Mining frequent subgraphs in a large and sparse biological graph

E-mail: pengjj@msu.edu
Plant Research Lab,
Michigan State University, East Lansing, MI 48823, USA

Mar 19, 2011

1 Introduction

Introduce the challenges to mine frequent subgraphs in a given large and sparse graph. And it's needs in biological graph mining applications. Introduce the basic idea of our algorithm, which is to partition the whole graph to size-k graphs and then efficiently build a level graph to find the relationships among these graphs to save the computational time in the subsequent frequent subgraph mining. The key to the success is how to efficiently build a level graph.

In cell biology, an organelle is a specialized subunit within a cell that has a specific function, and is usually separately enclosed within its own lipid bilayer. While some organelles are part of the endomembrane system and can receive and send things covered in membrane, other rely on proteins and enzymes regulated by feedback regulation and allosteric regulation mechanisms. The underlining mechanism of how organelles communicate in a cell is largely unknown. Communication patterns in social and biological networks have been extensively studied, such as structure and tie strengths in mobile communication networks[1], communication patterns in data centers [2], behavior

of techno-social systems[3], communication in neuronal networks[4], Inter-plant communication through common mycorrhizal networks[5] and local communities in protein networks[6]. Communication patterns discovered in the above communication networks are tightly associated with node or edge topological properties, such as betweenness, degree, etc., especially frequent subgraphs.

Literature reveals that complex interactions between organelles are often vital for the survival of organisms. The organelles communicate with each other to collaborate and perform complex functions that they can not do individually[7]. Biological networks are employed to model the communication pattern in many applications[8]. Discovering organelle communication patterns is promising because biological networks are robust and flexible but how this is achieved with regard to information exchange is largely unknown. As learned from the known biological communication patterns, sparse (rather than dense) interactions usually indicate important communication functions[9, 10]. Furthermore, we proposed a method to construct networks of organelle functional modules, called organelle network, based on gene expression data in previous work[]. In the network, nodes represent organelle functional modules or genes and edges represent the interaction of gene to gene, organelle to organelle or gene to organelle. Organelle communication patterns can be mining from the organelle network. Therefore, a critical computational problem is to find repeated sparse interacting entities in the organelle network.

To detect frequent subgraphs in a organelle network is a computationally challenging task. First, given that it is common for biological entities (genes and organelle modules) and their interactions in a organelle to participate in multiple biological functional modules, multiple biological entities or edges overlap subgraphs may be active at any time[11]. Hence, the downward closure of subgraph frequency, which is used to efficiently decrease the problem search space[12], does not hold,

which means we cannot detect frequent subgraphs from frequent subgraphs with less number of nodes[]. Second, the frequency counting procedure involves graph isomorphism testing[13] and subgraph isomorphism testing [14]. It is known that latter one is an NP-complete problem [15, 16] and former one is a special problem neither known to be solvable in polynomial time nor NP-complete[16].

This paper proposed an efficient algorithm, named E-FSM (efficiently frequent subgraph mining), to detect organelle communication pattern in a unlabeled undirected large sparse organelle network while allowing occurrences of the subgraphs to be arbitrarily overlapping. The main process of our algorithm is to partition the whole graph to size-k subgraphs and then efficiently group the subgraphs to save the computational time in the subsequent frequent subgraph mining, where k is a user defined subgraph size. The key of the whole algorithm is how to efficiently decrease the times of graph isomorphism testing.

The main contributions of this paper are as follows.

- First, a novel degree-sort based method is used to generated the adjacent matrix string of a subgraph.
- Second, an efficient frequent subgraph mining algorithm is proposed to find communication pattern in the organelle network.
- Third, the identified communication pattern between organelles shows that XXX

This paper is organized as follows. In section 2 we provide background for frequent subgraph detection and necessary definitions. Section 3 describes our method in detail. Section 4 shows some experiments and performance, which compared with XXX and XXX. Section 5 summarizes our work and shows some future work.

2 Background

In this paper, the communication pattern discovery problem is equal to mining frequent subgraphs from a organelle network, which is considered as unlabeled . To make the description clearer, the problem is defined as follows.

