

NemoProfile: Effective representation for network motif and their instances

Wooyoung Kim and Lynnette Haukap

Division of Computing and Software Systems
School of Science, Technology, Engineering, and Mathematics
University of Washington Bothell
18115 Campus Way NE, Bothell, WA 98011-8246, USA
{kimw6,lrd3}@uw.edu

Abstract. In systems biology, network motif is defined as a frequent and unique subgraph pattern in a network and has been applied in various biological and medical problems. However, identifying network motifs is intrinsically very costly, which restricts extensive and exhaustive experiments for real problems. Furthermore, most algorithms and tools provide limited information regarding network motifs, which necessitates the more expensive post-processing step to apply to real-world problems. Consequently, these problems bring a certain amount of skepticism about the relevance of network motifs in investigating real biological problems. Therefore, this paper introduces NemoProfile (network motif profile) as a new and effective network motif representation. Experimental results show that NemoProfile is generated almost effortlessly during motif-finding process, but it significantly reduces post-processing time for real applications. We also provide a case study to demonstrate the usability of NemoProfile for predicting essential proteins from protein-protein interaction networks.

Keywords: NemoProfile, NemoCollect, ESU, systems biology, biological network, network motif, essential protein

1 Introduction

Systems biology emphasizes on the interconnections of molecules, and produces biological networks where nodes are molecules and edges are interactions between them. Network motif analysis is one of the graph theory methods used to find biologically relevant functions in networks [8]. A network motif is defined as an overly frequent and unique subgraph pattern in a network, and it has been applied to solve various biological and medical problems [1, 2, 4–6, 12, 13, 15, 16, 18, 23–26].

In general, identifying network motifs is intrinsically very costly, and this high computational cost restricts extensive and exhaustive experiments for real problems. Consequently, various heuristic methods and parallel algorithms have been proposed that alleviate the performance concerns of exhaustive search methods

[3, 7, 9–11, 17, 19–21]. However, most motif-finding algorithms provide only frequency and statistical significance of each pattern, which restricts its usability for real-world problems. Therefore, we introduce a new network motif representation to overcome this problem, and define it as *NemoProfile*.

2 Method

NemoProfile relates each node to network motifs as a profile matrix while identifying network motifs. To show the effectiveness of *NemoProfile*, we also designed and implemented *NemoCollect* that collects instances of network motifs with *NemoProfile*. Figure 1 illustrates *NemoProfile* and *NemoCollect* process.

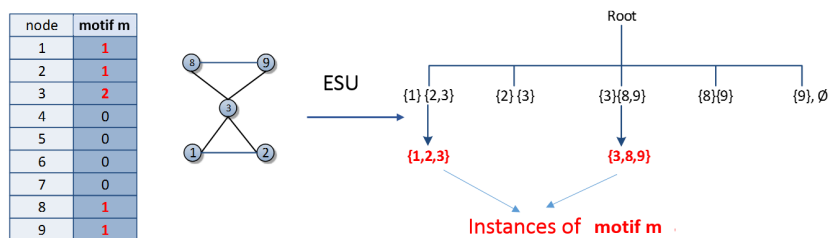


Fig. 1: *NemoProfile* (left) helps to effectively identify an induced subgraph for a “motif m ”. *NemoCollect* method induces subgraph with nodes 1, 2, 3, 8, and 9 will be processed to collect all the instances for a motif m .

Experimental results show that *NemoProfile* is efficiently generated but contains much richer information than ESU [22], and effectively collects network motifs. In addition, the test results on “*NemoCollect*” show that *NemoProfile* significantly alleviates the memory overhead problem when collecting network motif instances. We also demonstrated the usability of *NemoProfile* for real-world applications, specifically predicting essential proteins in a PPI network where network motif analysis has been applied previously [12, 13].

3 Conclusion and Future Study

Network motifs are structurally and statistically unique patterns in a network, and finding them involves computationally challenging tasks: subgraph enumeration, classification, and statistical validation processes. Various strategies have been introduced, leading to significant improvement of computational aspects. However, researchers still have skepticism about the biological relevance of the network motifs due to their lack of usability with current programs [14].

Therefore, we focused on relating their usability by introducing *NemoProfile* as an efficient and effective network motif representation, and demonstrated that

it is effortlessly generated but resolves the serious memory overhead problem while collecting network motif instances. Unlike other representations that only provide the frequencies and significance of subgraph patterns, NemoProfile is effectively reused to collect all the instances of network motifs, which can be directly applied to real-world problems.

To test the effectiveness of NemoProfile, we implemented a program and tested it with a number of PPI networks. Experimental results show that the construction of NemoProfile does not add extra computational costs, but it compactly represents network motifs as well as their instances. To show its effectiveness, we defined a process of collecting instances from NemoProfile as *NemoCollect* and tested its effective performance with real biological networks. We also demonstrated the usability of NemoProfile by providing a case-study that predicts essential proteins in PPI networks. The study demonstrates that NemoProfile can easily be converted to data feature space to apply various machine learning algorithms.

NemoProfile is a new representation. Therefore, future works include the following: (1) further improvement on the NemoCollect process by combining it with a symmetry breaking or mapping process; (2) design of a parallelized NemoCollect method that uses each individual column in NemoProfile, and; (3) design of a framework to apply NemoProfile to, both existing and new, real biological problems, thus significantly reducing the skepticism towards network motif analysis in the context of biology.

