Discovering Recursive Patterns in Biological Networks

Jacek Kukluk, Chang Hun You, Lawrence Holder, and Diane Cook {kukluk, changhun, holder,cook}@cse.uta.edu, Department of Computer Science and Engineering, University of Texas at Arlington www.subdue.org

Graph grammars can represent relations, hierarchical structures and generalize knowledge in graph domains which string grammars cannot. We introduce an algorithm for graph grammar inference which builds on previous work in discovering frequent subgraphs in a graph

Algorithm

The algorithm generates candidate substructures and evaluates them using the following measure of compression.

$$\frac{size(G)}{size(S)+size(G\,|\,S)}$$

G-graph
S-substructure

$size(t)=vertices(t) + edges(t)$ or alternatively we compute $size(t)$ as Minimum Description Length *(MDL)* of t. We check for overlap among the instances of the subgraphs in the input graph. If subgraphs overlap by one node we propose a node replacement grammar production. Below, the subgraphs overlap at nodes 3 and 4.

A graph with overlapping substructures and a graph grammar representation of it

Connection instructions
1-3
1-4
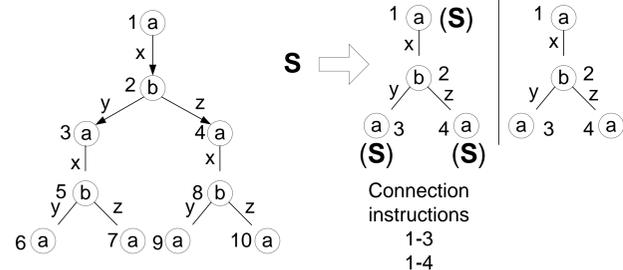
The embedding mechanism of the grammar consists of connection instructions. Every connection instruction is a pair of vertices that indicate where the production graph can connect to itself in a recursive fashion.

### Substructure and its instances while determining connection instructions
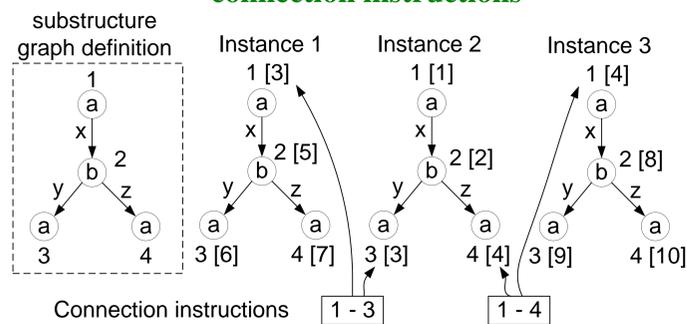
Experiments on biological networks

The KEGG database has 37,253 pathways generated from 296 reference pathways. A graph consisting of vertices and edges between these vertices is a natural data sturcture to represent biological networks.

### TCA cycle metabolic pathway of Homo Sapiens
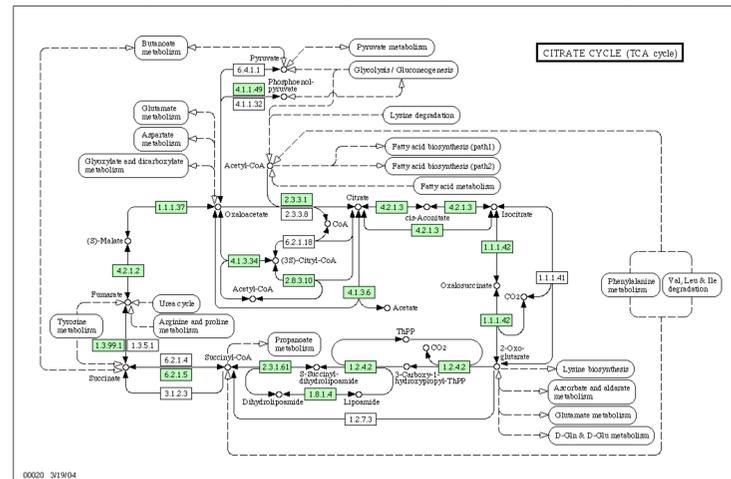
### Application of graph-based data mining to metabolic pathways
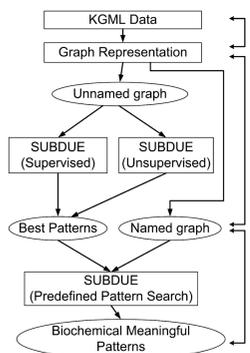
Supervise learning finds the unique substructures in a specific type of pathway, which help us to understand better how pathways differ. Unsupervise learning shows hierarchical clusters that describe the common substructures in a specific type of pathway which allow us to better understand the common features.

### The named graph representation of a metabolic pathway
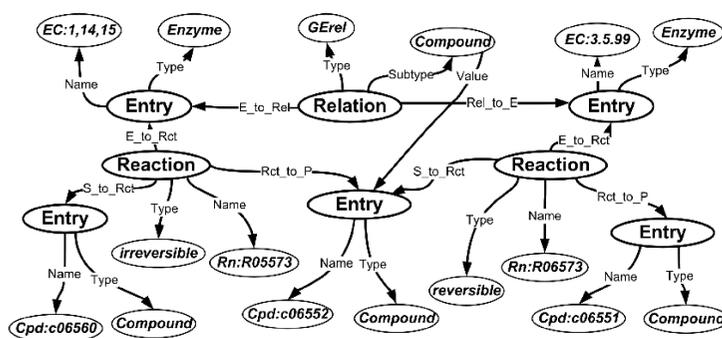
Hierarchical Clusters of metabolic pathways in fruit fly

