Substructure Analysis of Metabolic Pathways by Graph-Based Relational Learning

Chang hun You, Lawrence B. Holder, and Diane J. Cook

School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164-2752
{changhun,holder,cook}@eecs.wsu.edu

Systems biology has become a major field of post-genomic bioinformatics research. A biological network containing various objects and their relationships is a fundamental way to represent a bio-system. A graph consisting of vertices and edges between these vertices is a natural data structure to represent biological networks. Substructure analysis of metabolic pathways by graph-based relational learning provides us biologically meaningful substructures for system-level understanding of organisms.

This chapter presents a graph representation of metabolic pathways to describe all features of metabolic pathways and describes the application of graph-based relational learning for structure analysis on metabolic pathways in both supervised and unsupervised scenarios. We show that the learned substructures can not only distinguish between two kinds of biological networks and generate hierarchical clusters for better understanding of them, but also have important biological meaning.

1 Introduction

A biological organism has one ultimate goal: to continue the life of its species and itself in the natural environment. This goal requires two important activities, maintaining low entropy in the environment and reproducing oneself. Biological organisms need to process various functions to maximize free energy and minimize entropy. These two basic processes are preformed by a variety of interactions in an organism.

Fundamentally the organism is a system itself as well as a property of a bio-ecosystem. A biological organism is not just composed of various objects, but also has dynamic and interactive relationships among them. A system-level understanding is a more efficient way to approach the organisms. With advances in computer science, bioinformatics plays a central role in life science. Traditional bioinformatics has been focused on molecular-level research. Genomics and proteomics, main areas in molecular-level research, have studied function and structure of macromolecules in organisms, and produced a huge amount of results. However almost every biomolecule plays its role only in

A.S. Sidhu et al. (Eds.): Biomedical Data and Applications, SCI 224, pp. 237–261.
springerlink.com

harmony with other components of the cytoplasmic environment. Molecular-level understanding is definitely a fundamental step, but it is not the final step. It is time to steer our steps to system-level approaches to bio-systems.

A biological network is a fundamental and indispensable way to describe a complex system in terms of both the structure and its dynamics. The final goal of systems biology is to model and simulate the biological networks for better understanding of bio-systems and contribution for drug discovery. An efficient way to model unrecognized biological networks is to discover patterns in existing biological networks. Biological networks contain various biomolecules and their relationships. The patterns of relationships in biological networks are crucial to understanding organisms and to modeling them at the system-level. Structure analysis of biological networks is a primary movement in systems biology.

Logic-based and graph-based approaches, as subfields of multi-relational data mining, are applied to mine patterns in biological networks. Logic-based data mining, also called inductive logic programming, represents networks using logic [22]. But this approach requires complicated syntax and the definition of prior knowledge. A graph has been widely used to represent a variety of relational data such as computer networks, social networks, and biological data. A biological network is another appropriate domain to be represented as a graph. Graph-based relational learning can be applied to find the meaningful patterns in the biological network that is represented as a graph.

In this paper, we review systems biology and some multi-relational data mining approaches applied to biological networks, and we describe the knowledge discovery approach used in the SUBDUE graph-based relational learning system. We then show the application of SUBDUE to metabolic pathways, which are downloaded from the KEGG PATHWAY database and represented as a graph. The goal of this research is to show that the learned substructures describe the system-level features in metabolic pathways and convey important biological meaning. Structure analysis on the same metabolic pathways from different species finds the substructure showing the unique features for specific species. Supervised learning shows that the learned substructures can identify what is unique about a specific type of pathway, which allows us to understand better how pathways differ. Unsupervised learning produces hierarchical clusters that describe what is common about a specific group of pathways, which provides us better understanding of common structure in pathways. Ultimately, these substructures can improve both our ability to understand recognized pathways and categorize unrecognized ones.

2 Systems Biology and Biological Networks

Systems biology is a novel stream of life science focusing on a comprehensive bio-system including integrated and interacting biological networks which are relations of genes, proteins and other biomolecules [15]. A system should be studied in a system-level manner including using comprehensive

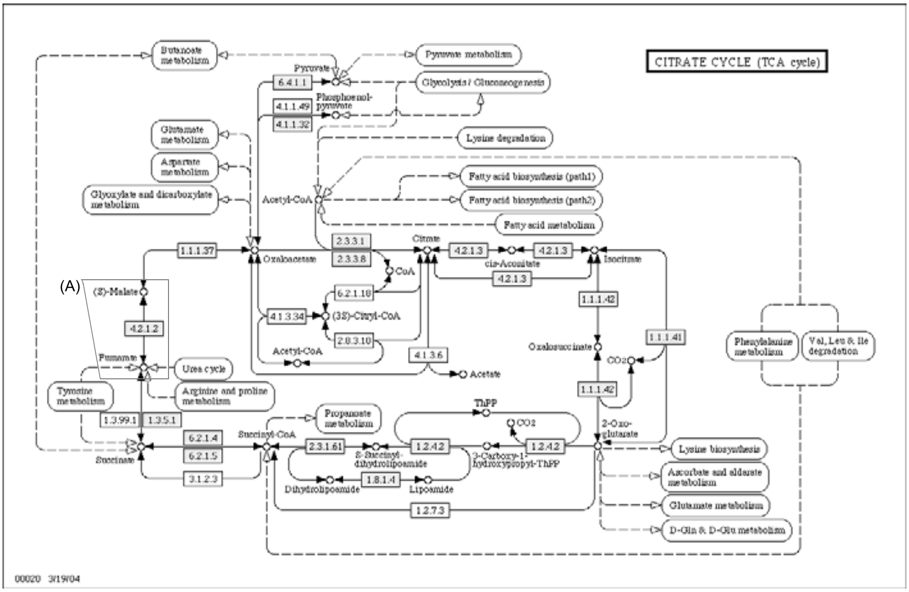

Fig. 1. TCA cycle biological network of *Homo Sapiens* [26]

methodologies, integrating heterogeneous data and understanding interactions with other data and various conditions. A system cannot be comprehended as a part but as a whole system.

Fundamentally an organism is a system itself as well as a property of a bio-ecosystem. The organism has a systematically well-organized structure consisting of multi-level compositions such as tissue, organ and organ system, all of which are based on a cell as a functional and structural basic unit. Each constituent is cooperating with others interactively as well as organized systematically. Even the basic unit, the cell, is also a system itself. A cell has a variety of biomolecules that are working with interactive relationships among them.

A huge amount of biological data has been generated by long-term research. Each result cannot allow us to understand a whole biological system, because any molecule or constituent in the organism never works alone. They always interact with others in the system. For this reason an organism should be explored as a system.

A biological system can be described as a large biological network, which consists of numerous small networks. Each network has various biomolecules and their relationships. Generally a cellular system is represented by three kinds of biological networks: metabolic pathways, protein-protein interactions and gene regulatory networks [16]. Our research is currently focused on the metabolic pathways.

Figure 1 shows a metabolic pathway called the TCA cycle which is a metabolic pathway for the production of ATP (a fundamental energy molecule in a cell). A rectangle represents an enzyme (protein) or a gene, and a circle represents a chemical compound. Each arrow describes a relationship between these molecules. In marked area (A), a compound, (S)-Malate (L-Malic acid), as a substrate is changed to another compound, Fumarate, as a product by an enzyme, ec:4.2.1.2. This is a basic biochemical reaction. The metabolic pathway is a complex network of biochemical reactions. A fundamental step to study metabolic pathways is the identification of structures covering a variety of biomolecules and their relationships. Dynamics and control methods of metabolic pathways are also included, because biological systems are interactive and well-controlled optimized systems [15]. Our current research is focused on identifying the structure. Our ultimate goal is to make a blueprint for system-level understanding and its application based on an understanding of the structure, dynamics and control of metabolic pathways.