References

1. Albert, I., Albert, R.: Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics* 20(18), 3346–3352 (December 2004)
2. Callebaut, W.: Scientific perspectivism: A philosopher of science’s response to the challenge of big data biology. *Studies in history and philosophy of biological and biomedical sciences* 43(1), 69–80 (Mar 2012)
3. Chen, J., Hsu, W., Lee, M., Ng, S.: Nemofinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY pp. 106–115 (2006)
4. Chen, J., Hsu, W., Lee, M.L., Ng, S.K.: Labeling network motifs in protein interactomes for protein function prediction. *Data Engineering, International Conference on* 0, 546–555 (2007)
5. Conant, G.C., Wagner, A.: Convergent evolution of gene circuits. *Nature genetics* 34(3), 264–266 (July 2003)
6. Dobrin, R., Beg, Q.K., Barabasi, A.L., Oltvai, Z.N.: Aggregation of topological motifs in the escherichia coli transcriptional regulatory network. *BMC Bioinformatics* 5, 10 (2004)
7. Grochow, J.A., Kellis, M.: Network motif discovery using subgraph enumeration and symmetry-breaking. In: *Proceedings of the 11th annual international conference on Research in computational molecular biology*. pp. 92–106. RECOMB’07, Springer-Verlag, Berlin, Heidelberg (2007)
8. Junker, B.H., Schreiber, F.: *Analysis of Biological Networks*. Wiley (2008)

9. Kashani, Z., Ahrabian, H., Elahi, E., Nowzari-Dalini, A., Ansari, E., Asadi, S., Mohammadi, S., Schreiber, F., Masoudi-Nejad, A.: Kavosh: a new algorithm for finding network motifs. *BMC Bioinformatics* 10(1), 318 (2009)
10. Kashtan, N., Itzkovitz, S., Milo, R., Alon, U.: Efficient sampling algorithm for estimating sub-graph concentrations and detecting network motifs. *Bioinformatics* 20, 1746–1758 (2004)
11. Kim, W., Diko, M., Rawson, K.: Network motif detection: Algorithms, parallel and cloud computing, and related tools. *Tsinghua Science and Technology* 18(5), 469–489 (Oct 2013)
12. Kim, W.: Prediction of essential proteins using topological properties in go-pruned ppi network based on machine learning methods. *Tsinghua Science and Technology* 17(6), 645–658 (Dec 2012)
13. Kim, W., Li, M., Wang, J., Pan, Y.: Essential protein discovery based on network motif and gene ontology. In: *Proceedings of IEEE Bioinformatics and Biomedicine*. pp. 470–475 (2011)
14. Konagurthu, A.S., Lesk, A.M.: On the origin of distribution patterns of motifs in biological networks. *BMC Systems Biology* 2(1), 1–8 (2008)
15. Lee, W.P., Jeng, B.C., Pai, T.W., Tsai, C.P., Yu, C.Y., Tzou, W.S.: Differential evolutionary conservation of motif modes in the yeast protein interaction network. *BMC Genomics* 7(1), 89 (2006)
16. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., Alon, U.: Superfamilies of evolved and designed networks. *Science* 303(5663), 1538–1542 (March 05 2004)
17. Omidi, S., Schreiber, F., Masoudi-Nejad, A.: Moda: An efficient algorithm for network motif discovery in biological networks. *Genes and Genetic Systems* 84(5), 385–395 (2009)
18. Przulj, N., Corneil, D.G., Jurisica, I.: Modeling interactome: scale-free or geometric? *Bioinformatics* 20(18), 3508–3515 (2004)
19. Ribeiro, P., Silva, F., Kaiser, M.: Strategies for network motifs discovery. In: *e-Science, 2009. e-Science '09. Fifth IEEE International Conference on*. pp. 80–87 (2009)
20. Schreiber, F., Schwobbermeyer, H.: Mavisto: a tool for the exploration of network motifs. *Bioinformatics* 21, 3572–3574 (2005)
21. Wernicke, S., Rasche, F.: Fanmod: a tool for fast network motif detection. *Bioinformatics* 22, 1152–1153 (2006)
22. Wernicke, S.: Efficient detection of network motifs. *IEEE/ACM Trans.Comput.Biol.Bioinformatics* 3(4), 347–359 (2006)
23. Wuchty, S., Oltvai, Z.N., Barabasi, A.L.: Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature genetics* 35(2), 176–179 (print 2003)
24. Xie, Z.R., Hwang, M.J.: An interaction-motif-based scoring function for protein-ligand docking. *BMC Bioinformatics* 11(1), 298 (2010)
25. Zhang, L., King, O., Wong, S., Goldberg, D., Tong, A., Lesage, G., Andrews, B., Bussey, H., Boone, C., Roth, F.: Motifs, themes and thematic maps of an integrated *saccharomyces cerevisiae* interaction network. *Journal of Biology* 4(2), 6 (2005)
26. Zhang, Y., Xuan, J., de los Reyes, B.G., Clarke, R., Ransom, H.W.: Network motif-based identification of breast cancer susceptibility genes. In: *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. pp. 5696–5699. IEEE (aug 2008)