EPiC Series in Computing

Volume 70, 2020, Pages 99–108

Proceedings of the 12th International Conference on Bioinformatics and Computational Biology

Clique Selection and its Effect on Paraclique Enrichment: An Experimental Study

Yuping Lu[1], Charles A. Phillips[2], Elissa J. Chesler[3] and Michael A. Langston[2]

[1] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[2] University of Tennessee, Knoxville, TN 37996, USA
[3] The Jackson Laboratory, Bar Harbor, ME 04609, USA
yupinglu89@gmail.com, cphill25@tennessee.edu, Elissa.Chesler@jax.org, langston@tennessee.edu

Abstract

The paraclique algorithm provides an effective means for biological data clustering. It satisfies the mathematical quest for density, while fulfilling the pragmatic need for noise abatement on real data. Given a finite, simple, edge-weighted and thresholded graph, the paraclique method first finds a maximum clique, then incorporates additional vertices in a controlled manner, and finally extracts the subgraph thereby defined. When more than one maximum clique is present, however, deciding which to employ is usually left unspecified. In practice, this frequently and quite naturally reduces to using the first maximum clique found. In this paper, maximum clique selection is studied in the context of well-annotated transcriptomic data, with ontological classification used as a proxy for cluster quality. Enrichment p-values are compared using maximum cliques chosen in a variety of ways. The most appealing and intuitive option is almost surely to start with the maximum clique having the highest average edge weight. Although there is of course no guarantee that such a strategy is any better than random choice, results derived from a large collection of experiments indicate that, in general, this approach produces a small but statistically significant improvement in overall cluster quality. Such an improvement, though modest, may be well worth pursuing in light of the time, expense and expertise often required to generate timely, high quality, high throughput biological data.

1 Introduction

Clustering is a core task in biological network analysis, whereby a cluster is typically defined as a dense subgraph extracted from high throughput omics data using some measure of pairwise similarity between genes, proteins, metabolites or other biological entities. Popular similarity metrics include Pearson's product-moment correlation, Spearman's and Kendall's rank correlations, and methods better

suited for handling nonlinear relationships such as mutual information. The highest possible density occurs when every vertex is connected to every other vertex. A subgraph with this property is called a clique. Although challenging to compute, clique is thus the gold standard of clustering.

An oft-used example is based on DNA microarray and gene co-expression analysis [1-3] in the context of the relevance network framework [4, 5]. In this classic setting, we begin with a complete graph whose vertices denote probe sets (gene surrogates), each of whose edges is assigned a weight equal to the similarity across all samples of the expression levels of its endpoints. Via thresholding [6] we next retain an edge if and only if its weight satisfies some predetermined lower bound, thereby producing an incomplete, unweighted graph on which advanced, scalable, clique-centric algorithms [7] can then be applied.

The biological fidelity provided by these powerful algorithms has previously been studied [8] and shown to be superior to that provided by wide assortments of well-known and popular but less computationally-intensive methods, including K-means Clustering, NNN (Nearest Neighbor Networks), SOM (Self-Organizing Maps) and WGCNA (Weighted Gene Co-expression Network Analysis). This has further motivated the use of clique-based techniques, as exemplified by the bottom-up approach originally called k-clique communities [9] (subsequently renamed clique percolation), and the more efficient top-down strategy known as paraclique first introduced in [10]. The paraclique algorithm in particular has found utility in numerous application domains, ranging from network science [11] to transportation planning [12] to anomaly detection [13]. In the health sciences alone, it has been employed in the study of lung cancer [14] and the exposome [15], as well as in transcriptomics [16], proteomics [17], epigenetics [18, 19], diabetes [20], allergic rhinitis [21], obesity [22], community-acquired pneumonia [23] and even in studying the impact of low dose ionizing radiation on living organisms [24].