The KEGG PATHWAY is a widely known database which contains information on various kinds of pathways including pathway image files (figure 1) [14]. The KEGG PATHWAY database has 84,441 pathways generated from 344 reference pathways (on December, 2008). It has six fundamental categories of pathways: Metabolism, Genetic information processing, Environmental information processing, Cellular processes and human diseases and Drug development . This database contains not only various information on pathways, but also plentiful information of their components as linked databases. It also has the KGML (KEGG Markup Language) as an exchange format for KEGG pathways, based on XML.

3 Related Work

Biological networks consist of various molecules and their relationships. Each molecule can have its own properties and can also influence relationships with other molecules. For this reason, traditional data mining, focusing only on the properties, is not applicable to biological networks. Multi-relational data mining is focused not only on properties of molecules but also their relationships [7]. To apply multi-relational data mining, it is necessary to represent the data along with its multiple relations. First-order logics and graph representations are used for representation of multi-relational data. These representation methods lead to two general approaches of multi-relational data mining: logic-based data mining and graph-based data mining.

In this section, we introduce several approaches to analyze biological networks. Then, we will introduce the multi-relational data mining approach as a method to analyze biological networks including inductive logic programming as a logic-based data mining method and two categories of graph-based data mining: graph-based relational learning and frequent subgraph mining

[10]. Graph-based relational learning, as our main focus, will be described in section 4.

3.1 Analysis of Biological Networks

Aittokallio and Schwikowski survey recent works of graph-based methods for analysis of biological networks [1]. They categorize graph-based approaches into three levels: global structural property, local structural connectivity and hierarchical functional description. They also introduced some recent approaches including integrating data from multiple source, graph-based software tools for network analysis and network reconstruction by reverse engineering.

Cheng et al. [3] show an approach of mining bridges and motifs from biological networks. They use a statistical method to detect structural bridges and motifs, and compare their bridges and motifs with functional units in the biological networks. They suggest the structures of the discovered bridges and motifs are significantly related to their function. Hwang et al. [13] introduce an approach of detecting novel functional patterns in protein-protein interaction networks using graph properties and signal transduction behavior. They model the dynamic relationships between two proteins using a signal transduction model to represent functional similarity between proteins. Huan et al. [12] try to discover spatial patterns in protein structures using the subgraph mining approach. They use Delaunay tessellation to represent three-dimensional structure of proteins, where Delaunay tessellation is defined based on the coordinates of molecules in the proteins. If two Voronoi cells share a common face, two points are connected by an edge. In this way, they represent the protein structure as a graph, and apply the frequent subgraph mining approach.

Mathematical modeling abstractly describes a system using mathematical formulae [20, 27]. Most of these approaches, as a type of quantitative analysis, model the kinetics of pathways and analyze the trends in the amounts of molecules and the flux of biochemical reactions. But most of them disregard relations among multiple molecules.

Our research applies graph-based data mining to learn the patterns in the dynamics of biological networks [29, 30]. We introduce a dynamic graph containing a sequence of graphs that represent the time-based versions of biological networks to represent the structural changes of biological networks. Our approach first discovers graph rewriting rules between two sequential graphs to represent how one network is changed to another, and then discover more general transformation rules in the set of the graph rewriting rules to describe how biological networks change over time.

There are many computational approaches to analyze biological networks. Biosystems are organized as networks and perform their function in relations to molecules and systems. For this reason, we need to focus on the structural properties of biological networks rather than the molecules themselves.

3.2 Logic-Based Data Mining

Logic-based data mining is a multi-relational data mining technique using first order logic to represent data. Inductive Logic Programming (ILP), a typical logic-based data mining technique, is an approach to induce hypotheses (rules, concepts, or knowledge) from observations (examples, instances, or experiences) by using logic to represent hypotheses and observations [7]. ILP has been applied in several ways to biological domains such as genomics and proteomics. ILP in company with other approaches has been applied to metabolic pathways.

Support Vector Inductive Logic Programming (SVILP) is the intersection approach between Support Vector Machines (SVM) and ILP. By using logic to represent background knowledge, the SVILP technique has been applied to the prediction of toxicity of given materials [23]. Stochastic methods are also applied to logic programming to model metabolic pathways. This approach models rates of biochemical reactions using probabilities in addition to representation of metabolic pathways by logic [21]. A last approach is a cooperation between induction and abduction. This approach generates hypotheses from not only abductive reasoning from experimental data (concentrations of metabolites), but also inductive reasoning from general rules (from the KEGG database) to model inhibition in metabolic pathways [25]. An inhibitor is a chemical compound to control biochemical reactions.

Several ILP-related techniques have been successfully applied to metabolic pathways. Logic programming can efficiently represent relational data, but prior rules and examples may be necessary to represent entire pathways.

3.3 Frequent Subgraph Mining

The graph is an abstract data structure consisting of vertices and edges which are relationships between vertices. Graph-based data mining denotes a collection of algorithms for mining the relational aspects of data represented as a graph. Graph-based data mining has two major approaches: frequent subgraph mining and graph-based relational learning. Frequent subgraph mining is focused on finding frequent subgraphs in a graph. There are two well-known approaches to bioinformatics domains. Frequent SubGraph discovery, FSG, finds all connected subgraphs that appear frequently in a set of graphs. FSG starts by finding all frequent single and double edge graphs. During each iteration FSG expands the size of frequent subgraphs by adding one edge to generate candidate subgraphs. Then, it evaluates and prunes discovered subgraphs with user-defined constraints [19].

Graph-based Substructure Pattern Mining, gSpan, uses the depth-first search and lexicographic ordering. First gSpan sorts the labels, removes infrequent vertices and edges, and it relabels the remaining vertices and edges. Next it starts to find all frequent one-edge subgraphs. The labels on these edges and vertices define a code for each graph. Larger subgraphs map themselves to longer codes. If the code of B is longer than A, the B code is a child of the A code in a code tree. If there are two not-unique codes in the tree, one of them is removed during the depth-first search traversal to reduce the cost of matching frequent subgraphs [28]. There are a few approaches of frequent subgraph mining applied to metabolic pathways. Pathway Miner, a simplified graph-mining approach on metabolic pathways, proposes a simplified graph representation consisting of just enzyme relationships. In this way, the approach may avoid the NP-hard subgraph isomorphism problem and find frequent patterns quickly [17]. However, the over-simplified representation makes this approach focus on just enzyme relationships, missing some important features of metabolic pathways.

Mining coherent dense subgraphs uses the correlation of graphs which represent gene regulatory networks [11]. This approach compresses a group of graphs into two meta-graphs using correlated occurrences of edges for efficient clustering. This approach also loses some biological characteristics of gene regulatory networks, because its representation of a graph is derived only from the similarity of the gene expression patterns between two genes, not representing how the practical biological interactions exist.

4 Graph-Based Relational Learning

Graph-based relational learning is focused on novel and meaningful, but not necessarily most frequent, substructures in a graph representation of data. We use the SUBDUE graph-based relational learning approach to discover patterns which not only abstract instances of the patterns by compression, but also provide better understanding of the data [5]. SUBDUE can perform unsupervised learning and supervised learning by substructure discovery based on Minimum Description Length (MDL). Using background knowledge given as predefined substructures can guide graph-based relational learning to find more meaningful substructures. SUBDUE has been applied to a variety of areas such as Chemical Toxicity [4], Molecular Biology [24], Security [9] and Web Search [6].

SUBDUE accepts input data which is represented as a graph with labeled vertices and labeled, directed or undirected edges between vertices. The objects and attribute values of the data are usually represented as vertices, while attributes and relationships between objects are represented as edges. Figure 3 shows a graph representation of a KEGG biological network. There are five 'Entry' vertices which represents Enzyme or Compound. Each Entry has two attributes: name and type. Relationships are given as directed and labeled edges from Entry to its attributes. More detail on this graph representation will be provided later.

