A quasi-clique mining method and its role in analysis of the human protein-protein interaction network and leather industry

Conference: 13th AICLST, Kolkata, 6–8 March 2026

Paper reference: TI-63/P

Printed pages: 444–448

Source: *Leather 2050: Challenges & Opportunities*, ILTA (2026)

A quasi-clique mining method and its role in analysis of the human protein-protein interaction network and leather industry

Brijesh Kumar Sriwastava1, Srirupa Dasgupta2, Kalyan Mahata3,

Falguni Sinhababu4, Ipsita Sengupta5

1,4Department of Computer Science and Engineering, 2,3Department of Information Technology, 5Department

Ref. No. TI-63/P

of Electronics and Communication

Government College of Engineering and Leather Technology, Kolkata-700106, India.

Abstract

In graph theory and data mining technique, detection of quasi-clique helps to find densely connected structures in a network. Protein-protein interactions (PPI) are at the core of the entire interaction system of any living cell. We can view a PPI network (PPIN) as a complex system of proteins which are linked by interactions among themselves. It can also be used to optimize leather supply chains by finding hidden collaboration clusters among suppliers like tanners, identifying risk communities, or mapping customer segments for targeted green initiatives. A huge complex graph is formed by raw material suppliers, tanneries, chemical suppliers (e.g., chromium sulphate), manufacturers, and retailers. These highly dependent and interconnected groups of suppliers and manufacturers can be revealed by quasi-cliques. It helps in risk management and development of alternative strategies. We can also form a graph of product preferences of customer and their interactions, purchase behaviour. Here finding quasiclique can identify hidden or overlapping communities of customers with similar, dense purchasing patterns. It helps to industry to adapt specific product lines like certain leather types, finishes and designs by specific quasi-cliques, optimizing production planning and reducing excessive inventory costs. In this paper, we present a novel heuristic approach to mine all possible maximal quasi-cliques (functionally associated

Corresponding author email: bsriwastava@gmail.com

proteins) from the complex and sparse human PPIN and identify significant modules of co-functional proteins. Information mining from the PPIN assists in rational drug design by targeting significant hub proteins in the network.

Keywords— Protein-protein interaction, leather supply chain optimization. network analysis, quasi-clique

1. Introduction

The Proteins perform biological functions by interacting with other proteins, small chemical compounds, metabolites, and RNA/DNA molecules. Interactions are completed by the formation of complexes, either transient or more permanent e.g., the replisome, RNA polymerases, spliceosome, ribosome, haperonins and various complexes formed along signal transduction pathways, or during enzyme catalysis and inhibition. Such interaction networks mediate biological processes. Protein-protein interactions (PPI) are at the core of the entire interaction system of any living cell, making them the central hubs or major mediators for virtually every bio-chemical process. Therefore, for a given protein, in order to understand its biological function, it is important to identify its likely interactions with other proteins. Detailed information of protein-protein interactions, metabolic and signal transduction networks improves our understanding of diseases, perturbation of healthy states or processes,

therefore providing the theoretical basis for new therapeutic approaches, mutant engineering and design, high throughput screening for drug design [1].

We can view a PPI network as a complex system of proteins which are linked by interactions among themselves. Wagner et al. [2] have modelled the PPI network as an undirected graph consisting of nodes and edges, where proteins are represented as nodes and the interaction of two proteins is represented as adjacent nodes

connected by an edge. The protein complexes in a PPI network are collection of proteins that interact with each other at a given time and place, forming dense regions in the PPI networks.

Protein-protein interaction networks (PPINs) cover a significant biological information at the molecular level of activity between protein pairs [3]. We get important directions towards the study of biological pathways, protein complexes, and protein function by mining such networks [4]. In order to unveil functional and evolutionary mechanisms behind biological networks, we need a better understanding of higher-order organization of proteins into overlapping complexes as most of the cellular processes are carried out by multi-protein complexes. So, analyzing such networks for exploring significant modules from the protein interactome is a very relevant research area in computational biology.

