Int. J. Appl. Math. Comput. Sci., 2008, Vol. 18, No. 2, 241–250
DOI: 10.2478/v10006-008-0022-y

amcs

INFERRING GRAPH GRAMMARS BY DETECTING OVERLAP IN FREQUENT SUBGRAPHS

JACEK P. KUKLUK *, LAWRENCE B. HOLDER **, DIANE J. COOK **

* Dana-Farber/Brigham and Women's Cancer Center
Harvard Medical School
75 Francis Street, Boston, MA 02115, USA
e-mail: jkukluk@lroc.harvard.edu

** School of Electrical Engineering and Computer Science
Washington State University, Box 642752, Pullman, WA 99164, USA
e-mail: holder@wsu.edu, cook@eecs.wsu.edu

In this paper we study the inference of node and edge replacement graph grammars. We search for frequent subgraphs and then check for an overlap among the instances of the subgraphs in the input graph. If the subgraphs overlap by one node, we propose a node replacement graph grammar production. If the subgraphs overlap by two nodes or two nodes and an edge, we propose an edge replacement graph grammar production. We can also infer a hierarchy of productions by compressing portions of a graph described by a production and then inferring new productions on the compressed graph. We validate the approach in experiments where we generate graphs from known grammars and measure how well the approach infers the original grammar from the generated graph. We show graph grammars found in biological molecules, biological networks, and analyze learning curves of the algorithm.

Keywords: grammar induction, graph grammars, graph mining, multi-relational data mining.

1. Introduction

Noam Chomsky (1956) pointed out that one of the main concerns of a linguist is to discover simple grammars for natural languages and study those grammars with the hope of finding a general theory of linguistic structure. While string grammars represent language, we are looking for graph grammars that represent graph properties and can generalize these properties from finite graph examples into generators that can generate an infinite number of graphs. String grammars can be inferred from a finite number of sentences and generalize to an infinite number of sentences. Inferring graph grammars will generalize the knowledge from the examples into a concise form and generalize to an infinite number of entities from the domain.

We developed an algorithm to infer graph grammars from structured data represented as a graph. In this work we use graph grammar inference as a data mining tool. Graph grammars in our study show interesting patterns and organize data into a hierarchy. We implemented the algorithms and tested them on synthetic and nonsynthetic data. For this reason we developed and implemented a generator which generates a graph from a known graph grammar. We showed how inferring graph grammars depends on the presence of noise, the complexity of the graph grammar structure and the number of different labels present in the graph. We show how the algorithms perform in inferring grammars from biological networks and how the inference error in this domain depends on the number of examples in the input set.

2. Related work

A vast amount of research has been done in inferring grammars. These analyses focus on string grammars where symbols appear in a sequence. We are concerned with graph grammars, which can represent much larger classes of problems than string grammars. Only a few studies can be found in graph grammar inference.

Jeltsch and Kreowski (1990) did a theoretical study of inferring hyperedge replacement graph grammars from

simple undirected, unlabeled graphs. Their paper leads through an example where from four complete bipartite graphs $(K3,1; K3,2; K3,3; K3,4)$ the authors describe the inference of a grammar that can generate a more general class of bipartite graphs $(K3, n)$, where $n \geq 1$. The authors define four operations that lead to a final hyperedge replacement grammar. Jeltsch and Kreowski start the process from a grammar which has all the sample graphs in its productions. Then they transform the initial productions into productions that are more general but can still produce every graph from the sample graphs. Their approach guarantees that the final grammar will generate graphs that contain all sample graphs.

Oates, Doshi, and Huang (2003) discuss the problem of inferring probabilities of every grammar rule for stochastic hyperedge replacement context free graph grammars. They call their program Parameter Estimation for Graph Grammars (PEGG). They assume that the grammar is given. Given a structure of a grammar S and a finite set of graphs E generated by the grammar S, they ask what are the probabilities θ associated with every rule of the grammar. Their strategy is to look for a set of parameters θ that maximizes the probability $p(E \mid S, \theta)$.

In terms of similarity to string grammar inference we consider the Sequitur system developed by Nevill-Manning and Witten (1997). Sequitur infers a hierarchical structure by replacing substrings based on grammar rules. The new, compressed string is searched for substrings which can be described by the grammar rules, and they are then compressed with the grammar and the process continues iteratively. Similarly, in our approach we replace the part of a graph described by the inferred graph grammar with a single node and we look for grammar rules on the compressed graph, and repeat this process iteratively until the graph is fully compressed.

Jonyer *et al.*'s approach to node-replacement graph grammar inference (Jonyer *et al.* 2002; Jonyer *et al.* 2004) starts by finding frequently occurring subgraphs in the input graphs. They check if isomorphic instances of the subgraphs that minimize the measure are connected by one edge. If they are, a production $S \rightarrow PS$ is proposed, where P is the frequent subgraph. P and S are connected by one edge. Jonyer's method of testing if subgraphs are adjacent by one edge limits his grammars to descriptions of "chains" of isomorphic subgraphs connected by one edge. Since an edge of a frequent subgraph connecting it to the other isomorphic subgraph can be included in the subgraph structure, testing subgraphs for overlap allows us to propose a class of grammars that have more expressive power than the graph structures covered by Jonyer's grammars. For example, testing for overlap allows us to propose grammars which can describe tree structures, while Jonyer's approach does not allow for tree grammars.