2 The Paraclique Algorithm

A primary aim of the paraclique algorithm is to ameliorate the effects of noise, primarily by reducing type II errors (false negatives). It accomplishes this by expanding a maximum clique in a tightly controlled manner with non-clique vertices that are adjacent to most, but not necessarily all, elements of the clique. More precisely, let G denote a finite, simple, undirected, edge-weighted graph. A maximum clique M in G is a clique of largest size. A paraclique P consists of M augmented with non-clique vertices via a user-defined parameter, g. Any non-clique vertex missing no more than g edges to M is added to P. Once computed, P is removed from G, the next paraclique is sought in G-P, and the process is iterated until G is decomposed into a series of paracliques. We refer the reader to [25] for density analyses and a more formal description of the paraclique method. A major motivation for paraclique's augmentation strategy rests in the fact that clique-centric methods are highly sensitive to these missing edges, which may be lost due to noise, experimental data capture, the effects of thresholding, and a multitude of other factors dependent on the problem at hand.

The main feature of interest here is the selection criteria by which a maximum clique M is chosen for augmentation. This question may at first seem moot given the computational recalcitrance of finding even one maximum clique, a classic NP-complete problem [26]. But modern, practical algorithms now make it feasible not only to find a single maximum clique, but to enumerate all of them [27]. With such capability now at hand, we created a test suite of graphs to measure the significance and consistency of maximum clique selection on cluster quality. For these we retained original edge weights, employed the well-known Gene Ontology (GO) [28, 29] as a proxy for a ground truth, and performed enrichment analysis [30] to determine how likely a cluster's contents are to occur by mere chance alone. For each graph thus constructed, we compared paracliques expanded from a maximum clique with the highest average edge weight, from another with the lowest average edge weight, and from one chosen at

random. We note that, for a given graph, all maximum cliques have the same size, and thus a maximum clique with the highest (lowest) average edge weight will naturally also have the highest (lowest) total edge weight.

3 Source Data

We employed 28 Saccharomyces cerevisiae microarray expression datasets obtained from the Gene Expression Omnibus (GEO) [31-33]. S. cerevisiae is a sound choice [8], since it is one of the simplest and best-studied eukaryotic organisms, possessing numerous essential cellular processes analogous to those found in humans. The first column of Table 1, to follow, contains the GEO accession numbers for datasets used in this study. For each, we constructed 21 weighted graphs using Pearson's product-moment correlations, with thresholds set at uniform increments of 0.01 over the interval 0.70 to 0.90, which is a common range for threshold preference. This produced a total of 588 graphs ranging in size from 1893 to 9335 vertices. Densities ranged from roughly 0.09% to 25%, where we define density in the usual way as the number of edges present divided by the maximum number of edges possible.

4 Computational Milieu

On each graph we tested the three aforementioned maximum clique selection strategies. Because high performance computing was required for a study of this magnitude, we ran the paraclique algorithm using the ORNL CADES platform [34], and halted a run only if it failed to complete its task within 48 hours. All but 20 graphs were solved in this fashion, and these 20 were of course excluded from the analysis. Over the remaining 568 graphs, we then performed GO functional enrichment using the tools at DAVID [35] on the first paraclique produced in each of the 1704 resultant paraclique listings. To produce a single score for each paraclique, we extracted the enrichment p-value of its most significant GO term.

5 Experimental Results

In Table 1, we list results obtained for graphs constructed at a sample threshold 0.80. Often the choice between a highest, a lowest, and a randomly chosen maximum clique makes little difference in p-value. On the other hand, this difference can sometimes be quite large, as is seen for example in the case of GDS2267. Of these 28 graphs, 11 had a better p-value in the paraclique constructed using a highest weight maximum clique versus a lowest weight maximum clique, nine exhibited no difference, and in eight a maximum clique of lowest weight produced a paraclique with a better p-value than did a maximum clique of highest weight. Thus, the ratio 11/8=1.375 denotes a measure of how often a better p-value was obtained by choosing a highest versus a lowest weight maximum clique. If this ratio across all tests tends to be consistently greater than 1, then it may be viewed as a reliable indication that selecting a highest weight maximum clique generally produces more highly enriched paracliques, which may then result in improved average cluster quality.

Table 1: Experimental results obtained at a threshold of 0.80.

