connected component' and 'giant component'.

Another path mining task is to identify the location of the metabolites in the network. Based upon this information, various structural parameters can be found out. For instance, 'degree centrality', D_a, of a metabolite 'a' is defined as the fraction of metabolites that are connected to each metabolite, i.e. $D_a = \mathrm{n} / (\mathrm{N} - 1)$, where 'n' is the number of metabolites connected to metabolite a and N is the number of metabolites in the network (Mazurie et al., 2010). The measure, called 'closeness centrality', C_a, not only considers the directly connected metabolites, but also the metabolites connected with it through other metabolites, i.e. $C_a = (\mathrm{N} - 1) / \sqrt{[d(a,b)]}$, where 'a' and 'b' represent two metabolites (Raaf and Messabih, 2010). Furthermore, 'betweenness centrality', B_a, of a metabolite 'a' is defined as the fraction of the number of the shortest paths that go through the metabolite (Raaf and Messabih, 2010) (Fig. 16 in Supplementary Information). Also, global network connectivity can be found out in any metabolic network by applying certain graph-based approaches. For instance, two concepts are regularly used to describe the network connectivity, 'strongly connected component' and 'weakly connected component'. A subset of metabolites in a network is called a strongly connected component if from every metabolite of the subset we can reach every other metabolite belonging to the same subset through a directed pathway (Ding et al., 2009). A directed pathway is weakly connected if replacing all of its directed links with undirected links produces a undirected connected pathway (Zhao et al., 2007). The largest component in the network in terms of the largest fraction of the metabolites is known as 'giant component (GC)' (Ding et al., 2009). In this case, there are two associated subsets, namely, the one in which all the metabolites can be converted to metabolites in the 'GC' i.e. 'incoming' and another in which all of the metabolites can be produced from metabolites in the 'GC' i.e. 'outgoing'. All the other metabolites that are not connected with metabolites in the 'GC' form 'isolated metabolite subsets'. This structure is known as 'bow-tie' structure of network (Fig. 17 in Supplementary Information) (Zhao et al., 2007). For example, metabolic networks in bacteria have resemblance to bow-tie organization. For instance, certain nutrient re-sources are catabolized to produce some activated carriers like ATP, NADH and NADPH and 12 precursor metabolites like *glucose*-6-*phosphate*, *fructose* 6-*phosphate*, *phosphoenolpyruvate* and *pyruvate*, which are then synthesized into roughly 70 larger building blocks like amino acids, nucleotides, fatty acids and sugars. In this case, precursors and carriers act as two 'knots' of separate bow ties that are both fed by catabolism, whereas the former takes part locally to the biosynthesis of universal building blocks, and the latter provides energy to the cell.

5. Case study discussion

Now, we elaborate on a case study where we have used graph-based path mining for 'back-validating' some metabolic networks from a structural point of view. A structure-based study of a metabolic network deals with identifying sub-networks, paths, scope, similarities among metabolites, links among metabolites, centrality measures, to name a few. 'Back-validation' is a method of back-tracking, where certain known information of metabolic networks is considered as initial input and certain graph-based methods are implemented for validating some already known information. For this purpose, we perform an analysis of identifying certain essential metabolites and their ongoing interactions involved in the functioning of *INS* and *GAD* genes, believed to play important roles in the onset of Type 1 *Diabetes mellitus* (T1D) in *H. sapiens* (Reaven, 1988). It is a disorder characterized by high blood sugar, resulting in improper production of insulin. It also results in failure and dysfunction of multiple organs. Moreover, till date around 250 genes have been studied for analyzing their role with T1D, one of which is anti-Glutaric acid decarboxylase (*GAD*) leading autoimmune processes for β-cell destruction. Similarly, insulin gene (*INS*) is the second well established susceptibility locus in *Diabetes mellitus*, contributing to about 10% toward T1D susceptibility (Reaven, 1988). We have analyzed the onset of T1D in *H*. sapiens by studying the role of *GAD* and *INS* genes in four metabolic pathways, namely, *glutamate* metabolism, β - *alanine* metabolism, *taurine* & *hypotaurine* metabolism and *butanoate* metabolism, for the purpose of identifying the essentiality of metabolites and the level of interaction present within the metabolites from a path mining perspective (Fig. 18 in Supplementary Information).