4.1 Discovery Algorithm

The SUBDUE discovery algorithm is based on a beam search as shown in figure 2. The algorithm starts with three parameters: input graph, beam length

```
Subdue(graph $G$, beam $B$, limit $L$)
    $Q = \{v \mid v$ is a vertex in G with a unique label$\}$
    $bestSub$ = first substructure in $Q$
    repeat
        foreach substructure $S \in Q$ do
            add Extend($S$) into $extSubs$
            foreach $newSub \in extSubs$ do
                Evaluate($newSub$)
                add $newSub$ in $Q'$ (length $< B$)
        if best Sub $\in Q'$ better than $bestSub$ then
            $bestSub$ = best Sub $\in Q'$
        $Q = Q'$
    until $Q$ is empty or $Num.OfSubs.Extended > L$
    return $bestSub$

Extend($S$): extend Sub. S by one edge in all possible ways
Evaluate($S$): evaluate Sub. S using MDL
```

Fig. 2. SUBDUE's discovery algorithm

and limit value. The beam length limits the length of the queue which contains extended substructures, and the limit value restricts the total number of substructures considered by the algorithm. The initial state of the search is the set of substructures representing each uniquely labeled vertex and their instances. The *Extend*(S) function extends each instance of a substructure in the Q in all possible ways by adding a single edge and a vertex, or by adding a single edge if both vertices are already in the substructure. The substructures in the Q' are ordered base on their ability to compress the input graph as evaluated by *Evaluate*(S), using the minimum description length (MDL) principle. This search (repeat loop) terminates when the number of substructures considered reaches the limit value, or the algorithm exhausts the search space. Then it returns the best substructure.

SUBDUE can be given background knowledge in the form of predefined substructures. SUBDUE finds the instances of these substructures and compresses them. Using this approach, we can verify whether the patterns learned from a graph belong to another graph.

4.2 MDL and Heuristic Methods

The discovery algorithm of SUBDUE fundamentally is guided by the minimum description length principle. The heuristic evaluation by the MDL principle assumes that the best substructure is the one that minimizes the description length of the input graph when compressed by the substructure [5]. The description length of the substructure S is represented by $DL(S)$, the description length of the input graph is $DL(G)$, and the description length of the input graph after compression is $DL(G|S)$. SUBDUE's discovery algorithm tries to minimize $DL(S) + DL(G|S)$ which represents the description length of the graph G given the substructure S. The compression of the graph can be calculated as

$$Compression = \frac{DL(S) + DL(G|S)}{DL(G)}$$

where description length $DL()$ is calculated as the number of bits in a minimal encoding of the graph. Cook and Holder describe the detailed computation of $DL(G)$ in [5].

The discovery algorithm of SUBDUE is computationally expensive as other graph-related algorithms. SUBDUE uses two heuristic constraints to maintain polynomial running time: *Beam* and *Limit*. Beam constrains the number of substructures by limiting the length of the Q' in figure 2. Limit is a user-defined bound on the number of substructures considered by the algorithm.

4.3 Unsupervised Learning

Once the best structure is discovered, the graph can be compressed using the best substructure. The compression procedure replaces all instances of the best substructure in the input graph with a pointer, a single vertex, to the discovered best substructure. The discovery algorithm can be repeated on this compressed graph for multiple iterations until the graph cannot be compressed any more or on reaching the user-defined number of iterations. Each iteration generates a node in a hierarchical, conceptual clustering of the input data. On the *ith* iteration, the best substructure S_i is used to compress the input graph, introducing a new vertex labeled S_i to the next iteration. Consequently, any subsequently discovered subgraph S_j can be defined in terms of one or more S_i, where $i < j$. The result is a lattice, where each cluster can be defined in terms of more than one parent subgraph.

4.4 Supervised Learning

The SUBDUE discovery algorithm has been extended to perform graph-based relational concept learning, or supervised learning [8]. The main approach of supervised learning is to find a substructure that appears often in the positive examples, but not in the negative examples. The substructure value is increased when positive examples are covered by the substructure, but is decreased where negative examples are covered. Positive examples not covered by the substructure and negative examples covered by the substructure are considered errors. The substructure value is calculated by

$$value = 1 - error$$

where the error is calculated by

$$error = \frac{\#PosEgsNotCvd + \#NegEgsCvd}{\#PosEgs + \#NegEgs}$$

$\#PosEgsNotCvd$ is the number of positive examples not containing the substructure, and $\#NegEgsCvd$ is the number of negative examples containing the substructure. $\#PosEgs$ is the number of positive examples remaining in the experimental set, of which the positive examples that have already been covered in a previous iteration were removed, and $\#NegEgs$ is the total number of negative examples, which is constant, because negative examples are not removed.

SUBDUE's supervised learning uses two approaches to minimize error. First, by using the definition of description length SUBDUE tries to find a substructure S minimizing $DL(G^{+}|S) + DL(S) + DL(G^{-}) - DL(G^{-}|S)$, where the last two terms represent the incorrectly compressed negative example graph. This approach will lead the discovery algorithm toward a larger substructure that characterizes the positive examples, but not the negative examples.

In addition to the compression-based evaluation, SUBDUE can use a set-cover approach based on the error measure. At each iteration SUBDUE adds a new substructure to the disjunctive hypothesis and removes covered positive examples. This process continues until either all positive examples are covered or no substructure exists discriminating the remain positive examples from the negative examples.

5 Substructure Analysis in Metabolic Pathways

Our goal is the application of the SUBDUE graph-based relational learning system to the KEGG metabolic pathways to find better understanding and biologically meaningful substructures. These substructures can distinguish two pathways or provide the common features in several pathways. Research shows that topological features of biological networks are closely related to biological functions [13, 2].

A simple way to apply supervised or unsupervised learning to the pathways is based on molecules, such as genes, proteins and other macro molecules. Because each molecule has a specific structure and other biochemical features, we can easily distinguish two groups or find the common features in a group. But our research is focused on the pattern of the relationship between molecules for system-level understanding of pathways. The pattern of relationship can be shown in a variety of forms, such as biochemical reaction, enzyme activity and signal transduction.

This section first introduces our graph representation (section 5.1). As a preliminary task, we describe substructure analysis on individual metabolic pathways (section 5.2). Then we represent our main experiments in this research: supervised learning (section 5.3) and unsupervised learning (section 5.4) on groups of metabolic pathways. The ultimate goal of our exploration is to show that the substructures found by graph-based relational learning are biologically important and meaningful.

5.1 Graph Representation

Input graphs for SUBDUE are converted from KGML files. KGML is a standard data format to express and distribute a biological network from KEGG. There are three major entities in KGML: Entry, Relation and Reaction. Entry represents various biomolecules in the metabolic pathway, such as enzyme, gene, compound and so on. Relation denotes a relationship between two or more enzymes, genes and maps. The maps denote the types of the Entry nodes linked to the other pathways [26]. The names of these Entry nodes represent the name of the linked pathways. Reaction is a biochemical reaction between two or more compounds catalyzed by one or more enzymes. Detailed information on KGML is described in [26]. In biochemical semantics, Entries are nodes of metabolic pathways, and Relations and Reactions are relationships between two or more Entries.