Dataset	Maximum Clique		Average Paraclique Edge Weights and Enrichment Scores					
	Size	Number	Highest	P-value	Lowest	P-value	Random	P-value
GDS344	87	6	0.9111	1.10E-49	0.9099	5.30E-50	0.9105	5.30E-50
GDS362	304	75184	0.9267	2.60E-09	0.9259	1.90E-10	0.9267	1.90E-10
GDS600	1736	40	0.9584	1.50E-06	0.9584	1.40E-06	0.9584	1.50E-06
GDS772	78	6	0.9134	2.70E-26	0.9118	2.70E-26	0.9118	2.70E-26
GDS777	87	15	0.9101	2.00E-08	0.9096	2.00E-08	0.9096	2.00E-08
GDS922	450	2160	0.9235	5.20E-11	0.9230	5.80E-11	0.9232	6.50E-11
GDS991	317	2468	0.9245	1.10E-95	0.9224	1.70E-85	0.9243	6.90E-97
GDS1013	269	19152	0.9127	3.30E-127	0.9112	6.90E-123	0.9123	3.30E-127
GDS1103	312	672	0.9293	8.10E-20	0.9283	8.10E-20	0.9290	9.40E-20
GDS1534	154	180	0.9140	3.40E-08	0.9133	1.20E-06	0.9137	3.40E-08
GDS1550	361	240	0.9469	2.60E-05	0.9459	2.50E-05	0.9464	2.60E-05
GDS1551	453	48	0.9408	5.30E-06	0.9405	4.80E-06	0.9405	4.80E-06
GDS1611	182	258	0.8847	3.90E-05	0.8839	3.70E-05	0.8845	3.70E-05
GDS1674	93	160	0.9102	8.00E-14	0.9078	1.40E-13	0.9090	1.40E-13
GDS2050	617	1152	0.9365	2.10E-32	0.9363	2.10E-32	0.9364	2.80E-32
GDS2079	1611	16	0.9563	8.30E-07	0.9563	8.30E-07	0.9563	4.50E-07
GDS2267	168	312	0.9058	7.50E-103	0.9035	3.00E-98	0.9058	2.60E-101
GDS2462	1351	13	0.9538	3.10E-46	0.9535	1.30E-43	0.9537	3.10E-46
GDS2508	49	11	0.9036	1.40E-03	0.8980	1.50E-03	0.9002	1.50E-03
GDS2522	428	13724	0.9321	1.40E-03	0.9313	1.50E-03	0.9318	2.10E-04
GDS2625	309	80	0.9191	3.00E-06	0.9187	2.80E-06	0.9189	2.80E-06
GDS2663	282	600	0.9283	5.80E-18	0.9269	4.40E-16	0.9273	4.40E-16
GDS2925	89	60	0.8940	1.10E-03	0.8930	3.80E-03	0.8934	4.40E-03
GDS2969	119	24	0.9161	1.80E-12	0.9143	1.80E-12	0.9148	1.80E-12
GDS3061	181	152	0.9218	2.80E-25	0.9198	2.80E-25	0.9208	2.80E-25
GDS3137	562	1088	0.9354	1.00E-04	0.9350	1.00E-04	0.9353	1.50E-04
GDS3198	383	2184	0.9333	3.50E-06	0.9327	1.70E-06	0.9331	2.80E-06
GDS3438	3424	2	0.9898	8.50E-11	0.9898	8.50E-11	0.9898	8.50E-11

5.1 Highest versus Lowest Weight Maximum Cliques

In Table 2, we summarize results comparing a highest weight paraclique to a lowest weight paraclique for all 21 thresholds under study. For each threshold, we list the number of graphs in which a highest weight maximum clique produced a lower p-value paraclique than did a lowest weight maximum clique, the number of graphs in which the reverse was true, the number of graphs in which the p-values were no different, and a ratio denoting the number of times highest weight was better to the number of times lowest weight was better. Overall, highest weight was better in 234 graphs, there was no difference in 177 graphs, and lowest weight was better in 157 graphs. Interestingly, the ratio was greater than one at all 21 thresholds, suggesting that it is generally beneficial to select a maximum clique of highest weight over one of lowest weight. Over the 1136 graphs tested, choosing a highest versus a lowest weight maximum clique resulted in improved cluster quality 1.490 times more often than it resulted in worse cluster quality.