The metabolic pathways involved in the functioning of *GAD* and *INS* genes were identified from KEGG repository using 'key-based' mining approach. The approach followed here takes two routes for analysis. The first route (Path 1 or pre-validation analysis) deals with analyzing the structural dynamics of metabolic pathways using a connectivity-pattern identification based approach, followed by calculating the centrality features. The second route (Path 2) initiates by identifying certain distance-based features among the metabolites, followed by implementation of a 'Divide-and-conquer' (*DC*) strategy. The *DC* strategy groups metabolites based on similarity features, followed by identifying those metabolites whose participation is necessary for the overall functioning of the metabolic pathway. This is done by identifying *cuts* in metabolic pathways (discussed in details later) (Cormen et al., 2009). We have also kept in mind that the results interpreted by us should have a biological significance associated with them. For this purpose, some biological validation methods have been implemented.

The foremost task in analyzing the internal dynamics of metabolic pathways from structural point-of-view is analyzing the connectivity patterns among metabolites. For this purpose, we identified degree distribution, average degree, network density, average clustering coefficient, betweenness centrality and closeness centrality for all the four metabolic pathways. The reason for identifying these graph-based parameters is to validate whether the metabolic pathway datasets taken from KEGG could be actually used for further analysis and have a significance level associated with it. A metabolic pathway dataset is significant if it has lesser number of metabolites having high degree, as well as higher clustering coefficient, whereas the betweenness and closeness centrality values should follow a power law, i.e., the metabolic pathways are scale-free in nature (Jeong et al., 2000). These are some pre-validation analysis that we need to do before we actually initiate our work. Table 12 in (Supplementary Information) provides information about the average degree, network density and average clustering coefficient of the four metabolic pathways. We observe that average degree for *taurine & hypotaurine* and *butanoate* metabolisms are higher, i.e., 2.0 whereas that of *glutamate* and β - *alanine* metabolisms are 1.8 and 1.7 respectively. This illustrates that the robustness of *taurine & hypotaurine* and *butanoate* metabolisms is higher as a greater number of parallel reactions are present for the production of most metabolites, leading to the conclusion that in case of abnormal situation, like that of an infection spread, these metabolic pathways are capable of self-immunization and self-resistance (Choisy et al., 2007). Similarly, the average clustering coefficient for *glutamate* and *taurine & hypotaurine* metabolisms is higher, i.e., 0.02 and 0.01 whereas that of the other two metabolic pathways are lower. This illustrates that the degree of compactness for *glutamate* and *taurine & hypotaurine* metabolism is higher, i.e., in case of internal or external perturbation, their metabolic architecture is difficult to fragment. Furthermore, network density of a metabolic pathway specifies the localization of metabolites, i.e., higher the network density, more metabolic groups are existing (Klamt and von Kamp, 2009). We observe that the network density of *glutamate* and *taurine & hypotaurine* metabolism is higher, i.e., a large number of metabolite clusters are present, whereas a lesser number of isolated metabolites are existent (Figs. 19–22 in Supplementary Information). This signifies that there are a lesser number of metabolites which participate in large number of reactions. We can finally conclude from this pre-validation step that out of all four metabolic pathways *taurine & hypotaurine* metabolism is architecturally more compact and resistant to any kind of external or internal perturbation, as it has a higher average degree, as well as better average clustering coefficient and network density (Klamt and von Kamp, 2009).

We now initiate with Path 2 by identifying path-distances among metabolites involved in the reactions that are responsible for highlighting the functionality of *GAD* and *INS* genes. Table 13 (in Supplementary Information) provides a list of reactions occurring in the four metabolic pathways that play an important role in the functioning of *GAD* and *INS* genes. We can observe that for the four reactions identified in *glutamate* metabolism, L-glutamate is common for all whereas the other metabolites are *L-glutamine*, 2-*oxoglutarate*, 4-*aminobutanoate* and *L*-1 - *pyrroline* - 5 - *carboxylate*. Similarly, for the four reactions identified in *β - alanine* metabolism the metabolite that is found to be common for all is *L-aspartate*, whereas others are *L-asparagine*, *oxaloacetate*, *L-arginino succinate*, and *β - alanine*. Likewise, in *taurine & hypotaurine* metabolism the common metabolite is *taurine* whereas others are *L-cysteate*, *taurocholate*, *hypotaurine* and 5-*glutamyl-taurine*. Finally, for *butanoate* metabolism only two reactions are identified, where no common metabolite exist. The metabolites found are *acetyl CoA*, *acetoacetyl CoA*, *L-glutamate* and 4-*butanoate* respectively which are used for our further studies. The next step is identifying the *Hamming distance* (*HD*) (Hamming, 1950) as well as *Tanimoto coefficient* (*TC*) (Tanimoto, 1957) for all metabolite pairs. We first create an *incidence matrix* (*IM*) with metabolites as 'rows' and 'reactions' as columns, with a '1' placed in the matrix if a reaction occurs between the corresponding metabolites (Tables 14–17 in Supplementary Information). *HD* is calculated by comparing each row of the *IM*. For two rows in the *IM*, *HD* is 'the number of positions where values (1 or 0) are different' (Table 18 in Supplementary Information). The use of *HD* as a basis ensures a biologically significant pattern detection (Brandes et al., 2004). The metabolites so grouped are separated by same number of substitutions. This group when compared to others separated by same distance provides insight into the internal complexity of metabolic pathway. Similarly, to detect similarity among the compared metabolites, we also find *Tanimoto coefficient* (*TC*) (Table 18 in Supplementary Information). We observe that for *glutamate* metabolism, *HD* is 3 for all pairs whereas higher *TC* is found for (*L-glutamate*, *L-glutamine*) = 0.78, (*L-glutamate*, *L*-1-*pyrroline* 5-*carboxylate*) = 0.64, (*L-glutamate*, 4-*aminobutanoate*) = 0.61. In case of *β - alanine HD* is 3 for all pairs whereas higher *TC* is found for (*L-aspartate*, *L-asparragine*) = 0.75, (*L-aspartateβ - alanine*) = 0.64. Similarly, for *taurine & hypotaurine* metabolism *HD* is found to be 3 for all, whereas *TC* is found for (*taurine,hypotaurine*) = 0.82, (*taurine*, L-cysteate) = 0.75. Lastly, for *butanoate* metabolism *HD* is found to be 1 for all, whereas *TC* is found for (*acetyl - CoA*, *acetoacetyl - CoA*) = 0.79, (*L-glutamate*, 4-*butanoate*) = 0.52.