3. Definitions

We give the definition of a graph and a graph grammar which is relevant to our approach and the implemented system. The defined graph has labels on vertices and edges. Every edge of the graph can be directed or undirected. The definition of a graph grammar describes the class of grammars that can be inferred by our approach. We emphasize the role of recursive productions in the name of the grammar, because the type of inferred productions is such that the nonterminal label on the left side of the production appears one or more times in the node labels of a graph on the right side. This is the main characteristic of our grammar productions. Our approach can also infer nonrecursive productions. 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.

A *labeled graph* G is the sextuple $G = (V, E, \mu, \nu, \eta, L)$, where V is the set of nodes,$E \subseteq V \times V$ is the set of edges, $\mu : V \rightarrow L$ is a function assigning labels to the nodes, $\nu : E \rightarrow L$ is a function assigning labels to the edges, $\eta : E \rightarrow 0, 1$ is a function assigning direction property to the edges (0 if undirected, 1 if directed). L is a set of labels on the nodes and the edges.

Example 1. The graph on the left in Fig. 2 has the following components:

$$
\begin{aligned}
V &= \{1,2,3,4,5,6,7,8,9,10\},\\
E &= \{(1,2),(2,3),(2,4),(3,5),(5,6),(5,7),(4,8),\\
&\quad (8,9),(8,10)\},\\
L &= \{a,b,x,y,z\},\\
\mu &= (1:a, 2:b, 3:a, 4:a, 5:b, 6:a, 7:a, 8:b, 9:a,\\
&\quad 10:a),\\
\nu &= ((1,2):x,(2,3):y,(2,4):z,(3,5):x,(5,6):y,\\
&\quad (5,7):z,(4,8):x,(8,9):y,(8,10):z),\\
\eta &= ((1,2):1,(2,3):1,(2,4):1,(3,5):1,(5,6):1,\\
&\quad (5,7):1,(4,8):1,(8,9):1,(8,10):1).
\end{aligned}
$$

♦

A *node replacement recursive graph grammar* is the quadruple $Gr = (\Sigma, \Delta, \Gamma, P)$, where Σ is an alphabet of node labels, Δ is an alphabet of terminal node labels, $\Delta \subseteq \Sigma$, Γ is an alphabet of edge labels, which are all terminals, is a finite set of productions of the form (d, G, C), where $d \in \Sigma - \Delta$, G is a graph, C is an embedding mechanism with a set of connection instructions, $C \subseteq V \times V$, V being the set of nodes of G. A connection instruction $(v_i, v_j) \in C$ implies that derivation can take place by replacing v_i in one instance of G with v_j in another instance of G. All

the edges incident to v_i are incident to v_j. All the edges incident to v_j remain unchanged.

Example 2. The grammar on the right in Fig. 2 has a set P of two productions. If we refer to the nonterminal graph (with three (S)s next to the nodes on the left of the vertical line) as $G1$ and to the terminal graph (on the right of the vertical line) as $G2$, then the productions are

$$P = \Big\{(S, G1, 1\text{–}3, 1\text{–}4), (S, G2, 1\text{–}3, 1\text{–}4)\Big\},$$
$$\Sigma = \{a, b, S\}, \quad \Delta = \{a, b\}, \quad \Gamma = \{x, y, z\}.$$

♦

An *edge replacement recursive graph grammar* is the quintuple $Gr = (\Sigma, \Delta, \Gamma, \Omega, P)$, where Σ is an alphabet of node labels, Δ is an alphabet of terminal node labels, $\Delta \subseteq \Sigma$, Γ is an alphabet of edge labels, Ω is an alphabet of terminal edge labels, $\Omega \subseteq \Sigma$, P is a finite set of productions of the form (d, G, C), G is a graph, where $d \in \Gamma - \Omega$, C is an embedding mechanism with a set of connection instructions, $C \subseteq (V \times V; V \times V$, where V is the set of nodes of G. A connection instruction $(v_i, v_j; v_k, v_l) \in C$ implies that derivation can take place by replacing v_i and v_k in one instance of G with v_j and v_l, respectively, in another instance of G. All the edges incident to v_i are incident to v_j, and all the edges incident to v_k are incident to v_l. All the edges incident to v_j and v_k remain unchanged. If, in the derivation process after applying a connection instruction $(v_i, v_j; v_k, v_l)$, nodes v_i and v_j are adjacent by an edge, we call edge $e = (v_i, v_j)$ a *real edge*, otherwise edge $e = (v_i, v_i)$ is used only in the specification of the grammar and we call this edge a *virtual edge*.

In Fig. 1 we show an example of an edge replacement recursive graph grammar. This example corresponds to a graph in Fig. 4. The grammar has two productions: one nonterminal production with (S) on edges and one terminal production. All nonterminal edges identified by (S) are real edges. There are two connection instructions $(2\text{-}3, 6\text{-}4)$ and $(1\text{-}1, 2\text{-}2)$

We introduce the definition of two data structures used in our algorithm.

A substructure S of a graph G is a data structure which consists of: (i) a graph definition of a substructure SG which is a graph isomorphic to a subgraph of G, (ii) a list of instances $(I_1, I_2, \ldots, I_n)$ where every instance is a subgraph of G isomorphic to S_G.