To estimate statistical significance, we employed two binomial tests. The only difference between them was the probability of success. For the first test, shown in the last column of Table 2, we assumed an equal likelihood for each of three possible outcomes: a better, a worse , or an unchanged enrichment score. Because 234 outcomes were actually better, 157 were actually worse, and 177 turned out to be unchanged, this test produced a significant result, with $p = 0.0000163$. For the second test, we used the observed proportion of graphs for which there was no difference as an estimate of the proportion of "no difference" graphs in the population. This assumed that, for all other graphs, a paraclique constructed using a highest versus a lowest weight maximum clique had equal likelihood of producing a better p-value. This test also yielded a significant result, with $p = 0.000122$.

Table 2: Paraclique with highest weight maximum clique vs paraclique with lowest weight maximum clique.

Threshold	Highest Better	No Difference	Lowest Better	Highest Better / Lowest Better	Binomial P-value
0.70	16	6	4	4	2.14E-03
0.71	10	7	6	1.667	9.96E-02
0.72	10	4	8	1.25	8.44E-02
0.73	11	6	9	1.222	9.96E-02
0.74	13	8	6	2.167	4.31E-02
0.75	14	5	8	1.75	2.15E-02
0.76	11	8	8	1.375	1.12E-01
0.77	13	8	7	1.857	5.36E-02
0.78	15	7	6	2.5	1.34E-02
0.79	9	10	8	1.125	1.61E-01
0.80	11	9	8	1.375	1.23E-01
0.81	12	7	8	1.5	7.47E-02
0.82	12	8	8	1.5	8.72E-02
0.83	9	12	7	1.286	1.58E-01
0.84	10	9	9	1.111	1.50E-01
0.85	11	8	9	1.222	1.23E-01
0.86	11	8	9	1.222	1.23E-01
0.87	8	13	7	1.143	1.42E-01

0.88	10	10	8	1.25	1.50E-01
0.89	10	10	8	1.25	1.50E-01
0.90	8	14	6	1.333	1.42E-01
Total	234	177	157	1.490	1.63E-05

5.2 Highest versus Random Weight Maximum Cliques

In Table 3, we list the results of testing whether choosing a highest weight maximum clique may be superior to choosing an arbitrary maximum clique, a process we simulated by selecting a maximum clique at random from among all maximum cliques enumerated. Once again, all ratios in the penultimate column are greater than or equal to one, and so we conclude that choosing a highest weight maximum clique tends to be wiser than merely making an arbitrary choice. Overall, the highest weight was better in 216 graphs, there was no difference in 225 graphs, and a random choice was better in 127 graphs. At first these differences may not appear as striking as did the differences between using a highest versus a lowest maximum clique. For example, the number of graphs for which there was no difference is noticeably larger in Table 3 than it was in Table 2. On the other hand, choosing a highest weight maximum clique resulted in improved cluster quality 1.701 times more often than it resulted in worse cluster quality, which is a slightly higher ratio than that computed from Table 2. Moreover, repeating the two binomial tests just described, we obtained significant results for both, with $p = 0.00219$ and $p = 0.0000124$, respectively.

Table 3: Paraclique with highest weight maximum clique vs paraclique with random maximum clique.

Threshold	Highest Better	No Difference	Random Better	Highest Better / Random Better	Binomial P-value
0.70	17	3	6	2.833	6.29E-04
0.71	9	12	2	4.5	1.42E-01
0.72	8	6	8	1	1.67E-01
0.73	10	8	8	1.25	1.37E-01
0.74	11	12	4	2.75	1.12E-01
0.75	11	6	10	1.1	1.12E-01
0.76	13	9	5	2.6	4.31E-02
0.77	15	11	2	7.5	1.34E-02
0.78	11	10	7	1.571	1.23E-01
0.79	11	11	5	2.2	1.12E-01
0.80	9	10	9	1	1.58E-01
0.81	9	11	7	1.286	1.61E-01
0.82	10	11	7	1.429	1.50E-01
0.83	8	14	6	1.333	1.42E-01
0.84	12	12	4	3	8.72E-02
0.85	9	12	7	1.286	1.58E-01
0.86	7	14	7	1	1.09E-01
0.87	11	12	5	2.2	1.23E-01

0.88	9	13	6	1.5	1.58E-01
0.89	9	13	6	1.5	1.58E-01
0.90	7	15	6	1.167	1.09E-01
Total	216	225	127	1.701	2.19E-03

5.3 Random versus Lowest Weight Maximum Cliques

Lastly, we used the same approach to compare paracliques constructed using random versus lowest weight maximum cliques. A random choice was better in 191 graphs, there was no difference in 215 graphs, and a lowest choice was better in 162 graphs. Although the aforementioned ratio was still above one (at 1.179), neither binomial test reached the level of significance, with $p = 0.035$ and $p = 0.015$, respectively.