Next, we use the *HD* and *TC* values for the metabolite pairs for creating 'strong' and 'weak links' using 'divide-and-conquer' strategy (Cormen et al., 2009). 'Strong' links (+1) are those which are most essential for *GAD* and *INS* genes functionality, whereas 'weak' links (−1) are those which may be of lesser importance as compared to strong links. For *glutamate* metabolism, we consider *L-glutamate* as the 'basis node' initially. We assume that all metabolites that are paired with *glutamate* have 'weak' links initially. On the basis of higher values for *HD* and *TC* a weak link is converted to a strong link, whereas a lower *HD* and *TC* value result in a weak link. We observe that *L-glutamine* has a strong link with *L-glutamate*, whereas *L*-1-*pyrroline*-5-*carboxylate* is weakly linked to *L-glutamate*. Similarly, considering *L-glutamine*, 2-*oxoglutarate* and 4-*aminobutanoate* as the 'basis node' creates all weak links with their corresponding pairs (Fig. 23 in Supplementary Information). Likewise, for *β - alanine* metabolism considering *L-aspartate* as 'basis node' creates strong links with *L-asparagine* as well as *β - alanine* and weak links with *oxaloacetate* as well as *L-argininosuccinate* respectively. Similarly, considering *L-asparagine*, *oxaloacetate* and L-argininosuccinate as 'basis node' creates all weak links with their corresponding pairs (Fig. 24 in Supplementary Information). In case of *taurine & hypotaurine* metabolism considering *taurine* as the 'basis node' creates strong links with *hypotaurine* and *L-cysteate* whereas weak links with 5-*glutamyl-taurine*. Similarly, considering *L-cysteate*, *taurocholate* and *hypotaurine* as 'basis node' creates all weak links with their corresponding pairs (Fig. 25 in Supplementary Information). Lastly, for *butanoate* metabolism considering *acetyl CoA* as 'basis node' creates a weak links with *acetoacetyl CoA*, *L-glutamate* and 4-*butanoate*. Similarly, considering *acetoacetyl CoA* as 'basis node' creates a weak links with all, whereas for *L-glutamate* to be considered as 'basis node' creates strong link with 4-*butanoate* (Fig. 26 in Supplementary Information). On the basis of link type metabolite groups were formed, namely, *Group* 1 = {*L-glutamate*, *L-glutamine*, *L*-1-*pyrroline*-5-*carboxylate*}, *Group* 2 = {*β-alanine*, *L-aspartate*, *oxaloacetate*, *L-arginino-succinate*}, *Group* 3 = {*taurine*, *hypotaurine*, *L-cysteate*, 5-*glutamyl-taurine*}, and *Group* 4 = {*L-glutamate*, 4-*butanoate*, *acetyl-CoA*}.