Definition 1. *An **unlabeled undirected sparse graph** is a tuple with two elements $G = \{V, E\}$ where V is a set of vertices and $E \subseteq V \times V$ is a set of undirected edges and $|E| \ll \frac{1}{2}|V| \cdot (|V| - 1)$.*

Definition 2. *Given two unlabeled undirected sparse graphs $G = \{V, E\}$ and $G' = \{V', E'\}$, G is a **subgraph** of G' iff: $V \subseteq V'$ and $E \subseteq E'$.*

Definition 3. *An **isomorphism** of graphs $G = \{V, E\}$ and $G' = \{V', E'\}$ is a bijective function between the vertex sets of G and G' f: $V(G) \longrightarrow V(G')$ such that $\forall(u, v) \in E$ iff $(f(u), f(v)) \in E'$. There is a subgraph isomorphism from G to G' if G' contains a subgraph that is isomorphic to G.*

Definition 4. *Given a unlabeled undirected sparse graph G and a minimum support threshold t, **frequent subgraph mining on a graph** is to identify every subgraph g in G, such that the occurrence frequency of g in G is not less than t.*

There has been an increased interest in graph mining. One of the most important problems in graph mining is frequent subgraph detection. Frequent subgraphs are essential composite features for numerous graph mining applications, such as graph classification[17, 18] and graph search[19]. Frequent subgraphs are also the key components of the network motif discovery algorithms[11, 20, 21, 22, 23, 24, 25].

There are two distinct categories in frequent subgraph detection that referred to as the graph-data setting and the single-graph setting. In the graph-data setting, the input of a subgraph mining algorithm is a

set of relatively small graphs. In the single-graph setting the input data is a single large graph. The difference affects the way the frequency of the various subgraphs is determined. For the graph-data setting, the frequency of a subgraph is determined by the number of graphs that the subgraph occurs in, irrespective of how many times a subgraph occurs in a particular graph, whereas in the single-graph setting, the frequency of a subgraph is based on the number of its occurrences, which also varies with different definitions of frequency graph.

In this paper, we focus on frequent subgraph detection in an unlabeled undirected sparse graph to find the frequent communication pattern from the organelle network. Compared to the research on mining frequent subgraph from graph-data setting, single-graph based algorithms problem have not been thoroughly studied[26]. Note that due to the inherent differences of the features of the underlying dataset and the problem formulation, algorithms developed for the graph-data setting[27, 28, 29, 30, 31] cannot be used to solve the single-graph setting[12, 32].

In recent years, although few tools are developed for mining frequent subgraph on a graph, lots of methods are proposed to discovery network motif[33, 22, 34, 9], which is a well-known concept considered as the building blocks of network. The frequent subgraph mining problem includes two steps: (1) enumerate all subgraphs of a given size that occur in the input graph; (2) classify the enumerated subgraphs based on isomorphic test; (3) output the frequent subgraph satisfying the minimum support threshold. Since the isomorphism test is a NP problem, the current research on frequent subgraph mining on a graph focused on how to enumerate subgraphs efficiently and decrease the count of isomorphism test. In Kavosh, an efficient enumeration method was proposed. Comparing with three well-known methods Mfinder[23], MAVisto[35] and FANMOD[22], Kavosh is faster [21]. However, Kavosh did not decrease the amount of isomorphic test.

More recently, a Quaternary tree data structure was used to decrease the amount of isomorphism test in QuateXelero. The evaluation test showed that QuateXelero is more efficient than Kavosh[36]. In the Result section, we compare the performance of Kavosh and Quaternary, which are the state-of-art methods.

3 Methods

4 Results

4.1 Performance evaluation on general networks

Evaluate the methods on Ecoli network... et al. (used in the kavosh and QuateXelera paper) using different parameters (fix the subgraph size, change the frequency; fix the frequency, change the subgraph size). Compare the amount of calling nauty.

4.2 Performance evaluation on organelle networks

Evaluate the methods on organelle network... et al. (used in the kavosh and QuateXelera paper) using different parameters (fix the subgraph size, change the frequency; fix the frequency, change the subgraph size). Compare the amount of calling nauty.

4.3 Identified organelle communication patterns

Output the frequent subgraphs with high frequency and their instances; then, find the interesting communication patterns

5 Conclusion

References

[1] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[2] Maitreya Natu, Vaishali Sadaphal, Sangameshwar Patil, and Ankit Mehrotra. Mining frequent subgraphs to extract communication patterns in data-centres. In *Distributed Computing and Networking*, pages 239–250. Springer, 2011.

[3] Alessandro Vespignani et al. Predicting the behavior of techno-social systems. *Science*, 325(5939):425, 2009.

[4] Simon B Laughlin and Terrence J Sejnowski. Communication in neuronal networks. *Science*, 301(5641):1870–1874, 2003.

[5] Yuan Yuan Song, Ren Sen Zeng, Jian Feng Xu, Jun Li, Xiang Shen, and Woldemariam Gebrehiwot Yihdego. Interplant communication of tomato plants through underground common mycorrhizal networks. *PLoS One*, 5(10):e13324, 2010.

[6] Konstantin Voevodski, Shang-Hua Teng, and Yu Xia. Finding local communities in protein networks. *BMC bioinformatics*, 10(1):297, 2009.