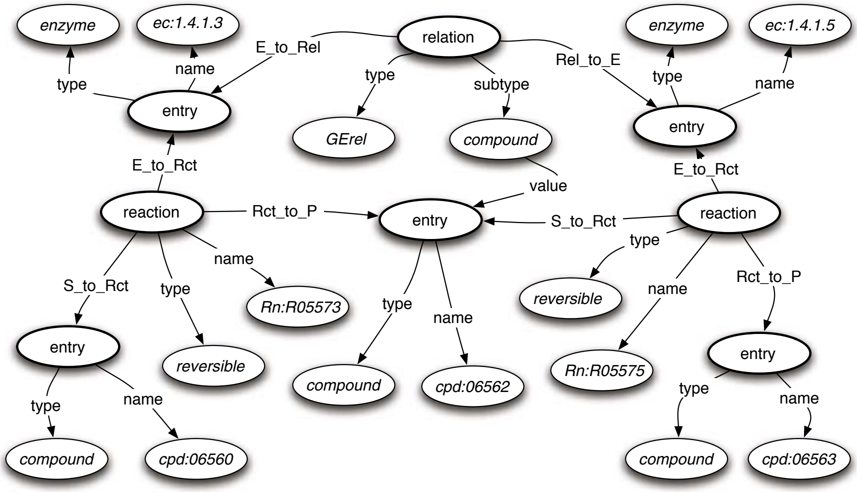

Fig. 3. A graph representation of a metabolic pathway

In our graph representation, Relations and Reactions are also represented as vertices in order to describe the properties of Relations and Reactions. Vertices representing major entities have two satellite vertices which are connected to their main vertex by edges, labeled as Name and Type, to explain its property. A name vertex linked by the Name edge denotes the KEGG ID, and a type vertex linked by the Type edge describes the property of the entity vertex. A Relation represents the association between two or more Entries (genes or enzymes) by an edge whose label represents a direction from one

Entry to another. Reaction also has connections with two compounds (described as a substrate and a product) and an enzyme (as a catalyst). Figure 3 shows an example of our graph representation, which has five Entries: three compounds and two enzymes. There is also a Relation between two enzymes, and two Reactions sharing a compound and having relationships with two other compounds.

Our research uses two kinds of graph representation: named and unnamed graph. The graph in figure 3 is the named graph, which includes the KEGG IDs. The unnamed graph is the same graph only excluding any unique IDs of each Entry and Reaction; all vertices and edges regarding "name" are removed from the graph in figure 3.

Unique IDs can pose potential problems when SUBDUE searches for substructures that distinguish two groups of graphs, which contain a group of pathways as positive examples and another group of pathways as negative examples. For example, when we try to distinguish two metabolic pathways G_1 and G_2, an enzyme which exists only in G_1, but not in G_2, is sufficient to distinguish. Because our research is focused on the pattern of metabolic pathways, a specific name is not useful for finding the pattern. For this reason, our supervised (section 5.3) and unsupervised (section 5.4) learning on groups of pathways use the unnamed graphs (the first phase). The named graphs are used at the second phase to verify the biological meaning of the patterns. The following sections describe this process in more detail.

5.2 Substructure in Metabolic Pathway

This section shows a brief example substructure analysis on metabolic pathways. Here we have two metabolic pathways. SUBDUE tries to find the substructures that exist in one pathway, but not in another. In this experiment, we show that the species-specificity is related to the structure of metabolic pathways. Species-specificity is one of the most important concepts in biology. Basically species-specificity is derived from protein structure and gene sequence.

We arrange two glycolysis pathways from Human and *E.coli*: hsa00010 and eco00010. Glycolysis is a representative energy generating process in almost every cell. We seek a slightly structural difference between these pathways from two species. A named graph of hsa00010 has 1047 vertices and 1208 edges, and the one of eco00010 has 1002 vertices and 1132 edges. The former is a positive example and the latter is a negative example for supervised learning of SUBDUE. Then we run SUBUDE to find substructures that exist in hsa00010, but not in eco00010.

As a result, the best substructure is found as shown in figure 4. While this substructure is discovered in eco00010 as well as hsa00010, it is still possible to show how the two pathways differ. The instances of the best substructure in figure 4 are found 10 times in hsa00010 and 6 times in eco00010. After

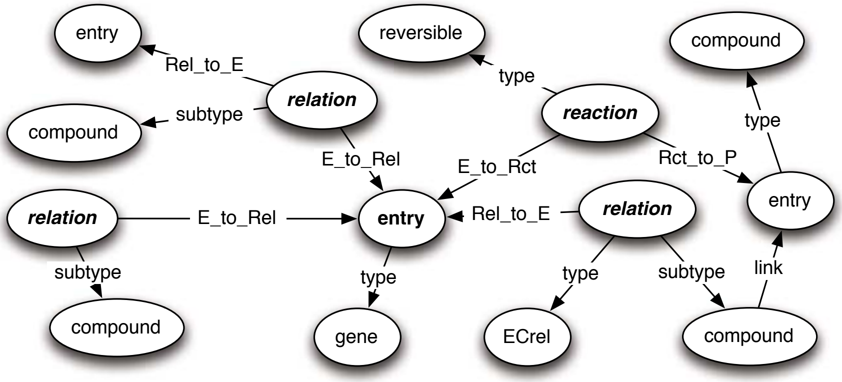

Fig. 4. An instance of the best substructure found in hsa00010

inspecting locations of these instances on each pathway, we can identify structural differences between two pathways.

Five cases out of 10 instances in hsa00010 are also found in eco00010. The other five instances are found only in hsa00010. These five instances show unique parts exist only in hsa00010. Figure 5 shows an example of one of these five instances. Figure 5 (A) and (B) show the part of the glycolysis pathways in human and *E.coli* respectively. Connected circled entities denote the instance of the best substructure. The marked instance includes a reaction (R00710) catalyzed by an enzyme (ec:1.2.1.5) and three relations related with the enzyme. The reaction R00710 exists only in hsa00010, but not in eco00010. We can confirm this concept in figure 5. Rectangles and circles denote enzymes and compounds. There is no symbol of relation in this figure. Grayed rectangles represent existing enzymes in the species and white rectangles denote only conceptual views (practically do not exist). As shown in this figure eco00010 does not include ec:1.2.1.5 and ec.1.1.1.2, which exist in hsa00010. The process of Acetaldehyde (NAD+ oxidoreductase) does not exist in eco00010. SUBDUE discovers this substructure existing in hsa00010, but not in eco00010. In this way structure analysis by SUBDUE can characterize the unique features of metabolic pathways.

We execute the same method on other experiment sets: 00020 (TCA cycle), 00051 (Fructose and mannose metabolism), 00061 (Fatty acid biosynthesis) and 00272 (Cysteine metabolism) from human and *E.coli*. The results are shown in table 1. The first column shows the name of the pathway. The second column shows the number of instances of the best substructure found in hsa pathways, and the third column shows the number of instances of the best substructure found in eco pathways.

The best substructure of the 00020 experiment is found in both pathways as in the 00010 experiment. But in the other three experiments the best substructures are found only in the hsa pathway. The best substructure existing only in one pathway precisely shows the unique features of the pathway. In