The metabolites having highest degree in each group is identified, namely, *L-glutamate* = 4 (*Group* 1), *L-aspartate* = 4 (*Group* 2), *Taurine* = 4 (*Group* 3), *acetyl-CoA* = 1, and *L-glutamate* = 1 (*Group* 4). The highest degree is calculated to identify the 'cuts' within the groups. A 'cut' is defined as the minimum number of links and/or metabolites that need to be removed to segment a metabolic pathway into two or more sub-pathway having similar size (Klamt and Gilles, 2013). Thus, for *Group* 1 the cuts are identified between *L*-1-*pyrrolime*5-*carboxylate* → *L*-glutamate, *L*-glutamine → *L*-glutamate whose removal causes a strong change in gene functionality and 2-*oxoglutarate* → *L-glutamate*, 4-*aminobutanoate* → *L-glutamate* whose removal causes a weak change in gene functionality (Fig. 27(a) in Supplementary Information). Thus, for *Group* 2 the cuts are identified between *β-alanine* → *L-aspartate*, *L-asparagine* → *L-aspartate* whose removal causes a strong change in gene functionality and *oxaloacetate* → *L-aspartate*, *L-arginino-succinate* → *L-aspartate* whose removal causes a weak change in gene functionality (Fig. 27(b) in Supplementary Information) (Reaven, 1988). Likewise, in *Group* 3 the cuts are identified between *taurocholate* → *taurine*, *hypotaurine* → *taurine* whose removal causes a strong change in gene functionality and 5 - *glutamyl* - *taurine* → *taurine*, *L-cysteate* → *taurine* whose removal causes a weak change in gene functionality (Fig. 27(c) in Supplementary Information). Finally, in *Group* 4 the cuts are identified between *acetylCoA* → *acetoacetylCoA*, *L-glutamatetarrow*4 - *butanoate* whose removal causes a strong change in gene functionality (Fig. 27(d) in Supplementary Information). The validation for 'divide-and-conquer' strategy, and 'cuts' is done by identifying the 'strongly connected components' for the four metabolic pathways. A strong component in a metabolic pathway consists of functionally significant metabolite groups, which have strong intra-interactions leading to essential gene functionality whereas a weak interaction leads to non-essential gene functionality (Gerlee et al., 2009). In case of *glutamate* metabolism, a single strong component exists, i.e., *D-glutamate*, *D-glutamine*, L-glutamine, 2-*oxoglutarate*, *L-glutamate*, *L-glutamine* (Fig. 28(a) in Supplementary Information). In case of β - *alanine* metabolism, 5 strong components exist, namely, 4 - *aminobutanoate*, 4 - *aminobutanal*, *acetyl CoA*, *malonyl CoA*, *spermindine*, *spermine*, β-alanine, *L-aspartate*; β-*alanine*, 3-*oxopropionate*, and *L-asparagine*, *L-aspartate*; 4-*aminobutanyl CoA*, *propenoyl CoA* (Fig. 28(b) in Supplementary Information). In case of *taurine & hypotaurine* metabolism, 5 strong components exist, namely, *succinate*, *succinate semialdehyde;* 4-*aminobutanoate*, *L-glutamate*, 2-*oxoglutarate*, 2-*hydroxyglutarate*, *taurocholate*, *taurine*, *hypotaurine*, *taurine*, and 2-*acetolactate*, *thiamine diphosphate* (Fig. 28(c) in Supplementary Information). Finally, in *butanoate* metabolism, only 1 strong component exists, namely, *butanoate*, *L-glutamate; acetoacetyl CoA*, *acetyl CoA* (Fig. 28(d) in Supplementary Information). Thus, we can observe that the most essential metabolites, namely, *L-glutamate*, *L-glutamine*, 2-*oxoglutarate*, 4-*aminobutanoate*, *L*-1-*pyrroline* 5-*carboxylate* (in *glutamate* metabolism), *L-aspartate*, *L-asparagine*, *oxaloacetate*, *L-arginino succinate*, β-*alanine* (in β - *alanine* metabolism), *taurine*, *L-cysteate*, *taurocholate*, *hypotaurine*, 5-*glutamyl-taurine* (in *taurine & hypotaurine* metabolism) and *acetyl CoA*, *acetoacetyl CoA*, *L-glutamate*, 4-*butanoate* (in *butanoate* metabolism), that are actively responsible for the functioning of tGAD and *INS* genes are present in strong components (Gerlee et al., 2009).

6. Conclusions

In this article we have discussed a number of graph-based approaches for representing and analyzing metabolic networks, and some issues regarding the analysis of metabolome network using path mining techniques. These strategies give in-depth insight into comprehending the molecular mechanisms of a particular organism for correlating the genome with molecular physiology. Biological processes differ from chemical experiments in that cells are able to regulate the concentration and/or activity of their enzymes and transporters Thus, a comprehensive analysis should integrate metabolic, genetic and physiological information. Also, a set of objective rules are needed to build network graphs that can be further used to establish a proper correlation among metabolites. Thus, a pathway inference framework based on the functional annotations of enzymes participating in a pathway can be presented. Given a pathway, one can create a pathway functionality template for each known organism-specific version of the pathway. Next, using the concept of graphs, frequent pathway functionality template patterns can be discovered that can be used for analysis of the pathway.