Applying clustering to metabolic pathways with hierarchical topologies provides a better understandable blueprint.
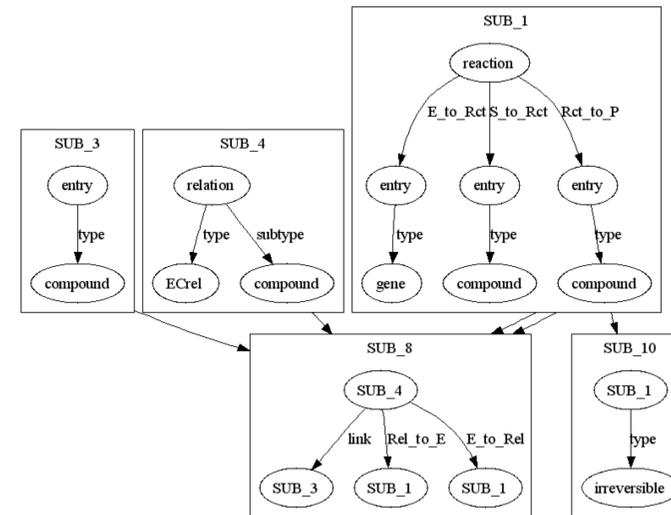

Recursive, hierarchical pattern discovered in Glycolysis / Gluconeogenesis - Escherichia coli pathway (eco00010)

We can find graph grammar in biological networks which show the hierarchical organization of it.
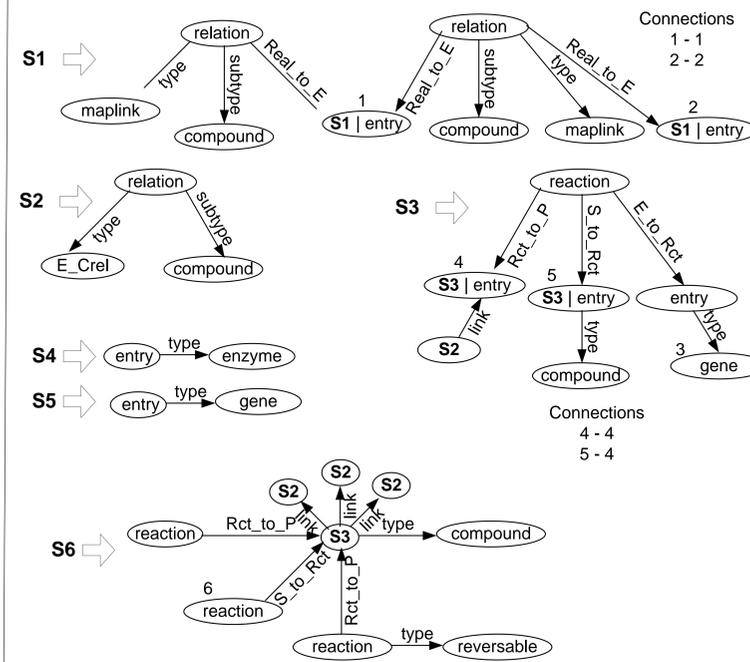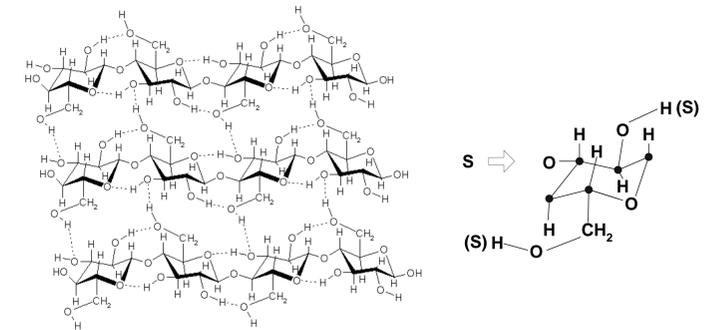

Experiments on chemical structures

The structure of dendronized polymer and its representation in hierarchical graph grammar productions.

Chemical structure   Inferred production
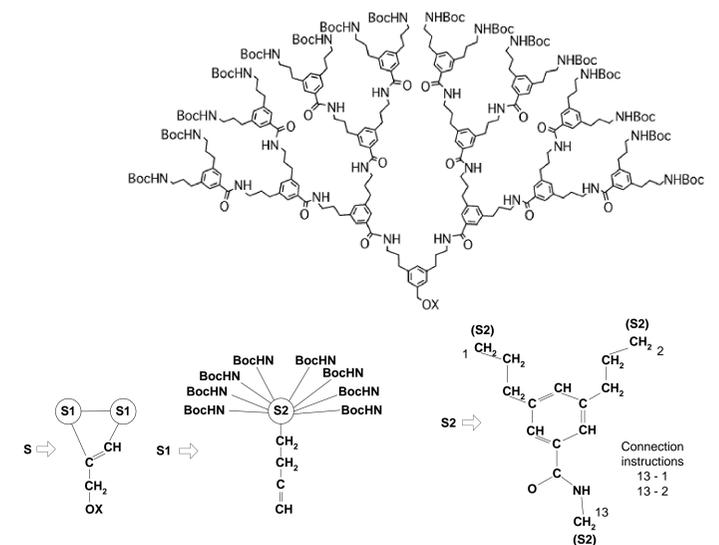
Cellulose with hydrogen bonding

water-soluble tin-based metallodendrimer