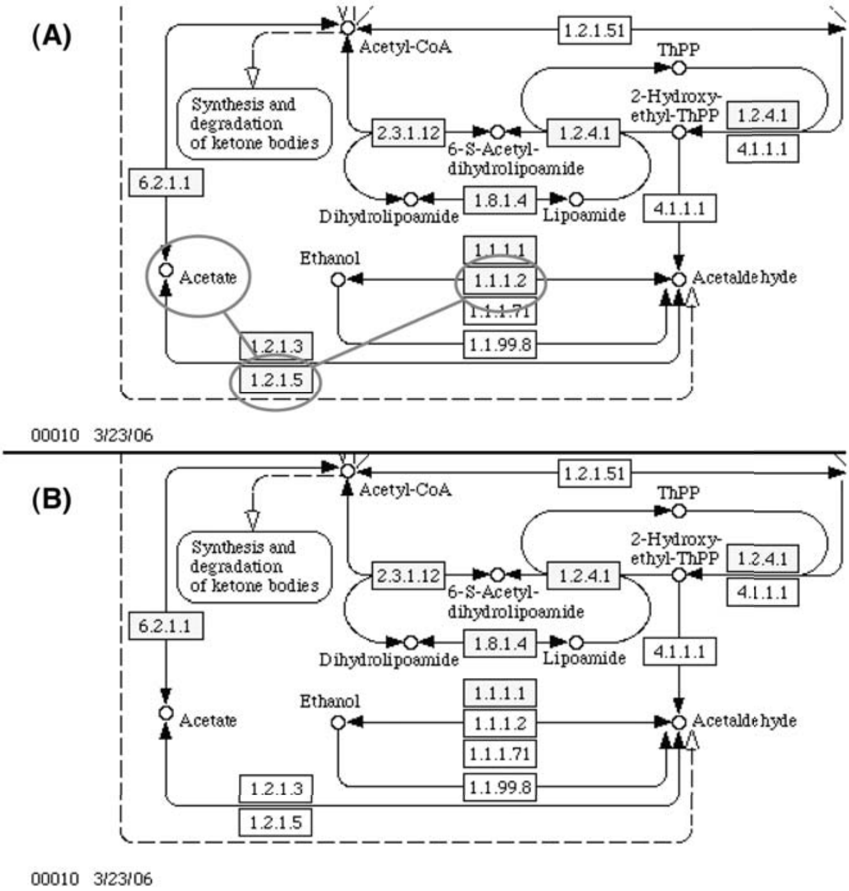

Fig. 5. Part of Glycolysis pathways in human (A) and *E.Coli* (B) [26]. Connected three circled entities represents an instance of the best substructure in figure 4.

Table 1. Results of structure analysis on pathways

Pathway	Number of instance in hsa	Number of instances in eco
00010	10	6
00020	66	47
00051	6	0
00061	171	0
00272	3	0

the case of the best substructure existing in both pathways, the number or the location of the instances can also indicate distinguishing features of the pathways: how many instances are in the pathway or what process is included into the pathway. In this way substructure analysis allows us to reveal the system-level features of metabolic pathways.

5.3 Supervised Learning in Metabolic Pathways

The main goal of supervised learning is to distinguish between a biological network in a species group and a different network in the same group. This task provides us the unique substructures in the specific group of pathways to understand better how pathways differ. The distinguishing pattern of relationships between molecules, with limited consideration of the features of each molecule, can also play an important role in system-level understanding of organisms.

As described previously supervised learning uses the unnamed graphs for substructure discovery (the first phase) and the named graphs for verification of the substructure (the second phase). Graphs of metabolic pathways are divided into two groups: positive and negative examples. SUBDUE searches for patterns which exist in positive examples, but not in negative examples. We then use SUBDUE to find the erased unique IDs or unfound (in the first phase) vertices and edges in a group of named graphs using the best substructure from the first phase as predefined substructures. The second phase takes aim at verifying the biological meaning of the discovered substructures. Linked databases of the KEGG PATHWAY are also used to identify biological meaning of the final substructures.

The discovery algorithm uses the set-cover approach and it is iterated for the number of positive examples. We use the heuristic constraint for polynomial running time as described in section 4.2. Our heuristic Limit, L, is calculated by

$$L = V + B(E\gamma - 1) \tag{1}$$

where V is the number of initial vertices, B is Beam length, E is the number of unique labeled edges, and γ is a heuristic constant. V and E are determined from the input graph (positive examples). B is set as 4 because this value is generally used in the successful application of SUBDUE to various domains. When we assume that the substructure learned by SUBDUE has the same number of edges as the number of unique labeled edges in the metabolic pathway, γ is 1.0. We try different γ values, and determine which value gives the best substructure in the shortest time. After several experiments, we found 1.5 supervised learning) and 1.0 (unsupervised learning) as the best choice for γ in this domain.

Results of supervised learning

Supervised learning in SUBDUE tries to find the substructures that exist in the positive examples, but not in the negative examples. The choice of which set of pathways is the positive example might affect the classification result. Since our goal is the better classification between two groups, we run two cases. First (A+), we make a positive example set and a negative example set. The second (B+) is vice versa. We present the classification accuracy and running time for both in each experimental set. Assuming that we run two cases in parallel, the maximum accuracy expresses the best case of the classification and the maximum running time represents the worst case of the running time.

Table 2. Results of supervised learning

Set	**Ex.**	**Size**	**Time**	**Acc.**	**Time**	**Acc.**	**Time**	**Acc.**
(A_B:src)	(A / B)	(V+E)	A+(s)	A+(%)	B+(s)	B+(%)	Max(s)	Max(%)
00300_00310:euk	9/16	14,715	1.26	44.00	1.17	64.00	1.26	64.00
00520_00530:euk	14/17	15,689	2.19	83.87	1.44	67.74	2.19	83.87
00010_00900:euk	17/16	38,804	79.00	100.00	8.84	100.00	79.00	100.00
00010_00061:euk	17/15	56,914	27.72	100.00	54.12	100.00	54.12	100.00
00230_00240:euk	17/17	75,086	49.38	100.00	111.31	55.88	111.31	100.00
00010_00230:euk	17/17	75,786	57.62	100.00	50.79	94.12	57.62	100.00
00300_00310:45	33/42	41,569	11.22	44.00	18.50	56.00	18.50	56.00
00520_00530:45	39/40	42,092	17.48	64.56	18.71	54.43	18.71	64.56
00010_00510:45	44/31	82,767	129.71	100.00	337.25	44.00	337.25	100.00
00010_00900:45	44/41	88,041	109.42	100.00	130.19	100.00	130.19	100.00
00010_00020:45	44/39	110,701	302.62	63.86	876.96	50.60	876.96	63.86
00251_00252:45	45/45	116,621	354.47	61.11	226.82	53.33	354.47	61.11
00010_00061:45	44/39	117,582	247.18	100.00	305.51	46.99	305.51	100.00
00010_00251:45	44/45	129,187	410.27	94.38	503.64	61.80	503.64	94.38
00010_00230:45	44/45	179,393	1322.95	76.40	650.40	91.01	1322.95	91.01
00230_00240:45	45/45	183,701	368.12	100.00	2349.60	60.00	2349.60	100.00
00520_00530:150	137/136	150,363	874.79	53.85	1236.14	53.41	1236.14	53.85
00300_00310:150	136/143	157,267	441.32	48.75	587.7	53.41	587.7	53.41
00010_00900:150	149/143	286,091	1610.45	95.21	1117.41	100.00	1610.45	100.00
00010_00061:150	149/140	371,032	3107.66	100.00	4013.80	48.44	4013.80	100.00

Table 2 shows the experimental sets and the results for supervised learning. The first column shows the name of the set which consists of three parts: A, B and source group. A and B represent two groups of pathways [26], and the source group represents the species set. The Eukaryote set consists of all eukaryote species (17) in the KEGG PATHWAY database. The 45 set has 45 species, and the 150 set has 150 species. The second column provides the number of pathways in each group. This number is less than or equal to the number of each source set, since the metabolic pathway may not yet be constructed (or not presented) in the specific species. For example all 17 species of the eukaryote cell have the 00010 network. But, *Encephalitozoon cuniculi* (fungi) and *Danio rerio* (Zebra fish) do not have the 00061 network. The third column shows the total size of the graphs, which is calculated as $size(G) = |V| + |E|$, where a graph $G = (V, E), V$ is the set of vertices and E is the set of edges. The 4th and the 6th columns show the running