A *recursive substructure* $recursiveSub$ is a data structure which consists of:

(i) a graph definition of a substructure S_G which is a graph isomorphic to a subgraph of G,

(ii) a list of connection instructions which are pairs of integer numbers describing how instances of the substructure can overlap to comprise one instance of the corresponding grammar production rule,

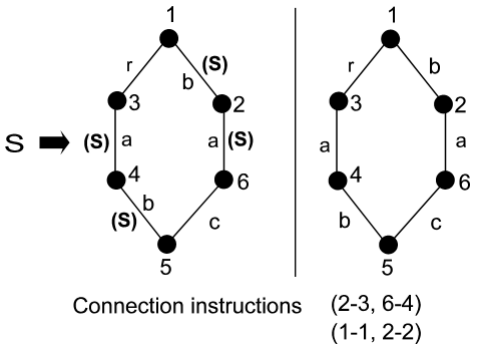

Fig. 1. Edge replacement recursive graph grammar example.

(iii) a list of recursive instances $(IR_1, IR_2, \ldots, IR_n)$ where every instance IR_k is a subgraph of G. Every instance IR_k consist of one or more isomorphic copies of S_G, overlapping by no more than one vertex in the algorithm for node graph grammar inference and no more than two vertices in edge grammar inference.

We show an example of a recursive substructure in Fig. 3.

In our definition of a substructure we refer to subgraph isomorphism. However, in our algorithm we are not solving the subgraph isomorphism problem. We are using a polynomial time beam search to discover substructures and a graph isomorphism to collect instances of the substructures.

4. Graph grammar inference algorithms

The example in Fig. 2 shows a graph composed of three overlapping substructures. The algorithm generates candidate substructures and evaluates them using any one of the learning biases, which are discussed later. The input to our algorithm is a labeled graph G which can be one connected graph or a set of graphs. G can have directed or undirected edges. The algorithm starts by creating a list of substructures where every substructure is a single node and its instances are all nodes in the graph with the same node label. Initially, the best substructure is the node with most instances. The substructures are ranked and placed on the expansion queue Q. It then extends all substructures in Q in all possible ways by a single edge and a node or only by a single edge if both nodes are already in the graph definition of the substructure. We keep all extended substructures in $newQ$. We evaluate substructures in $newQ$ according to the chosen evaluation heuristic.

The total number of substructures considered is determined by the input parameter $Limit$. The best substructure identified becomes the right side of the first grammar production, and the graph G is compressed using this

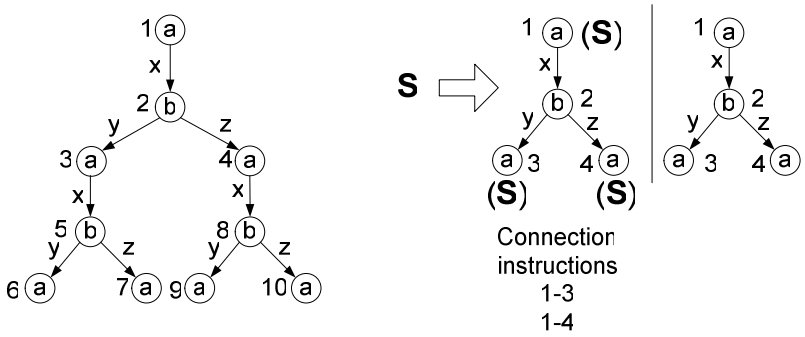

Fig. 2. Graph with overlapping substructures and its graph grammar representation.

best substructure. Compression replaces every instance of the best substructure with a single nonterminal node. This node is labeled with a nonterminal label. The compressed graph is further processed until it cannot be compressed any more, or some user-defined stopping condition is reached (e.g., the maximum number of productions). In consecutive iterations the best substructure can have one or more nonterminal labels. It allows us to create a hierarchy of grammar productions. The input parameter $Beam$ specifies the width of the beam search, i.e., the length of Q. Algorithm 1 shows the pseudocode.

Recursive productions are identified during the previously described search process by allowing instances to grow and overlap. Any two instances are allowed to overlap by only one vertex. The recursive substructure is evaluated along with nonrecursive substructures and is competing with nonrecursive substructures for placement on Q. Connection instructions are created by determining which nodes overlapped across instances. Figure 3 shows an example of a substructure that is the right side of a recursive rule, along with its connection instructions (Kukluk *et al.*, 2006).

The edge replacement algorithm operates on a data structure called a *substructure* (similarly to the algorithm for node replacement grammar inference). A substructure consists of a graph definition of the repetitive subgraph and its instances. We illustrate it in Fig. 4. We grow substructures similarly as in the algorithm for node replacement graph grammar inference, and then we examine instances for overlap. If nodes v_1 and v_2 in G belong to two different instances (two overlapping instances), we propose a recursive grammar rule. We determine the type of nonterminal edge. If v_1 and v_2 are adjacent by an edge, it is a real edge, and we determine its label which we use to specify the terminating production. If v_1 and v_2 are not adjacent, then the nonterminal edge is virtual. In Fig. 4 we illustrate how we determine connection instructions.

One advantage of our algorithm is its modular design in which the evaluation of candidate grammar rules is done separately from the generation of these candidates. The result is that any evaluation metric can be used to drive the search. Different evaluation metrics are part of the system and can be specified as arguments. We have had great success with the minimum description length (MDL) principle on a wide range of domains. MDL is an information theoretic approach (Rissanen, 1989). The description length of the substructure S given the input graph G is calculated as $DL(S,G) = DL(S) + DL(G|S)$, where $DL(S)$ is the description length of the subgraph, and $DL(G|S)$ is the description length of the input graph compressed by the subgraph (Cook and Holder, 1994, 2000). An alternative measure is the size heuristic which is computed as

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

where G is the input graph, S is a substructure and $G|S$ is the graph derived from G by compressing each instance of S into a single node. Here $\text{size}(t)$ can be computed simply by summing the numbers of nodes and edges: $\text{size}(t) = \text{vertices}(t) + \text{edges}(t)$. The third measure is called 'setcover', which is used for concept learning tasks employing sets of disconnected graphs. This measure maximizes the number of positive examples in which the grammar production is found while minimizing the number of such negative examples.

