Clique finding in *Tetraselmis* Subcordiformis: Final Project Report

Eli Spiliotopoulos
elispilio@reed.edu

ABSTRACT

In this project a predicted protein-protein interactome was seeded with transcriptome read data. This seeded data was taken and read through a maximal clique algorithm to find maximal cliques that included these known to be higher expressed genes. The goal of this project was to identify higher expressed gene groups through these more related groups of genes within the larger set.

Keywords

Cliques; *Tetraselmis Subcordiformis;* maximal cliques;

1. INTRODUCTION

The purpose of this project was to identify maximal cliques within the interactome of the marine algae *Tetraselmis Subcordiformis*. A maximal clique, the largest complete connected subgraph within a larger network, would act as an identifying factor for a large group of proteins that interact with each other for similar purposes. *Tetraselmis Subcordiformis*, besides having a protein-protein interactome available, is an important species due to its lipid profile, which makes it a good candidate for biofuel production[1]. In addition to its lipid content *Tetraselmis Subcordiformis* also has high starch content and productivity, which could be useful for low cost feed[2]. As such further exploration of the genome and proteome of this species is important, especially in light of computational methods possible on high throughput data. The predicted protein-protein interaction network used in this project is one example of this, created from a mixture of proteome and transcriptome data, and predicted from orthologous interaction in model organisms. The resulting interactome was created by a group of researchers at Dalian University of Technology[3]. This interactome, when combined with expression data, would allow the most highly expressed groups of genes to be determined. To be able to determine the cliques with the highly expressed genes, the input to a maximal cliques function was seeded with the 25 most highly expressed genes from a transcriptional study. After the function was seeded with the weighted nodes, cliques were found by running through unseen nodes and finding the intersection of nodes that are neighbors of each other. Once the

maximal cliques were found, the unigene ID's of significant cliques can be searched on the NCBI database to determine the nature of the cliques including highly expressed genes.

2. Methods

The protein-protein interactome data was first downloaded and read from the excel file into a csv document. The protein-protein interactome file was made up of confidence values, paper references to where the original orthologue was as well as other forms of identification that were not as consistent as Unigene ID's. This CSV file was then read into a list of edges and nodes for the graph, along with a list of nodes to be weighted. These nodes to be weighted were added to the beginning of a list of nodes, with the rest of the nodes being organized by degree. The maximal clique finding function was based on the one used in class, from the textbook "Analysis of Biological Networks" chapter 6.

Figure 1. The pseudocode of the maximal clique finding algorithm used. Image from Analysis of Biological Networks. Edited: Bjorn H Junker, Falk Schreiber, pub. 2007, DOI: 10.1002/9780470253489

This code is, in its most basic form, two while functions. The first one runs through every node that has not already been observed by the function. The while function adds to the queue from the unseen node and starts a new

---

[1] Huang, X., Huang, Z., Wen, W. et al. J Appl Phycol (2013) 25: 129. doi:10.1007/s10811-012-9846-9

[2] Changhong Yao, Jiangning Ai, Xupeng Cao, Song Xue, Wei Zhang, , Bioresource Technology, Volume 118, August 2012,

Pages 438-444, ISSN 0960-8524, http://dx.doi.org/10.1016/j.biortech.2012.05.030.

[3] Ji, C., Cao, X., Yao, C. et al. J Ind Microbiol Biotechnol (2014) 41: 1287. doi:10.1007/s10295-014-1462-z

cluster. Then the next while function begins and runs through the nodes that have been placed in the queue. The node being looked at than has its neighbors intersected with the members of the current set, and the unseen neighbors are added to the queue. The end result is a set that contains a clique of nodes, which is added to the larger list of cliques. When these while loops run their course, a list of sets is returned of maximal cliques in the graph, with the seeded nodes found first.