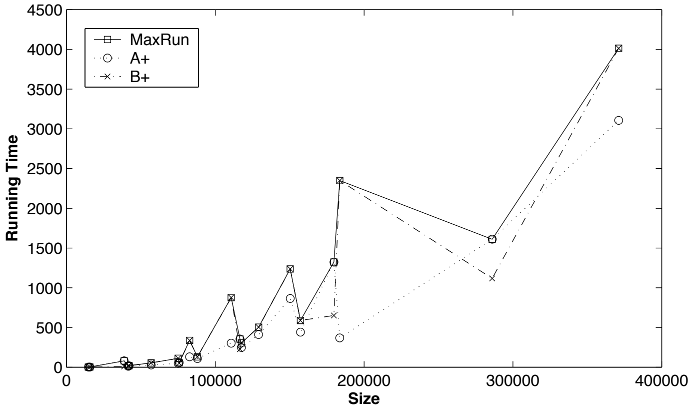

Fig. 6. Running time with graph size in supervised learning

time (seconds), and the 5th and the 7th columns show the accuracy. Accuracy is calculated as $(TP + TN)/(|A| + |B|)$, where TP is the number of the positive examples containing at least one of the best patterns from any iteration, and TN is the number of the negative examples containing none of the best patterns from any iteration. The 4th and the 5th columns (A+) represent experiments that have A as positive and B as negative examples. The 6th and the 7th columns (B+) represent the reverse experiments. The last two columns show the maximum running time and accuracy in each set. We use a convention A(+)_B(-):src when we denote each experimental case. For example, 00300(+)_00310(-):euk represent A+ case of 00300_00310:euk experiment.

Supervised learning uses 1.5 as the γ constant in the heuristic equation (1). Higher γ may sometimes afford better substructures which show more accuracy. But computational expense is usually not worth the small increase in accuracy. We make a compromise on this γ constant between running time and accuracy of classification. Each set has 11 $\sim$ 13 initial unique vertices and 8 $\sim$ 11 unique edges, so Limit, L, can be calculated as 55 $\sim$ 73.

Each case shows the different result in terms of running time and accuracy, dependent on what is used as the positive examples. The average accuracy of all experiments consisting of A+ and B+ is 71.76%. The average of maximum accuracy is 82.3%. There are discrepant cases between $A+$ and $B+$. 7 sets out of 20 show that one case is better than the average and another is worse. For instance, the 00010_00510:45 set has 100.00% accuracy in $A+$ case and 44.00% in $B+$. However, SUBDUE finds the substructures well to distinguish between two groups of pathways with more than 60% accuracy (17 sets out of 20). Figure 6 shows the running time with the graph size: $A+, B+$, and MaxRun (the maximum running time). SUBDUE's running time increased polynomially with the size of the graph.

Verification of the substructures

The goal of supervised learning is to find the patterns which are not only able to distinguish between two sets of examples, but are also biologically meaningful. The pattern found by SUBDUE can differentiate well between two examples. We verify biological meaning of these patterns by using the linked database of KEGG PATHWAY [26]. Figure 7 shows the best substructure, which is found in 40 instances of 40 examples in the first iteration of the 00010(+)_00900(-):45 experiment. This substructure which covers 90.9% of the positive examples (40 out of 44) is related to two reactions. Because the edge E_to_Rct represents a relationship between reaction and enzyme (gene), the entry should be the enzyme or the gene.

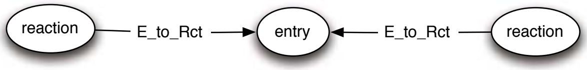

Fig. 7. First best pattern from supervised learning on 00010(+)_00900(-):45 set

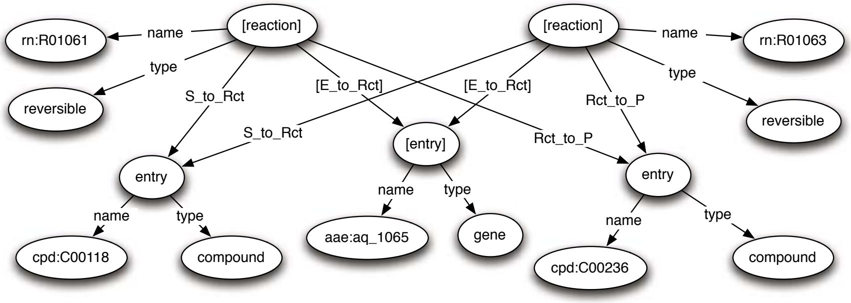

Fig. 8. Updated best pattern from supervised learning on 00010(+)_00900(-):45 set

In the second phase (verification phase) SUBDUE runs on the named graph of the same example set 00010(+)_00900(-):45 with the first best pattern (figure 7) as the predefined substructure. SUBDUE can find clearly all forty instances in the named graph. The second phase adds more vertices and edges which are erased in the unnamed graph or are not found at the first phase. The substructure in figure 8, which is the updated pattern from the result of the first phase, is the final result of this experiment. The vertices and edges marked by "[ ]" are included from the original substructure learned in the first phase. With this completed substructure, we can refer to linked databases in the KEGG PATHWAY database.

SUBDUE supervised learning finds the substructure representing that an enzyme catalyzing two reactions, which share the same substrate and product. Generally an enzyme catalyzes a reaction, but some enzymes can be related to two or more reactions. Figure 8 shows two reaction vertices are connected to an entry (enzyme) vertex by two E_to_Rct edges, which denote links between an enzymes and a reaction. The two reactions include a shared substrate (linked by a S_to_Rct edge) and product (linked by a Rct_to_P edge). The S_to_Rct edge denotes a link from a substrate to a reaction, and the Rct_to_P edge represents a link from a reaction to a product. SUBDUE finds that this substructure exists only in 00010 examples, not in 00090 examples.

In this substructure aae:aq_1065 which is the *gene* name, represents the *enzyme* ec:1.2.1.12 (glyceraldehyde-3-phosphate dehydrogenase). This enzyme catalyzes two reactions, R01061 and R01063, which are oxidoreductase reactions of NAD^+ and $NADP^+$ [26]. NAD^+ and $NADP^+$ are coenzymes that function as carriers of hydrogen atoms and electrons in some oxidation-reduction reactions, especially ATP (Adenosine TriPhosphate: energy material) related reactions. In our experiment the learned substructure is found only in the positive examples (*Glycolysis*), not in the negative examples (*Terpenoid biosynthesis*). Glycolysis is an energy generating process which degrades a molecule of glucose in a series of enzyme-catalyzed reactions to yield two molecules of the Pyruvates and ATPs. The conclusion of verification shows that the substructure found by SUBDUE can distinguish between two metabolic pathways and has an understandable biological meaning.

Two different metabolic pathways have unique relations as well as unique biochemical molecules. This research is focused on the unique relations. In case of 00010(+)_00900(-):45, an enzyme has relations with two reactions at the same time. The enzyme has an important feature called *substrate specificity*, which indicates that an enzyme can be active only when binding with a specific compound. For this reason, the enzyme, ec:1.2.1.12, catalyzes two reactions which have a common relation with the compound, cpd:C00118. In addition that identification of the unique biomolecules in each biological network is a fundamental step, but discovery of the unique relations is also important to classify metabolic pathways. The pattern of relations in the metabolic pathway can be a guide to model an unrecognized metabolic pathway.

5.4 Unsupervised Learning in Metabolic Pathways

Unsupervised learning tries to find common substructures in a set of different pathways of one species. The ultimate purpose of applying unsupervised learning to metabolic pathways is to provide a better understandable blueprint of metabolic pathways by using hierarchical topologies. This experiment allows us to understand what common structures the different networks have. The common patterns of relationships in metabolic pathways can contribute to biological network research accompanied with traditional bioinformatics.