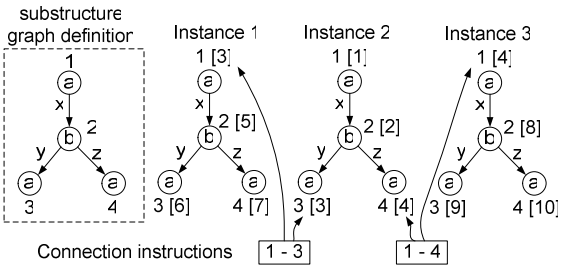

Fig. 3. Substructure and its instances while determining connection instructions (continuation of the example from Fig. 2).

Algorithm 1 Graph grammar discovery

procedure INFER_GRAMMAR(graph G, integer $Beam$, integer $Limit$)
  grammar $\leftarrow \{\}$
  repeat
    $Q \leftarrow \{v|v$ is a node in G having a unique label$\}$
    $bestSub \leftarrow$ first substructure in Q
    repeat
      $newQ \leftarrow \{\}$
      for each substructure $S \in Q$ **do**
        $newSubs \leftarrow$ extend S in all possible ways by a single edge and a node
        $recursiveSub \leftarrow$ RECURSIFY_SUBSTRUCTURE(S)
        $newQ \leftarrow newQ \cup newSubs \cup recursiveSub$
        $Limit \leftarrow Limit - 1$
        evaluate substructures in $newQ$, maintain $length(newQ) < Beam$ eliminating substructure with the lowest value if necessary
      end for
      if best substructure in $newQ$ is better than $bestSub$ **then**
        $bestSub \leftarrow$ best substructure in $newQ$
        $Q \leftarrow newQ$
      end if
    until Q is empty or $Limit \leq 0$
    $grammar \leftarrow grammar \cup bestSub$
    $G \leftarrow G$ compressed by $bestSub$
  until *bestSub* cannot compress the graph G
end procedure

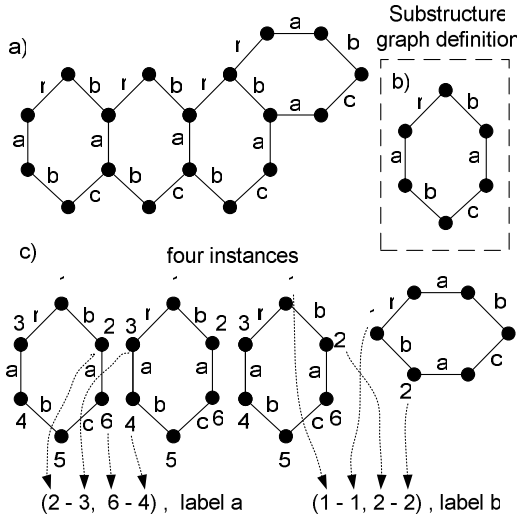

Fig. 4. Input graph (a), substructure graph definition (b) and four overlapping instances of the repetitive subgraph (c).

Our algorithms make use of the substructure discovery algorithm described in Cook and Holder (2000). This algorithm uses a heuristic search whose complexity is polynomial in the size of the input graph. The overlap test is the main computationally expensive addition of our grammar discovery algorithm and it does not change its complexity. The number of nodes of an instance graph is no larger than V, where V is the number of nodes in the input graph. Checking two instances for overlap will not take more than $O(V^2)$ time. The number of pairs of instances is no more than V^2, so the entire overlap test will not take more than $O(V^4)$ time.

5. Experiments

5.1. Methodology. Having our algorithm implemented, we are faced with the challenge of evaluating its performance. There are an infinite number of grammars as well as graphs generated from these grammars. We seek to understand the relationship between graph grammar inference and grammar complexity, and so need a measure of grammar complexity. One such measure is the minimum description length (MDL) of a graph, which is the minimum number of bits necessary to completely describe the graph.

In our experiments we measure an error based on the structural difference. Another approach to measuring the accuracy of the inferred grammar would be based on a graph grammar parser. We would consider accurate the inferred grammars that can parse the input graph. A graph grammar parser would require a subgraph isomorphism test which is computationally expensive and much more difficult in implementation than the error measure we are using. For these reasons we did not pursue the implementation of a graph grammar parser.

We would like our error to be a value between 0 and 1. Therefore, we normalize the error by having in the denominator the sum of the size of the graph used in the original grammar and the number of nonterminals. We do not allow an error to be larger than 1. Therefore, we take the minimum of 1 and our measure as a final value. The restriction that the error is no larger than 1 prohibits unnecessary influence on the average error taken from several values by the inferred graph structure significantly larger than the graph used in the original grammar. We have

$$\text{Error} = \min\left(1, \frac{\text{matchCost}(g_1, g_2) + \mid \#CI - \#NT \mid}{\text{size}(g_1) + \#NT}\right),$$

where $\text{matchCost}(g_1, g_2)$ is the minimal number of operations required to transform g_1 to a graph isomorphic to g_2, or g_2 to a graph isomorphic to g_1. The operations are: the insertion of an edge or node, the deletion of a node or an edge, or the substitution of a node or edge label. More-

over, $\#CI$ is the number of inferred connection instructions, $\#NT$ is the number of nonterminals in the original grammar, and $size(g_1)$ is the sum of the number of nodes and edges in the graph used in the grammar production.