We focussed on some path mining approaches like pattern-based mining, process-based mining and multi-method mining. All these approaches need to be executed one after the other for achieving a particular task. Furthermore, we discussed about some other recent methods related to path mining, namely, grammar-based, regular expression-based, key-based, pattern-based, neighborhood-based and index-based mining respectively. Moreover, we used some of the path mining strategies for elucidating certain structural features of metabolic networks. We have also highlighted the importance of path mining using graph-based models. We also discussed a case study using glycolysis metabolism in *H. sapiens*. We used similarity measures for predicting similarity among metabolites, followed by predicting links and sub-networks. We also found various path-based measures along with scope, directionality, contribution and functioning of metabolites for easier analysis of networks. Moreover, an important aspect of using graph-based approaches in metabolic networks like applying filters on the metabolomics data. Some of these are withdrawing reactions involving large molecules such as proteins, followed by removing reactions involved in unidentified pathways and lastly avoiding ubiquitous compounds, like ATP and NADH forming hubs and leading to difficulty in discriminating them from useful metabolites. Path-mining also help in building hypothetical models and for analyzing topological parameters in metabolome networks. Lastly, it can also be used to predict the outcome of various processes by performing perturbations within the network and choosing an appropriate path-finding technique depends on the type of network and the objective.

Conflict of interest

The authors declare no conflicts of interest.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2013.10.056.

References

Abulaish, M., Dey, L., 2005. An ontology-based pattern mining system for extracting information from biological texts. Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, b3642 420–429.

Adam, N.R., Janeja, V.P., Atluri, V., 2004. Neighborhood based detection of anomalies in high dimensional spatio-temporal sensor datasets. SAC '04 Proceedings of the 2004 ACM Symposium on Applied Computing, New York, NY, USA.

Agarwal, R.C., Agarwal, C.C., Prasad, V.V.V., 2001. A tree projection algorithm for generation of frequent itemsets. J. Parallel Distrib. Comput. 61 (Suppl. 3), 350–371.

Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J., 2009. Describing linked datasets: on the design and usage of voiD, the 'Vocabulary of Interlinked Datasets'. 2nd International Workshop on Linked Data on the Web, Madrid, Spain.

Angles, R., Gutierrez, C., 2005. Querying RDF data from a graph database perspective. Semantic Web Res. Appl. 346–360.

Angles, R., Gutierrez, C., 2008. Survey of graph database models. ACM Comput. Surv. 22 (Suppl. 1), 346–360.

Ayres, J., Gehrke, J., Yiu, T., Flannick, J., 2002. Sequential pattern mining using bitmaps. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, pp. 429–435.

Bairoch, A., Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28 (Suppl. 1), 45–48.

Barba, M., Dutoit, R., Legrain, C., Labedan, B., 2013. Identifying reaction modules in metabolic pathways: bioinformatic deduction and experimental validation of a new putative route in purine catabolism. BMC Syst. Biol. 7, 99.

Becker, M.Y., Rojas, I., 2001. A graph layout algorithm for drawing metabolic pathways. Bioinformatics 17 (Suppl. 5), 461–467.

Berman, H.M., et al., 2000. The Protein Data Bank. Nucleic Acids Res. 28 (Suppl. 1), 235–242.

Berman, H.M., Westbrook, J., Feng, Z., Iype, L., Schneider, B., Zardecki, C., 2003. The nucleic acid database. Methods Biochem. Anal. 44, 199–216.

Bichindaritz, I., Akkineni, S., 2005. Concept mining for indexing medical literature. Mach. Learn. Data Min. Pattern Recog. 3587, 682–691.

Bille, P., 2006. New algorithms for regular expression matching. Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, Reykjavik, Iceland.

Bille, P., Farach-Colton, M., 2005. Fast and compact regular expression matching. Theor. Comput. Sci. 409 (Suppl. 3), 57–71.

Brandes, U., Dwyer, T., Schreiber, F., 2004. Visual understanding of metabolic pathways across organisms using layout in two and a half dimensions. J. Integr. Bioinform. 1 (Suppl. 1), 2004.

Brun, C., Herrmann, C., Gunoche, A., 2004. Clustering proteins from interaction networks for the prediction of cellular functions. BMC Bioinforma. 5, 95.