Like supervised learning, unsupervised learning also employs the unnamed graphs in the first phase and the named graphs in the second phase. In the first phase SUBDUE discovers the substructures and generates hierarchical

Table 3. Results of unsupervised learning

Set	**Number of examples**	**Size**	**Running time**
(Species)	(Number)	(V+E)	(sec.)
ath	100	68,585	267.81
dme	92	58,166	204.52
eco	102	78,252	418.42
rno	96	61,409	183.60
sce	86	63,078	249.69
mmu	106	81,634	556.81
hsa	110	90,157	598.99

clusters using iterative discovery. Then SUBDUE adds eliminated unique IDs or unfound vertices and edges from the first phase. This process uses the substructures discovered from the first phase as predefined substructures in the second phase. The second phase also tries to verify the biological meaning of the discovered substructures by referring to the linked databases of the KEGG PATHWAY, which are used to identify biological meaning of the final substructures. The same heuristic equation (1) is used to compute the limit, L, as in supervised learning.

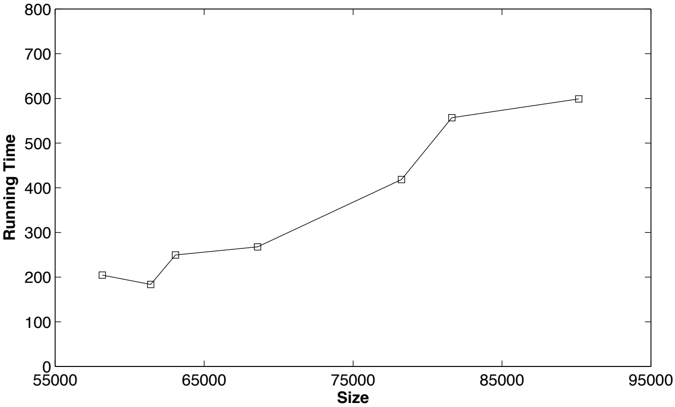

Fig. 9. Running time with graph size

Results of unsupervised learning

Table 3 shows the experimental sets used in unsupervised learning. Set represents the name of the species [26]. The number of examples denotes the number of metabolic pathways which the species has in the KEGG PATHWAY database. The 110 metabolic pathways in hsa (*Homo Sapiens*) is the largest number in the KEGG, when we include just metabolic pathways, not regulatory networks. Other species have fewer metabolic pathways, because they do not exist or are yet to beconstructed. Size is the size of the graph as described above. The last column shows the running time. Each run iterates 10 times to construct hierarchical clusters. Unlike supervised learning, this experiment uses MDL as the evaluation method. Unsupervised learning uses 1.0 as the γ constant in the heuristic equation (1). Each set has 14 $\sim$ 16 initial unique labeled vertices and 8 $\sim$ 11 unique labeled edges. Limit, L, is calculated as 42 $\sim$ 54. SUBDUE runs in polynomial time with the size of the graph as shown in figure 9.

Verification of the substructures

The purpose of SUBDUE unsupervised learning is to find the common substructures, which describe the regular features in a group of metabolic pathways. Moreover, hierarchical clusters of the common substructures show a blueprint of metabolic pathways. We provide hierarchical clusters learned by SUBDUE and verify them using the linked databases of KEGG PATHWAY. Partial hierarchical clusters of substructures learned from the dme (fruit fly) set are shown in figure 10. SUB_*i* denotes the best substructure in the *i*-th iteration.

The hierarchical clusters show that the substructures in the upper level are contained in lower level. For example, SUB_8 includes two SUB_1, one SUB_3 and one SUB_4. The general substructures are used to compose more specific substructures. This is how SUBDUE shows the common relational patterns of the metabolic pathways and how the patterns relate to each other hierarchically. SUB_8 is discovered in three metabolic pathways of fruit fly (dme). This substructure is not only the common feature in these three pathways, but also the distinct property from other pathways.

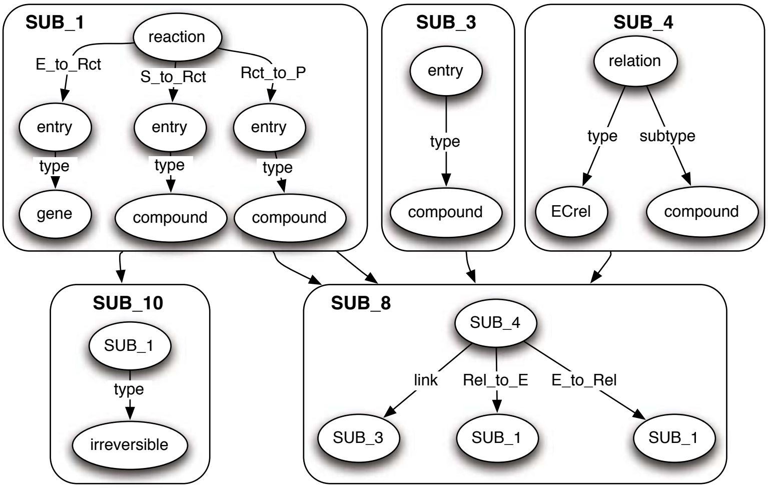

Fig. 10. Partial Hierarchical Clusters of metabolic pathways in fruit fly

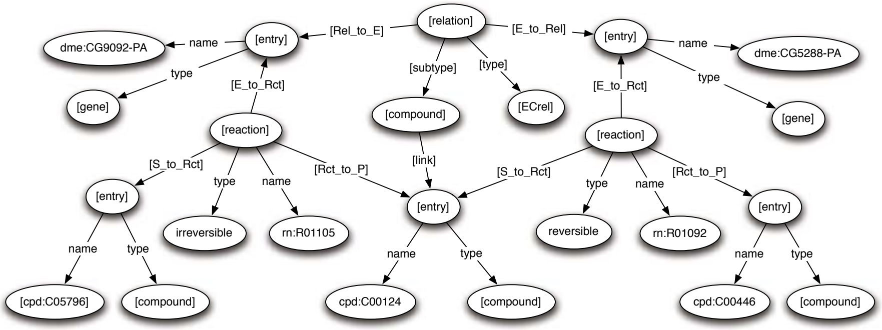

Fig. 11. Updated eighth substructure from figure 10

SUB_1 shows a basic reaction that is found in 972 instances of 90 examples. SUB_3 is found in 3,659 instances of 47 examples at the third iteration. SUB_4, found in 1,136 instances of 21 examples, represents a relation with the ECrel property. The ECrel relation is an enzyme-enzyme relation where two enzymes catalyze successive reaction steps [26]. SUB_8 discovered in 264 instances of 3 examples includes one relation of two enzymes which catalyze two successive reactions. Moreover, SUB_8 has an additional meaning to SUB_4 such that the "link" edge connects to a compound which is a product of the first reaction of this relation and a substrate of the second reaction at the same time [26].

SUB_8 includes an enzyme-enzyme relation which relates three consecutive chemical compounds. Figure 11 shows an example of SUB_8 which is found in the dme00052, Galactose metabolic pathway of the fruit fly. Figure 11 is a fully updated substructure thorough the second phase. Like the previous case, the nodes and edges checked with "[ ]" are found in the first phase; others were added in the second phase. This substructure shows a relation between two enzymes which shares a compound as a substrate by one and a product by another. The enzyme-enzyme relation has a relationship with two reactions: R01092 and R01105 [26]. R01092 is catalyzed by the enzyme of the gene, dme:CG5288-PA, and R01105 is catalyzed by the enzyme of the gene, dme:CG9092-PA. The substrate of R01092 is the C05796 compound (Galactin). The product of this reaction is C00124 (D-Galactose), which is also the substrate of R01092. R01092 produces C00446 (alpha-D-Galactose 1-phosphate) as the product compound. The relation in this substructure has the *link* as a pointer to C00124 because this compound is the shared metabolite in two reactions catalyzed by two enzymes connected within this relation.

SUB_1 and SUB_4 are found in all experimental sets (species), and SUB_8 is commonly found in ath, dme, eco and sce. A hierarchical clustering presents the common relations, which shows how biological molecules work interactively with others in the different species. This provides system-level understanding of metabolic pathways.