5.2. Error as a function of noise and complexity of a graph grammar. We used twenty nine graphs from Fig. 6 in grammar productions. We assigned different labels to nodes and edges of these graphs except three nodes used for nonterminals. As noise we added nodes and edges to the generated graph structure. We compute the number of added nodes from the formula $(noise/(1 - noise)) \times number\_of\_nodes$. A similar formula is employed for edges. We generated graphs with noise from 0 to 0.9 in 0.1 increments. For each value of noise and MDL we generated thirty graphs from the known grammar and inferred the grammar from the generated graph. We computed the inference error and averaged it over thirty examples. We generated 8700 graphs to plot each of the three graphs in Fig. 5. The first plot shows the results for grammars with one nonterminal. The second and third plots show the results for grammars with two and three nonterminals.

We averaged the value of an error over ten values of noise, which gave us the value we could associate with the graph structure. It allowed us to order graph structures used in the grammar productions based on the average inference error. In Fig. 6 we show all twenty nine connected simple graphs with three, four and five nodes used in productions ordered in nondecreasing MDL values of a graph structure.

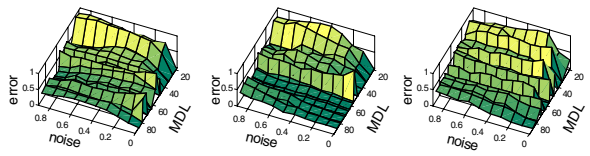

Fig. 5. Error as a function of noise and MDL where the graph structure was not corrupted (one, two and three nonterminals, respectively).

5.3. Error as a function of the number of labels. We would like to evaluate how the error depends on the number of different labels used in a grammar. We restricted graph structures used in productions to graphs with five nodes. Every graph structure was labeled with $1, 2, 3, 4, 5$ or 6 different labels. For each value of MDL and the number of labels we generated 30 different graphs from the grammar and computed average errors between them and the learned grammars. The generated graphs were without noise. We show the results for one, two, and three nonterminals in Fig. 7. For clarity, below the three-dimensional plots, we give two-dimensional plots with triangles representing the errors. The larger and lighter the triangle, the larger the error. We see that the error increases as the number of different labels decreases. On the two-dimensional plots we see the shift in the error towards graphs with higher MDL when the number of nonterminals increases.

5.4. Learning curves. We wanted to examine the learning process on a graph grammar with several productions. Since there are an infinite number of different graph grammars, we decided to select one example with several different graph structures used in the grammar productions. We show this example in Fig. 8, where we see the graph grammar used to generate graphs. There are five productions. The last production with only one node is a terminating production. Each graph in the first four productions had two nonterminal nodes. The first four productions are chosen with probability 0.1 in the generation process. The terminating production is chosen with probability 0.6.

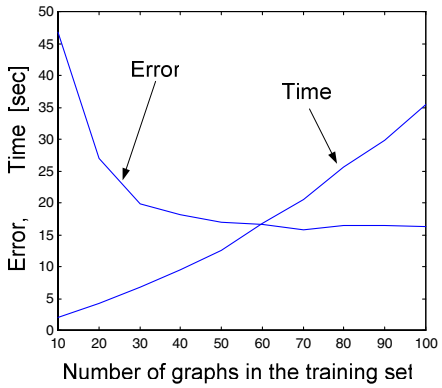

Fig. 9. Error and time as a function of the number of graphs in the training set.

We generate sets of graphs with 10, 20, 30, and up to 100 graphs generated from the grammar of Fig. 8. Every graph in the set has 30 to 40 nodes. We compare the first four grammar productions found by our algorithm to the original grammar of Fig. 8. As a measure of the error, we use the minimal match cost of a transformation from one graph structure to the other, as described in Section 5.1 where we discuss the measure of the error. We calculate the match cost of the structure of the graph from the first inferred grammar production to the four structures of the original productions and choose the smallest value. Then, we calculate the match cost of the structure from the second inferred production to the three structures from the original grammar not selected before and select the smallest value. Similarly, we find the smallest match cost between the structure of the third inferred production and the two structures left. We compare the last inferred production with the remaining production from the original grammar. The inference error was computed as a sum of the four errors we just introduced. We repeat generation

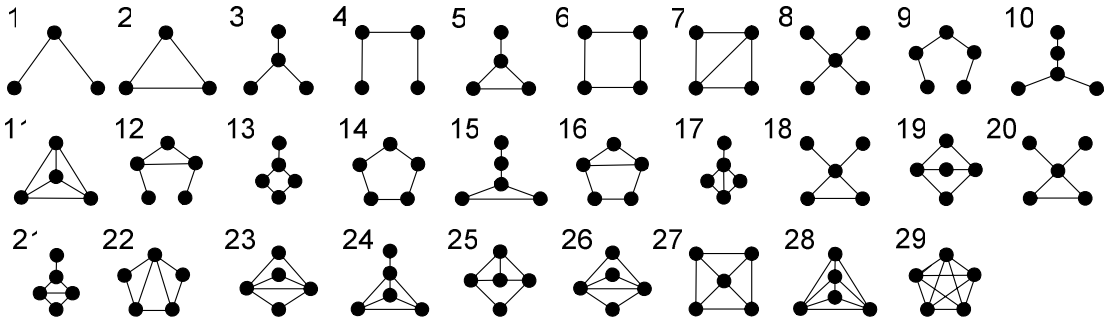

Fig. 6. Twenty nine simple connected graphs ordered according to nondecreasing MDL values.