Calders, T., Goethals, B., 2002. Mining all non-derivable frequent itemsets. Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery. Springer.

Choisy, M., Gugan, J.F., Rohani, P., 2007. Mathematical modeling of infectious diseases dynamics. Encyclopedia of Infectious Diseases: Modern Methodologies.John Wiley & Sons, Inc. 379404.

Ciaccia, P., 1998. Processing complex similarity queries with distance-based access methods. Adv. Database Technol. — EDBT 98 (1337), 9–23.

Cook, D.J., Holder, L.B., 2000. Graph-based data mining. IEEE Intell. Syst. 15 (Suppl. 2), 32–41.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2009. Introduction to Algorithms. MIT Press 1180.

Cummings, M.D., Maxwell, A.C., DesJarlais, R.L., 2007. Processing of small molecule databases for automated docking. Med. Chem. 3 (Suppl.), 107–113.

Dimitropoulos, D., Ionides, J., Henrick, K., 2006. Using PDBeChem to search the PDB ligand dictionary. Current Protocols in Bioinformatics. John Wiley & Sons, p. 14.3.1-.

Ding, D.W., Ding, Y.R., Li, L.N., Cai, Y.J., Xu, W.B., 2009. Structural and functional analysis of giant strong component of *Bacillus thuringiensis* metabolic network. Braz. J. Microbiol. 40 (Suppl. 2), 411–416.

Dries, A., Nijssen, S., 2007. Analyzing graph databases by aggregate queries. MLG'10 Proceedings of the Eighth Workshop on Mining and Learning with Graphs, Washington DC, U.S.A. pp. 37–45.

Ferro, A., Giugno, R., Mongiov, M., Pulvirenti, A., Skripin, D., Shasha, D., 2008. GraphFind: enhancing graph searching by low support data mining techniques. Bioinformatics 9 (Suppl. 4), S10.

Flesca, S., Furfaro, F., Greco, S., 2006. A graph grammars based framework for querying graph-like data. Data Knowl. Eng. 59 (Suppl. 3), 652–680 (Special issue: ER 2003 archive).

Frenz, C.M., 2007. Deafness mutation mining using regular expression based pattern matching. BMC Med. Inform. Decis. Mak. 7, 32.

Garofalakis, M.N., Rastogi, R., Shim, K., 1999. SPIRIT: sequential pattern mining with regular expression constraint. Proc. Intl Conf. Very Large Databases (VLDB 1999), pp. 223–234.

Genc, B., Dogrusoz, U., 2004. A constrained, force-directed layout algorithm for biological pathways. Graph Drawing 5 (Suppl. 4), 314–319.

Gerlee, P., Lizana, L., Sneppen, K., 2009. Pathway identification by network pruning in the metabolic network of *Escherichia coli*. Bioinformatics 25 (Suppl. 24), 3282–3288.

Gifford, E., Johnson, M., Tsai, C., 2001. A graph-theoretic approach to modeling metabolic pathways. J. Comput. Aided Mol. Des. 5 (Suppl. 4), 303–322.

Goertzel, B., 2006. Patterns, hypergraphs & embodied general intelligence. IEEE World Congress on Computational Intelligence, Vancouver, BC, Canada, pp. 455–458.

Hamming, R.W., 1950. Error detecting and error correcting codes. Bell Syst. Tech. J. 29 (Suppl. 2), 147–160.

Han, J., Cheng, H., Xin, D., Yan, X., 2007. Frequent pattern mining: current status and future directions. Data Min. Knowl. Disc. 15 (Suppl. 1), 55–86.

Helms, B., Eben, K., Shea, K., Lindemann, U., 2009. Graph grammars — a formal method for dynamic structure transformation. 11th International design strictire matrix conference, DSM'09, Greenville, South Carolina, U.S.A. pp. 93–103.

Hirsh, H., 2005. Evaluation of biomedical text-mining systems: lessons learned from information retrieval. Brief. Bioinform. 6 (Suppl. 4), 344–356.

Hirsh, H., 2008. Data mining research: current status and future opportunities. Stat. Anal. Data Min. 1 (Suppl. 2), 104–107.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G., Vriend, G., 1992. A database of protein structure families with common folding motifs. Protein Sci. 1 (Suppl. 12), 1691–1698.

Hu, X., Wu, D.D., 2007. Data mining and predictive modeling of biomolecular network from biomedical literature databases. IEEE/ACM Trans. Comput. Biol. Bioinform. 4 (Suppl. 2), 251–263.