6 Conclusion

Systems biology views an organism as a system. System-level understanding indispensably involves integrating heterogeneous data and a variety of relations among the entities. The biological network is a crucial way to describe the biological system. Biological networks include various biomolecules and assorted relationships among molecules. Structure analysis of metabolic pathways allows us to understand how biomolecules interact with others. The research on the relations can play a contributive role in systems biology.

This research shows several methods of structure analysis on metabolic pathways. Substructure discovery on the same metabolic pathways from two species reveals the unique features of the pathways related to the species. Even in the cases that SUBDUE cannot find a unique substructure distinguishing two pathways, the number or the location of the instances of the substructure is able to distinguish them; how many specific relations or what specific relations are included into the pathway. Supervised learning shows the substructures that can identify what is unique about a specific type of pathway, which allows us to understand better how pathways differ. Unsupervised learning generates hierarchical clusters that reveal what is common about a specific type of pathways, which provides us better understanding of the common structure in pathways.

Moreover, our results show that the substructures discovered by SUBDUE have understandable biological meaning. These substructures, when considered as building blocks, can be used to construct new metabolic pathways. Ultimately, we can consider these substructures as guides to define a graph grammar for metabolic pathways that would improve both our ability to generate new networks and our comprehension of pathways [18]. These building blocks of metabolic pathways open our sights to an advanced application: drug discovery. The substructure of metabolic pathways learned by SUBDUE allows us to identify the target place of the drug in pathways. In addition a graph grammar of relational patterns on metabolic pathways can guide us to simulate the drug interaction on pathways.

Our future works include graph-based relational learning on graphs representing dynamics of biological networks and association with other methodologies for efficient learning on biological networks.

References

1. Aittokallio, T., Schwikowski, B.: Graph-based methods for analysing networks in cell biology. Briefings in Bioinformatics 7(3), 243–255 (2006)
2. Bu, D., Zhao, Y., Cai, L., et al.: Topological structure analysis of the protein-protein interaction network in budding yeast. Nucleic Acids Research 31, 2443–2450 (2003)

3. Cheng, C.Y., Huang, C.Y., Sun, C.T.: Mining bridge and brick motifs from complex biological networks for functionally and statistically significant discovery. IEEE Transactions on Systems, Man, and Cybernetics, Part B 38(1), 17–24 (2008)
4. Chittimoori, R., Holder, L., Cook, D.: Applying the subdue substructure discovery system to the chemical toxicity domain. In: Proceedings of the Florida AI Research Symposium, pp. 90–94 (1999)
5. Cook, D., Holder, L.: Substructure discovery using minimum description length and background knowledge. Journal of Artificial Intelligence Research 1, 231–255 (1994)
6. Cook, D., Manocha, N., Holder, L.: Using a graph-based data mining system to perform web search. International Journal of Pattern Recognition and Artificial Intelligence 17(5) (2003)
7. Dzerosk, S.: Multi-relational data mining: an introduction. SIGKDD Explorations Newsletter 5(1), 1–16 (2003)
8. Gonzalez, J., Holder, L., Cook, D.: Graph-based relational concept learning. In: Proceedings of the International Conference on Machine Learning, pp. 219–226 (2002)
9. Holder, L., Cook, D., Coble, J., Mukherjee, M.: Graph-based relational learning with application to security. Fundamenta Informaticae Special Issue on Mining Graphs, Trees and Sequences 6, 83–101 (2005)
10. Holder, L., Cook, D., Gonzalez, J., Jonyer, I.: Structural Pattern Recognition in Graphs. In: Pattern Recognition and String Matching, pp. 255–280. Springer, Heidelberg (2003)
11. Hu, H., Yan, X., Huang, Y., Han, J., Zhou, X.J.: Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics 21(1), 213–221 (2005)
12. Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J., Tropsha, A.: Mining protein family specific residue packing patterns from protein structure graphs. In: Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB), pp. 308–315 (2004)
13. Hwang, W., Cho, Y.R., Zhang, A., Ramanathan, M.: A novel functional module detection algorithm for protein-protein interaction networks. Algorithms for Molecular Biology 1 (2006)
14. Kanehisa, M., Goto, S., Kawashima, S., Okuno, U., Hattori, M.: KEGG resource for deciphering the genome. Nucleic Acids Research 32, 277–280 (2004)
15. Kitano, H.: Systems biology: A brief overview. Science 295, 1662–1664 (2002)
16. Klipp, E., Herwig, R., Kowald, A., Wierling, C., Lehrach, H.: Systems Biology in Practice: Concepts, Implementation and Application, 1st edn. WILEY-VCH, Weinheim (2005)
17. Koyuturk, M., Grama, A., Szpankowski, W.: An efficient algorithm for detecting frequent subgraphs in biological networks. In: Proceedings of the International Conference on Intelligent Systems for Molecular Biology, vol. 20, pp. 200–207 (2004)
18. Kukluk, J., You, C., Holder, L., Cook, D.: Learning node replacement graph grammars in metabolic pathways. In: Proceedings of International Conference on Bioinformatics and Computational Biology, BIOCOMP 2007 (2007)
19. Kuramochi, M., Karypis, G.: Frequent subgraph discovery. In: Proceedings of the IEEE Conference on Data Mining, pp. 313–320 (2001)

20. Laub, M., Loomis, W.: A molecular network that produces spontaneous oscillations in excitable cells of dictyostelium. Mol. Biol. Cell 9(12), 3521–3532 (1998)
21. Lodhi, H., Muggleton, S.H.: Modelling metabolic pathways using stochastic logic programs-based ensemble methods. In: Danos, V., Schachter, V. (eds.) CMSB 2004. LNCS (LNBI), vol. 3082, pp. 119–133. Springer, Heidelberg (2005)
22. Muggleton, S.: Inductive logic programming. New Generation Computing 8, 295–318 (1991)
23. Muggleton, S.H., Lodhi, H., Amini, A., Sternberg, M.J.E.: Support Vector Inductive Logic Programming. In: Hoffmann, A., Motoda, H., Scheffer, T. (eds.) DS 2005. LNCS, vol. 3735, pp. 163–175. Springer, Heidelberg (2005)
24. Su, S., Cook, D., Holder, L.: Application of knowledge discovery to molecular biology: Identifying structural regularities in proteins. In: Proceedings of the Pacific Symposium on Biocomputing, vol. 4, pp. 190–201 (1999)
25. Tamaddoni-Nezhad, A., Kakas, A., Muggleton, S., Pazos, F.: Modelling inhibition in metabolic pathways through abduction and induction. In: Camacho, R., King, R., Srinivasan, A. (eds.) ILP 2004. LNCS, vol. 3194, pp. 305–322. Springer, Heidelberg (2004)
26. KEGG, http://www.genome.jp
27. Wolf, J., Sohn, H., Heinrich, R., Kuriyama, H.: Mathematical analysis of a mechanism for autonomous metabolic oscillations in continuous culture of saccharomyces cerevisiae. FEBS Lett. 499(3), 230–234 (2001)
28. Yan, X., Han, J.: Gspan: Graph-based substructure pattern mining. In: Proceedings of the IEEE Conference on Data Mining, pp. 721–724 (2002)
29. You, C., Holder, L., Cook, D.: Graph-based data mining in dynamic networks: Empirical comparison of compression-based and frequency-based subgraph mining. In: IEEE International Conference on Data Mining (ICDM) Workshop on Analysis of Dynamic Networks (2008)
30. You, C., Holder, L., Cook, D.: Graph-based temporal mining of metabolic pathways with microarray data. In: ACM SIGKDD Workshop on Data Mining in Bioinformatics, BIOKDD (2008)