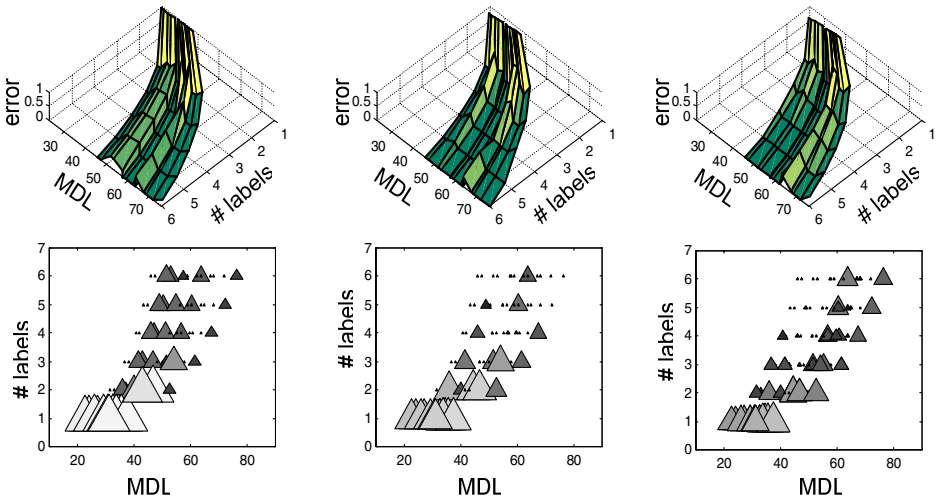

Fig. 7. Error as a function of MDL and the number of different labels used in a grammar definition (one, two and three nonterminals, respectively).

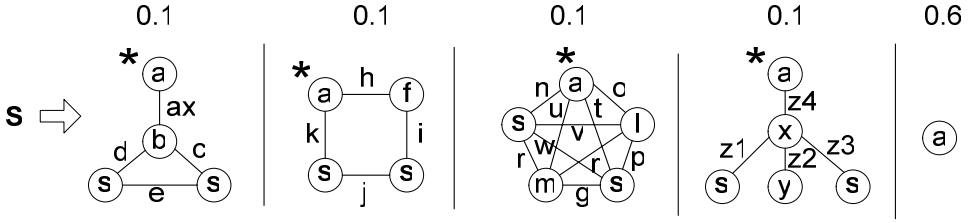

Fig. 8. Graph grammar used for graph generation.

and error determination thirty times and compute the average value of the error. In Fig. 9 we show the grammar inference error and time as a function of the number of graphs in the input set. We see that time in the range 10 to 100 graphs increases almost linearly. The error decreases sharply as we increase the set of graphs from 10 to 30. The error does not reach zero. The input graph has now four patterns. We often infer productions which contain two of the patterns or a portion of two patterns which causes the error.

5.5. Biological networks. The biological networks used in our experiments were from the Kyoto Encyclopedia of Genes and Genomes. (KEGG) (Kanehisa *et al.*, 2006). We use a graph representation which has labels on vertices and edges. The graphs represent processes like metabolism, membrane transport, and biosynthesis. We group the graphs into sets which allow us to search for common recursive patterns that can help to understand the basic building blocks and hierarchical organization of processes. The label entry represents a molecule, a molecule group or a pathway. A node labeled entry can be connected to a node labeled type. The type can be a value of the set: enzyme, ortholog, gene, group, compound, or map. A reaction is a process where a material is changed to another material catalyzed by an enzyme. For example, a reaction can have one or more enzyme entries, and one or more compounds. Labels on edges show relationships between entities. The meanings are as follows: Rct_to_P : reaction to product , S_to_Rct : substrate to reaction, E_to_Rct : enzyme (gene) to reaction, E_to_Rel : enzyme to relation, Rel_to_E : relation to enzyme. Nodes labeled $ECrel$ indicate an enzyme–enzyme relation meaning that two enzymes catalyze successive reactions.

In our experiments we use ten species. The abbreviated names of the species and their meanings are as follows: bsu – Bacillus subtilis, sty – Salmonella enterica serovar Typhi CT18, xcc – Xanthomonas campestris pv. campestris ATCC 33913, pto – Picrophilus torridus, mka – Methanopyrus kandleri, pho – Pyrococcus horikoshii, sfx – Shigella flexneri 2457T (serotype 2a), efa – Enterococcus faecalis, bar – Bacillus anthracis Ames 0581.

The species were selected randomly from the database. The number of networks was different for each species. We wanted to see how our algorithm performs when we increase the sample size of graphs supplied to our inference algorithm. For this purpose, we divided all the networks into 11 sets such that the last set (11-th) has all the species. Set 10 excludes the 11-th portion of all networks. Set 9 excludes 2/11 of all networks and Set 1 has 1/11 of all networks. If all networks in the species do not divide by 11 evenly, we distribute the remaining networks randomly to the eleven sets.