3. Results

The function resulted in a maximal clique that was greater than 300 nodes, as well as 8 cliques that were far less than that. After running the function a second time, a similar result was found, though the makeup of the cliques were slightly different. As a result of this, all of the nodes that were seeded as being highly expressed were either a part of the largest clique (hereby known as the 'extreme' node) or not part of a clique.
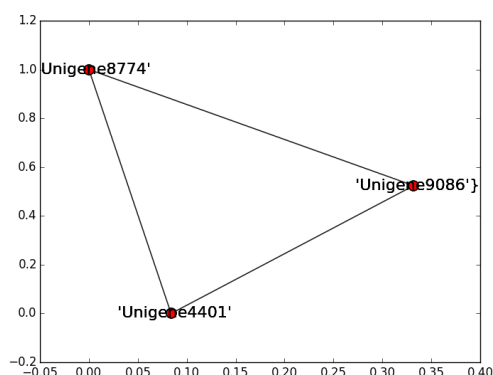

Figure 2. An example of one of the cliques found by the algorithm.

As the cliques found by the algorithm are maximal, the algorithm does not succeed at finding cliques related to the more highly expressed nodes. The more highly expressed nodes inside of the maximal clique were not found, only the entirely connected part. As a result, the smaller cliques, such as the one in Figure 2, may accurately represent the part of proteins of similar types. In the case of figure 2, the three proteins are Histone H3, putative modification tRNA to GTPase, and Pyrroline-5-carboxylate reductase, the three of which have different functions.

4. Discussion

Unfortunately, the results were not helpful in determining protein types from their cliques. Not only did the algorithm find few cliques outside of the extreme, it also

5. SECTIONS

The heading of a section should be in Times New Roman 12-point bold in all-capitals flush left with an additional 6-points of white space above the section head. Sections and subsequent subsections should be numbered and flush left. For a section head and a subsection head together (such as Section 3 and subsection 3.1), use no additional space above the subsection head.

5.1 Subsections

The heading of subsections should be in Times New Roman 12-point bold with only the initial letters capitalized. (Note: For subsections and subsubsections, a word like *the* or *a* is not capitalized unless it is the first word of the header.)

5.1.1 Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized and 6-points of white space above the subsubsection head.

5.1.1.1 Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

5.1.1.2 Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] Huang, X., Huang, Z., Wen, W. et al. *Journal of Applied Phycology*(2013) *Effects of Nitrogen Supplementation of the culture Medium on the Growth, Total Lipid Content and Fatty Acid Profiles of Three Microalgae (Tetraselmis Subcordiformis, Nannochloropsis Oculata and Pavlova Viridis)* in *Journal of Applied Phycology*(2013) 25: 129. doi:10.1007/s10811-012-9846-9

[2] Changhong Yao, Jiangning Ai, Xupeng Cao, Song Xue, Wei Zhang, *Enhancing starch production of a marine green microalga Tetraselmis subcordiformis through nutrient limitation*, Bioresource Technology, Volume 118, August 2012, Pages 438-444, ISSN 0960-8524, http://dx.doi.org/10.1016/j.biortech.2012.05.030.

[3] Analysis of Biological Networks. Edited: Bjorn H Junker, Falk Schreiber, pub. 2007, DOI: 10.1002/9780470253489

[4] Tao-Wei Huang, Chung-Yen Lin and Cheng-Yan Kao, *Reconstruction of Human Protein Interolog Network Using Evolutionary Conserved Network*, BMC Bioinformatics20078:152. DOI: 10.1186/1471-2105-8-152

[5] Yu, Y. T. and Lau, M. F. 2006. A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions. *J. Syst. Softw.* 79, 5 (May. 2006), 577-590. DOI= http://dx.doi.org/10.1016/j.jss.2005.05.030.

[6] Spector, A. Z. 1989. Achieving application requirements. In *Distributed Systems*, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33. DOI= http://doi.acm.org/10.1145/90417.90738.

Columns on Last Page Should Be Made As Close As Possible to Equal Length