In order to discover the cluster of protein complexes in PPI networks, a number of techniques have been used by researchers. These include clustering of subgraph, finding dense regions [5-7], or clique finding [8]. Bader et al. [9] presented a molecular complex detection (MCODE) approach, that detects densely connected regions in large protein-protein interaction networks for molecular complexes. In this method, they weighted the vertex by local neighbourhood density and outward traversal from a locally dense seed protein to isolate the dense regions. Dongen et al. [10] proposed markov cluster algorithm (MCL) which converges toward a partitioning of the graph, with a set of high-flow regions (the clusters) separated by boundaries with no flow.

The current work is based on the clique property of a graph which is defined as a complete subgraph of a graph. The maximum Clique Problem (MCP), which is a wellknown combinatorial optimization problem, finds the largest complete subgraph of a graph. Bomze et al. and Du et al. [3, 4] provided detailed discussion regarding finding of cliques in a graph. Finding the maximum clique in a graph is known to be NP-complete [11].

Matsuda et al. [12] first introduced the concept of maximum quasi-clique problem (MQP), which can be thought as the largest constrained association of vertices in a graph, where constrained association of vertices referred to here is bounded by a threshold of minimum degree value for each vertex within the group. A quasiclique is an almost clique in a graph. In a quasi-clique, each vertex satisfies a minimum degree value with respect to the other vertices within the quasi-clique. The constraint of minimum degree in a quasi-clique , leads to a - quasi-clique.

Matsuda et al. [12] have used greedy approximation algorithm having time complexity of the algorithm is , where is the order of the graph. The approximation algorithm finds the minimum number of quasi-cliques but not the complete set of quasi-cliques. Later, Abello et al. [11], presented a greedy randomized adaptive search procedure (GRASP) for finding the set of quasi-cliques (for a given) in relatively larger graphs (having number of vertices in the order of). But, neither of these earlier works finds the complete set of quasi-cliques, nor addresses how to mine the largest quasi-clique. Subsequently Pie et al. [13] have proposed an efficient mining algorithm ( ) to find the complete set of quasi-cliques. The time complexity of this algorithm linearly grows with the number of quasi-cliques present in the graph. Subsequently an improved version of the Crochet algorithm, , with a similar goal was proposed [14]. However, these approaches have some limitations in finding the maximum quasi-clique. In, one has to sort the complete set of maximal quasi-cliques for getting the optimized solution. Brunato et al. [15], have proposed a new definition quasi-clique that associates two separate

parameters. It has defined a subgraph induced by the vertex subset as a quasi-clique if the minimum degree of all the vertices within is at least and the number of edges it contains is at least, where is the cardinality of the graph.

In view of the above facts, it is evident that there is an increasing necessity to localize those significant modules or impactful key protein clusters in an interaction network. Existing algorithms for determination of cliques are not sufficient to address the complexity of many networks. In case of homo-sapiens, simple analysis of tightly coupled cliques from the protein-protein networks may not be sufficient for investigation of key disease pathways. Therefore, we attempted to relax the constraints in an attempt to find all possible maximal quasi-cliques in the networks. The work presented here is found to be computationally efficient from the earlier works (DA)[16], Quick[17] and Cocain[18] and the experimental results validates our claims.

In this work, we attempt to search for the largest proteinprotein interaction cluster using a new quasi-clique algorithm (qCliP). In the following, the problem is formally defined accompanying the necessary preliminaries. Then, our proposed mining method is presented along with detailed pathway analysis.

Algorithm 3. □□□□□□□□□□□□□□□□□□(□)

1. if (□= ∅) then 2. Add each node □□ ∈□ in the singleton vector □□ of candi

dates list □□

3. else

4. //permute all candidate vectors, satisfying the property of □□

5. Compute □□= ⌈□× (□□□□□−1)⌉

6. for (∀ □□□□ □□ ∈(□−□),□□□□□□□□□≥2) do 7.

□□□= ∑□□□□□□□□□□□□□ ,□□ □,∀ □□ ∈□ □

8.

if (□□□≥1 and □□□≥□□) then

9.

create a new vector □□ in □□ by joining

vector □ and node □□ from (□−□)

10.

end if

11. end for 12. end if

13. return candidate List □□

3. Conclusion

In this paper, we have developed a heuristic approach to mine the largest quasi-clique from large and sparse scale free networks. This work is novel in its application in computational biology.