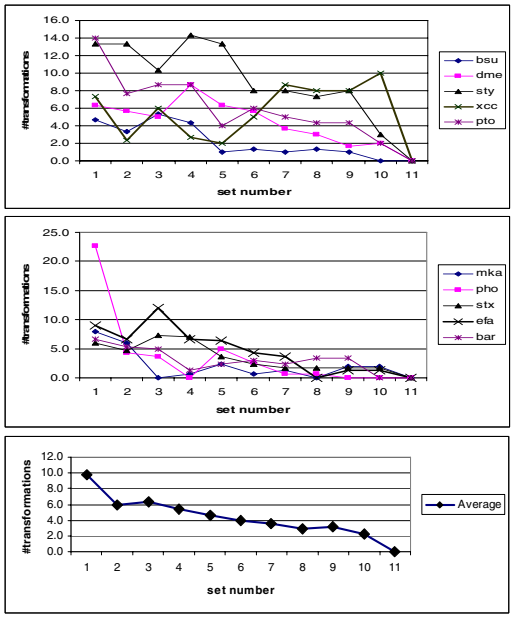

Fig. 10. Change in the inferred grammar measured in reference to the biggest set in networks of ten species.

We would like to compare our inferred grammar from sets of different sizes to the original, true, ideal grammar which represents the species. However, such a graph grammar is not known. In the first experiment, as an original grammar we adopted the grammar inferred from the last set. From each set we infer four grammar productions which score the highest in the evaluation. We compute the error (distance) of an inferred grammar to the grammar inferred from the set with all networks. The computation of an error is the same as that described in Section 5.4. The error is the minimal number of edges, vertices, and labels required to be changed or removed to transform the structure of graph productions from one grammar to another. In figures we refer to it as #transformations. In Fig. 10 we show the results of the experiment. Each value in Fig. 10 is an average from three runs. In every run we randomly shuffle the networks over 11 sets such that sets are different in every run. In Fig. 11 we show the graph grammar inferred from a set of thirty and a set of one hundred and ten graphs of Picrophilus torridus (*pto*).

The experiments on the biological network domain give us insight into the performance of the algorithm and to the biological networks. Examining Fig. 10 we notice that some species, like *dme*, have a very regular set of biological networks. Increasing the size of the set does not change the inferred grammar. Whereas in other species, like xcc, the set of biological networks is very diverse resulting in significant changes in the curve. Several curves, pto, pho, efa, gradually decrease with the last values be-

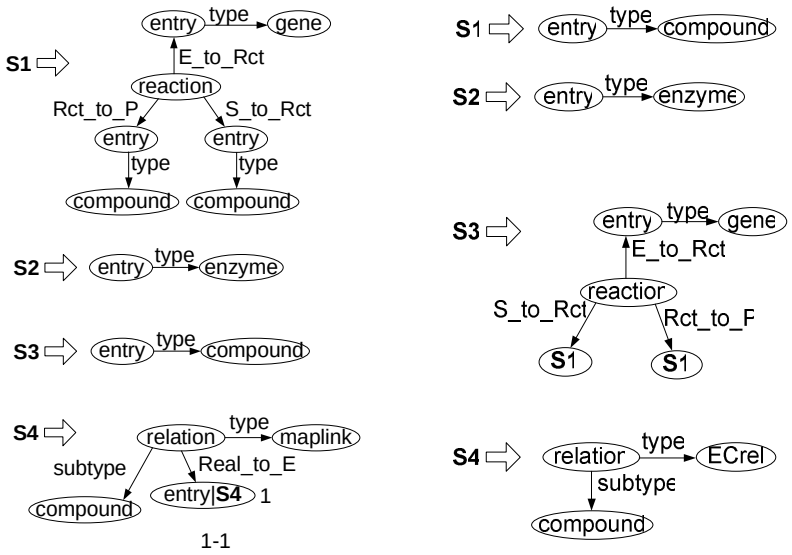

Fig. 11. Graph grammar inferred from a set of thirty (a), one hundred and ten (b) graphs of Picrophilus torridus (pto).

ing zero. This shows us that our algorithm performed well, and with an increasing number of graphs in the input set we find a grammar which does not change more with increased number of graphs, which indicates that the grammar found represents the input set well. The very bottom chart in Fig. 10 shows the average change. We see that with an increasing number of graphs in the input sets the curve declines to zero, which says that with an increasing number of graphs we infer a more accurate grammar.

6. Conclusions and future work

We have studied an algorithm for inferring node and edge replacement graph grammars. The algorithm starts from all nodes with the same label and grows them by adding to them one edge or a node and an edge at a time. We developed a substructure which consists of the definition of a graph and all subgraphs appearing in the input graph that are isomorphic to this graph definition (i.e., instances). The overlap of instances proposes a recursive graph grammar production which expresses the concepts of 'one or more' of the same substructures. The input graph to our algorithm is an arbitrary directed or undirected graph with labels on nodes and edges.

The node replacement recursive graph grammar inference algorithm limits productions to one single node on the left-hand side. The algorithm infers either recursive or nonrecursive productions depending on whether frequent subgraphs in the input graphs overlap or not. A smaller inference error occurs when the inferred pattern has a higher MDL value, i.e., it is more complex.

We proposed the inference of edge replacement recursive graph grammars as an extension to the algorithm for node replacement inference. We allowed for overlap by two nodes and we inferred grammars with a real or virtual edge. With this approach we can infer the grammar generating chains of squares overlapping on one edge, which was not possible with node replacement grammars. Patterns often overlap on two nodes in chemical structures. Therefore, we have an approach which can find and represent important patterns in the chemical domain.