[7] Ganesh Kumar Agrawal, Jacques Bourguignon, Norbert Rolland, Geneviève Ephritikhine, Myriam Ferro, Michel Jaquinod, Konstantinos G Alexiou, Thierry Chardot, Niranjan Chakraborty, Pascale Jolivet, et al. Plant organelle proteomics: collaborating for optimal cell function. *Mass spectrometry reviews*, 30(5):772–853, 2011.

[8] A.L. Barabási and Z.N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[9] Elisabeth Wong, Brittany Baur, Saad Quader, and Chun-Hsi Huang. Biological network motif detection: principles and practice. *Briefings in bioinformatics*, 13(2):202–215, 2012.

[10] Mohsen Bayati, David F Gleich, Amin Saberi, and Ying Wang. Message-passing algorithms for sparse network alignment. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(1):3, 2013.

[11] Jin Chen, Wynne Hsu, and Mong Li Lee et al. Nemofinder: dissecting genome wide protein protein interactions with meso scale network motifs. *SIGKDD*, pages 106–115, 2006.

[12] M. Kuramochi and G. Karypis. Finding frequent patterns in a large sparse graph*. *Data mining and knowledge discovery*, 11(3):243–271, 2005.

[13] S. Fortin. The graph isomorphism problem. *Department of Computing Science, University of Alberta*, 1996.

[14] JR Ullmann. An algorithm for subgraph isomorphism. *Journal of the ACM (JACM)*, 23(1):42, 1976.

[15] Bruno T. Messmer and Horst Bunke. Efficient subgraph isomorphism detection: A decomposition approach. *TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 12:307–324, 2000.

[16] M.R. Garey and D.S. Johnson. Computers and intractability: A guide to the theory of np completeness. *Freeman*, 1979.

[17] H. Fei and J. Huan. Structure feature selection for graph classification. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 991–1000. ACM, 2008.

[18] N. Jin, C. Young, and W. Wang. Graph classification based on pattern co-occurrence. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 573–582. ACM, 2009.

[19] R. Zhou and E.A. Hansen. Structured duplicate detection in external-memory graph search. In *Proceedings of the National Conference on Artificial Intelligence*, pages 683–689. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[20] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824, 2002.

[21] Zahra RM Kashani, Hayedeh Ahrabian, Elahe Elahi, Abbas Nowzari-Dalini, Elnaz S Ansari, Sahar Asadi, Shahin Mohammadi, Falk Schreiber, and Ali Masoudi-Nejad. Kavosh: a new algorithm for finding network motifs. *BMC bioinformatics*, 10(1):318, 2009.

[22] S. Wernicke and F. Rasche. FANMOD: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152, 2006.

[23] Nadav Kashtan, Shalev Itzkovitz, Ron Milo, and Uri Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, 2004.

[24] J. Grochow and M. Kellis. Network motif discovery using subgraph enumeration and symmetry-breaking. In *Research in Computational Molecular Biology*, pages 92–106. Springer, 2007.

[25] N. Alon, P. Dao, I. Hajirasouliha, F. Hormozdiari, and S.C. Sahinalp. Biomolecular network motif counting and discovery by color coding. *Bioinformatics*, 24(13):i241, 2008.

[26] C. Chen, X. Yan, F. Zhu, and J. Han. gapprox: Mining frequent approximate patterns from a massive network. In *icdm*, pages 445–450. IEEE Computer Society, 2007.

[27] A. Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. *Principles of Data Mining and Knowledge Discovery*, pages 13–23, 2000.

[28] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *icdm*, page 313. Published by the IEEE Computer Society, 2001.

[29] J. Huan, W. Wang, and J. Prins. Efficient mining of frequent subgraphs in the presence of isomorphism. 2003.

[30] Xifeng Yan Jiawei Han. gspan: Graph-based substructure pattern mining. *ICDM*, pages 633–642, 2002.

[31] Jianzhong Li Zhaonian Zou, Hong Gao. Discovering frequent subgraphs over uncertain graph databases under probabilistic semantics. *ICDM*, pages 706–715, 2010.

[32] R. Zou and L.B. Holder. Frequent subgraph mining on a single large graph using sampling techniques. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 171–178. ACM, 2010.

[33] Shmoolik Mangan and Uri Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.

[34] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Nemofinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 106–115. ACM, 2006.

[35] Falk Schreiber and Henning Schwöbbermeyer. Mavisto: a tool for the exploration of network motifs. *Bioinformatics*, 21(17):3572–3574, 2005.

[36] Sahand Khakabimamaghani, Iman Sharafuddin, Norbert Dichter, Ina Koch, and Ali Masoudi-Nejad. Quatexelero: an accelerated exact network motif detection algorithm. *PloS one*, 8(7):e68073, 2013.