Inokuchi, A., Washio, T., Motoda, H., 2003. Complete mining of frequent patterns from graphs: mining graph data. Mach. Learn. 50 (Suppl. 3), 321–354.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.L., 2000. The large-scale organization of metabolic networks. Nature 407 (Suppl. 6804), 411–416.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 12, 2577–2637.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M., 2012. KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res. 40, D109–D114.

Klamt, S., Gilles, E.D., 2013. Minimal cut sets in biochemical reaction networks. Bioinformatics 20 (Suppl. 2), 226–234.

Klamt, S., von Kamp, A., 2009. Computing paths and cycles in biological interaction graphs. BMC Bioinforma. 10, 181.

Krallinger, M., Valencia, A., 2005. Text-mining and information-retrieval services for molecular biology. Genome Biol. 6, 224.

Kramer, S., De Raedt, L., Helma, C., 2001. Molecular feature mining in HIV data. KDD-2001 Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, Washington DC, U.S.A. pp. 136–143.

Kriegel, H., Borgwardt, K., Kroger, P., Pryakhin, A., Schubert, M., Zimek, A., 2007. Future trends in data mining. Data Min. Knowl. Disc. 15 (Suppl. 1), 87–97.

Kuramochi, M., Karypis, G., 2001. Frequent subgraph discovery. Proceedings of IEEE International Conference on Data Engineering. IEEE, Computer Society, Washington DC, U.S.A., pp. 313–320.

Laskowski, R.A., Hutchinson, E.G., Michie, A.D., Wallace, A.C., Jones, M.L., Thornton, J.M., 1997. PDBsum: a Web-based database of summaries and analyses of all PDB structures. Trends Biochem. Sci. 22, 488–490.

Laskowski, R.A., Watson, J.D., Thornton, J.M., 2005. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res. 33, W89–W93.

Li, M.H., Ung, P.M., Zajkowski, J., Garneau-Tsodikova, S., Sherman, D.H., 2009. Automated genome mining for natural products. BMC Bioinforma. 10, 185.

Lin, D., 1998. An information-theoretic definition of similarity. ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning, San Fransisco, U.S.A. pp. 296–304.

Liu, T., Lin, Y., Wen, X., Jorissen, R.N., Gilson, M.K., 2006. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucl. Acids Res. 35 (Suppl.1), D198–D201.

Lu, R., Sadiq, S., 2006. Managing process variants as an information resource. Bus. Process. Manage. 9 (Suppl. 1), 426–431.

Martinez-Bazan, N., Muntes-Mulero, V., Gomez-Villamor, S., Nin, J., Sanchez-Martinez, M., Larriba-Pey, J., 2007. Dex: high-performance exploration on large graphs for information retrieval. Proceedings of the Sixteenth ACM Conference on Conference on information and Knowledge Management, Lisbon, Portugal, pp. 573–582.

Mazurie, A., Bonchev, D., Schwikowski, B., Buck, G.A., 2010. Evolution of metabolic network organization. BMC Syst. Biol. 4, 59.

Myers, E.W., Miller, W., 1989. Approximate matching of regular expressions. Bull. Math. Biol. 51 (Suppl. 1), 5–37.

Nam, J.M., 2007. Comparison of validity of assessment methods using indices of adjusted agreement. Stat. Med. 26 (Suppl. 3), 620–632.

Netzer, M., et al., 2012. A network-based feature selection approach to identify metabolic signatures in disease. J. Theor. Biol. 310, 216–222.

Osterman, A., Overbeek, R., 2003. Missing genes in metabolic pathways: a comparative genomics approach. Curr. Opin. Chem. Biol. 7, 238–251.

Raaf, H., Messabih, B., 2010. Betweenness centrality of event graph application to metabolic network modelled by elementary net system. J. Appl. Sci. 10 (Suppl. 15), 1610–1615.

Rao, C.R., 1948. The utilization of multiple measurements in problems of biological classification. J. R. Stat. Soc. Ser. B 10, 159–193.

Reaven, G.M., 1988. Role of insulin resistance in human disease. Diabetes 37, 1595–1607.

Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M., Karp, P.D., 2004. Computational prediction of human metabolic pathways from the complete human genome. Genome Biol. 6 (Suppl. R2), 1–17.

Rosemann, M., Zur Muehlen, M., 2000. Workflow-based process monitoring and controlling — technical and organizational issues. Proceedings of the 33rd Hawaii international conference on system science (HICSS-33). IEEE Computer Society Press, p. 6032.

Rubin, V., Gnther, C., van der Aalst, W., Kindler, E., van Dongen, B., Schfer, W., 2007. Process mining framework for software processes. Softw. Process. Dyn. Agility 4470 (Suppl. 1), 169–181.