It can also be used to optimize leather supply chains by finding hidden collaboration clusters among suppliers like tanners, identifying risk communities, or mapping customer segments for targeted green initiatives. A huge complex graph is formed by raw material suppliers, tanneries, chemical suppliers (e.g., chromium sulphate), manufacturers, and retailers. These highly dependent and interconnected groups of suppliers and manufacturers can be revealed by quasi-cliques. It helps in risk management and development of alternative strategies. We can also form a graph of product preferences of customer and their interactions, purchase behaviour. Here finding quasiclique can identify hidden or overlapping communities of customers with similar, dense purchasing patterns. It helps to industry to adapt specific product lines like certain leather types, finishes and designs by specific quasicliques, optimizing production planning and reducing excessive inventory costs.

References

1. V. Chelliah, L. Chen, T. Blundell et al., “Distinguishing structural and functional restraints in evolution inorder to identify interaction sites,” Journal of Molecular Biology, vol. 342, pp. 1487–1504, 2004.

2. A. Wagner, “How the global structure of protein interaction networks evolves,” Proceeding of Royal Society London Ser. B Biological Science, vol. 270, pp. 457–466, 2003. 3. D. Du, and P. M. Pardalos, Handbook of combinatorial optimization: Springer, 1998.

4. I. M. Bomze, M. Budinich, P. M. Pardalos et al., “The maximum clique problem,” Handbook of combinatorial optimization, vol. 4, no. 1, pp. 1-74, 1999.

5. M. Altaf-Ul-Amin, Y. Shinbo K. Mihara et al., “Development and implementation of an algorithm for detection of proten complexes in large interaction networks,” BMC Bioinformatics, vol. 7, no. 1, pp. 207, 2006.

6. S. Brohee, and J. Van Helden, “Evaluation of clustering algorithms for protein-protein interaction networks,” BMC Bioinformatics, vol. 7, no. 1, pp. 488, 2006.

7. J. B. Pereira Leal, A. J. Enright, and C. A. Ouzounis, “Detection of functional modules from protein interaction networks,” PROTEINS: Structure, Function, and Bioinformatics, vol. 54, no. 1, pp. 49-57, 2003.

8. V. Spirin, and L. A. Mirny, “Protein complexes and functional modules in molecular networks,” Proceedings of National Academy of Science, U. S. A. , vol. 100(21), pp. 12123–12128, 2003.

9. G. D. Bader, and C. W. V. Hogue, “An automated method for finding molecular complexes in large protein interaction networks,” BMC Bioinformatics, vol. 4, no. 1, pp. 1471–2105, 2003.

10. S. Van Dongen, A new cluster algorithm for

graphs: Citeseer, 1998.

11. J. Abello, M. Resende, and S. Sudarsky, “Massive

quasi-clique detection,” LATIN 2002: Theoretical Informatics, pp. 598-612, 2002.

12. H. Matsuda, T. Ishihara, and A. Hashimoto, “Classifying

molecular sequences using a linkage graph with their pairwise similarities,” Theoretical Computer Science, vol. 210, no. 2, pp. 305-325, 1999

13. J. Pei, D. Jiang, and A. Zhang, “On mining cross-

graph quasi-cliques.” pp. 228-238.

14. D. Jiang, and J. Pei, “Mining frequent cross-graph

quasi-cliques,” ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 2, no. 4, pp. 16, 2009.

15. M. Brunato, H. Hoos, and R. Battiti, “On

effectively finding maximal quasi-cliques in graphs,” Learning and Intelligent Optimization, pp. 41-55, 2008.

16. M. Bhattacharyya, and S. Bandyopadhyay,

“Mining the largest quasi-clique in human protein interactome.” pp. 194-199.

17. G. Liu, and L. Wong, “Effective pruning

techniques for mining quasi-cliques,” Machine Learning and Knowledge Discovery in Databases, pp. 33-49: Springer, 2008.

18. Z. Zeng, J. Wang, L. Zhou et al., “Coherent closed

quasi-clique discovery from large dense graph databases.” pp. 797-802.

19. L. Salwinski, C. S. Miller, A. J. Smith et al., “

The Database of Interacting Proteins: 2004 update,” Nucleic Acids Research, vol. 32, pp. D449-D451, 2004.

20. B. K. Sriwastava, S. Basu, U. Maulik et al.,

“PPIcons: identification of protein-protein interaction sites in selected organisms,” Journal of Molecular Modeling, pp. 1-12, 2013.