Experiments with biological networks showed that our algorithm performed well. As the number of input graphs increases, the inferred grammar does not change, which indicates that the grammar found represents the input set well. We can use inferred grammar productions not only to provide an abstraction of recognized metabolic pathways for better understanding, but also to construct unknown metabolic pathways based on molecular-level experimental data.

Grammars inferred by the approach developed by Jonyer *et al.* (2004) were limited to chains of isomorphic subgraphs which must be connected by a single edge. Since the connecting edge can be included in the production subgraph, and isomorphic subgraphs will overlap by one vertex, our approach can infer Jonyer's class of grammars.

We would like to indicate general future directions in graph grammar inference research. They are as follows:

- Develop algorithms which allow for learning larger classes of graph grammars. We extended classes of presently learnable graph grammars. It is possible to extend it even further into context sensitive graph grammars where we could still replace nodes and edges, but whether or not this replacement takes place depends on the neighborhood of the replaced

node or edge. In order to regenerate structures, we would need a more sophisticated generation mechanism with a context sensitive embedding mechanism. This mechanism, inferred during induction, would indicate nodes to merge during the generation process. We can explore other techniques like the decomposition of graphs in searching for the best grammar which describes the data.

- Investigate learnable properties of graphs from the perspective of graph grammars.
- Identify experimental areas and show the significance of graph grammar inference in these domains. One of the new domains we approach (Ates *et al.*, 2006) are visual languages, where graph grammar inference from a sample of a language can give a grammar to be used to check newly written programs.
- Use graph grammar inference to identify building blocks, modularity and motifs in biology, software, social networks, and electronics circuits. We did experiments in biology and XML domains (Kukluk *et al.*, 2007). Biological and chemical structures are still very promising areas of the application of recursive graph grammars. Social networks, VLSI circuits, and the Internet are domains with relational data whose hierarchy and recursive properties can be explored with graph grammars.
- Expand graph grammar inference to learning stochastic graph grammars. This extension would require assigning a probability to each production. We can evaluate this probability based on the portion of the input graph covered by the inferred production (Oates, Doshi, and Huang, 2003).

References

Ates K., Kukluk J., Holder L., Cook D. and Zhang K. (2006). Graph grammar induction on structural data for visual programming, *Proceedings of the Conference Tools with Artificial Intelligence, Washington DC, USA*, pp. 232–242.

Chomsky N. (1956). Three models of language, *IRE Transactions on Information Theory* **2**(3): 113–24.

Cook D. and Holder L. (1994). Substructure discovery using minimum description length and background knowledge, *Journal of Artificial Intelligence Research* **1**: 231–255.

Cook D. and Holder L. (2000). Graph-based data mining, *IEEE Intelligent Systems* **15**(2): 32–41.

Jeltsch E. and Kreowski H. (1990). Grammatical inference based on hyperedge replacement. Graph-Grammars, *Lecture Notes in Computer Science* **532**: 461–474.

Jonyer I., Holder L. and Cook C. (2002). Concept formation using graph grammars, *Proceedings of the KDD Workshop on Multi-Relational Data Mining, Edmonton, Alberta, Canada*, pp. 71–79.

Jonyer I., Holder L. and Cook D. (2004). MDL-based context-free graph grammar induction and applications, *International Journal of Artificial Intelligence Tools*, **13**(1): 65–79.

Kanehisa M., Goto S., Hattori M., Aoki-Kinoshita K.F., Itoh M., Kawashima S., Katayama T., Araki M. and Hirakawa M. (2006). From genomics to chemical genomics: New developments in KEGG, *Nucleic Acids Res.* **34**: D354-357.

Kukluk J., Holder L. and Cook D. (2006). Inference of node replacement recursive graph grammars, *Proceedings of the 6-th SIAM International Conference on Data Mining*, Washington, USA, pp. 544–548.

Kukluk J., Hun You C., Holder L. and Cook D. (2007). Learning node replacement graph grammars in metabolic pathways, *International Conference on Bioinformatics & Computational Biology, (BIOCOMP'07), Las Vegas, NV, USA*, pp. 44–50.

Kuramochi M. and Karypis G. (2001). Frequent subgraph discovery, *Proceedings of the IEEE 2001 International Conference on Data Mining (ICDM '01)*, San Jose, CA, USA, pp. 313–320.

Neidle S. (Ed.) (1999). *Oxford Handbook of Nucleic Acid Structure*, Oxford, Oxford University Press.

Nevill-Manning G. and Witten H. (1997). Identifying hierarchical structure in sequences: A linear-time algorithm, *Journal of Artificial Intelligence Research*, **7**: 67–82.

Phan A., Kuryavyi V., Ma J., Faure A., Andreola M. and Patel D. (2005). An interlocked dimeric parallel-stranded DNA quadruplex: A potent inhibitor of HIV-1 integrase, *Proceedings of the National Academy of Sciences* **102**(3): 634–639.

Oates T., Doshi S. and Huang F. (2003). Estimating maximum likelihood parameters for stochastic context-free graph grammars, *Lecture Notes in Artificial Intelligence* **2835**: 281–298.

Rissanen J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific Company.

Yan X. and Han J., gSpan (2002): Graph-based substructure pattern mining, *Proceedings of the IEEE International Conference on Data Mining, Maebashi City, Japan*, pp. 721–724.

Received: 25 July 2007
Revised: 26 November 2007