Saha, B., Hoch, A., Khuller, S., Raschid, L., Zhang, X., 2010. Dense subgraphs with restrictions and applications to gene annotation graphs. Res. Comput. Mol. Biol. 6044, 456–472.

Sander, C., Schneider, R., 1994. The HSSP database of protein structure–sequence alignments. Nucleic Acids Res. 22 (Suppl. 17), 3597–3599.

Schomburg, I., et al., 2013. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res. 41 (Database Issue), D764–D772.

Seth, S., Rping, S., Wrobel, S., 2009. Metadata extraction using text mining. Stud. Health Technol. Inform. 147, 95–104.

Smalter, A.M., Huan, J., Lushington, G.H., 2008. Chemical compound classification with automatically mined structure patterns. Proc. Asia Pac. Bioinform. Conf. 6, 39–48.

Sorensen, T., 1957. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. K. Dan. Vidensk. Selsk. 5 (Suppl. 4), 1–34.

Tagari, M., Newman, R., Chagoyen, M., Carazo, J.M., Henrick, K., 2002. New electron microscopy database and deposition system. Trends Biochem. Sci. 27 (Suppl. 11), 589.

Tan, P., Kumar, V., Srivastava, J., 2002. Selecting the right interestingness measure for association patterns. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM New York, NY, U.S.A. , pp. 32–41.

Tanimoto, T., 1957. An elementary mathematical theory of classification and prediction. Internal IBM Technical Report, 8, p. 12.

Thornton, J.M., 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res. 32, D129–D3329.

Tomasulo, P., 2002. ChemIDplus-super source for chemical and drug information. Med. Ref. Serv. Q. 21 (Suppl.1), 53–59.

Van der Aalst, W., Rubin, V., Verbeek, H., van Dongen, B., Kindler, E., Gnther, C., 2009. Process mining: a two-step approach to balance between underfitting and overfitting. Softw. Syst. Model. 9 (Suppl. 1), 87–111.

Van Helden, J., Wernisch, L., Gilbert, D., Wodak, S.J., 2002. Graph-based analysis of metabolic networks. Bioinforma. Genome Anal. 14 (Suppl. 4), 245–274.

Velankar, S., et al., 2011. PDBe: Protein Data Bank in Europe. Nucl. Acids Res 39 (Suppl. 1), D402–D410.

Weckwerth, W., 2010. Metabolomics: an integral technique in systems biology. Bioanalysis 2 (Suppl. 4), 829–836.

Weijters, A., van der Aalst, W., van Dongen, B., Herbst, J., Maruster, L., Schimm, G., 2003. Workflow mining: a survey of issues and approaches. Data Knowl. Eng. b47 (Suppl. 2), 237–267.

Wheeler, D.L., et al., 2003. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 31, 28–33.

Wishart, D.S., 2010. Computational approaches to metabolomics. Methods Mol. Biol. 593, 283–313.

Witten, I.H., Frank, E., Hall, M.A., 2011. Data mining: practical machine learning tools and techniques. Morgan Kaufmann Series in Data Management Systems. 664.

Wolda, H., 1981. Similarity indices, sample size and diversity. Oecologia 50 (Suppl. 3), 296–302.

Xue, L., Godden, J.W., Stahura, F.L., Bajorath, J., 2001. Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. *ct.* J. Chem. Inf. Comput. Sci. 41 (Suppl. 2), 394–401.

Xue, L., Godden, J.W., Stahura, F.L., Bajorath, J., 2003. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *otect.* J. Chem. Inf. Comput. Sci. 43 (Suppl. 4), 1218–1225.

Yoshii, F., Yamada, Y., Hoshi, T., Hagiwara, H., 2002. The creation of a database of odorous compounds focused on molecular rigidity and analysis of the molecular features of the compounds in the database. Chem. Senses 27 (Suppl. 5), 399–405.

You, C.H., Holder, L.B., Cook, D.J., 2006. Application of graph-based data mining to metabolic pathways. 6th IEEE International Conference on Data Mining — Workshops (ICDMW'06), Hong Kong, China. 169–173.

Zaki, M.J., Hsiao, C.J., 2002. CHARM: an efficient algorithm for closed itemset mining. Proceedings of the Second SIAM International Conference on Data Mining, Chicago, U.S.A. , pp. 457–473.

Zhang, Y., et al., 2006. Phylophenetic properties of metabolic pathway topologies as revealed by global analysis. BMC Bioinforma. 7 (Suppl. 1), 252.

Zhao, J., Tao, L., Yu, H., Luo, J., Cao, Z., Li, Y., 2007. Bow-tie topological features of metabolic networks and the functional significance. Chin. Sci. Bull. 52 (Suppl. 8), 1036–1045.