6 Discussion

As can be seen in Table 1, there is sometimes little difference in enrichment p-values. And indeed, as can be seen in Tables 2 and 3, there are instances for which the choice makes no difference at all. Close scrutiny reveals that this is usually due to significant overlap between maximum cliques. In GDS344, for example, it turns out that 84 (of 87) vertices appear in all maximum cliques at a threshold of 0.8. We also note that the number of maximum cliques can vary greatly between datasets, and even between graphs constructed at different thresholds from the same dataset. In Table 1, for instance, we witnessed from 2 to 75184 maximum cliques at a single threshold. And GDS2925 had but one maximum clique when thresholded at 0.89, but 95044 when thresholded at 0.74.

These issues are relevant because large numbers of maximum cliques can dramatically increase computational costs. Thus, we tested only the first paraclique produced under each criterion, else time requirements quickly become prohibitive. To see this, note that not only is clique extraction an expensive operation in its own right, but a sample graph with, say, 100 different maximum cliques will yield 100 different first paracliques that, once deleted, leave a set of 100 new graphs, each of which may again have 100 different maximum cliques, paracliques and so on ad infinitum.

7 Conclusions and Directions for Future Research

In summary, these comprehensive tests provide convincing evidence that selecting a highest weight maximum clique tends to produce more functionally enriched paracliques than does choosing either a lowest weight or an arbitrary maximum clique. While this seems rather intuitive and to be expected, the effect size has been small, and so a large number of graphs has been required to confirm this relationship. Across Tables 2 and 3, for example, only two thresholds are significant at $p = 0.01$. Every other result, when analyzed alone, is non-significant. It is therefore only when results at many thresholds are combined that we reach a large enough sample size for the maximum clique choice to meet the standards of statistical significance.

Future research directions beckon, motivated in no small part by the significance of incremental improvements in solution quality given the enormous costs and staggering delays frequently encountered in producing high quality data. This general line of work could be applied, for example, to proteomic, metabolomic, epigenetic and other sorts of high throughput biological data to which the paraclique algorithm has already shown utility. This study might also be extended to more than just the

first paraclique distilled from each graph. Similarly, a variety of alternatives to Pearson product-moment correlation could be tested.

Acknowledgments

This research has been supported in part by the National Institutes of Health under grant R01AA018776 and by the Environmental Protection Agency under grant G17D112354237.

References

[1] E. Alm, and A. P. Arkin, "Biological networks," *Current Opinion in Structural Biology,* vol. 13, no. 2, pp. 193-202, Apr, 2003.

[2] A. J. M. Walhout, "Gene-centered regulatory network mapping," *Caenorhabditis Elegans: Molecular Genetics and Development, Second Edition,* vol. 106, pp. 271-288, 2011.

[3] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science,* vol. 302, no. 5643, pp. 249-255, Oct 10, 2003.

[4] C. J. Wolfe, I. S. Kohane, and A. J. Butte, "Systematic Survey Reveals General Applicability of "Guilt-by-Association" within Gene Coexpression Networks," *BMC Bioinformatics,* vol. 6, no. 227, 2005.

[5] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane, "Discovering Functional Relationships between RNA Expression and Chemotherapeutic Susceptibility using Relevance Networks," *Proc Natl Acad Sci U S A,* vol. 97, no. 22, pp. 12182-6, Oct 24, 2000.

[6] A. D. Perkins, and M. A. Langston, "Threshold selection in gene co-expression networks using spectral graph theory techniques," *BMC Bioinformatics,* vol. 10, 2009.

[7] E. Tomita, Y. Sutani, T. Higashi, S. Takahashi, and M. Wakatsuki, " A Simple and Faster Branch-and-Bound Algorithm for Finding a Maximum Clique," *Lecture Notes in Computer Science,* vol. 5942, pp. 191–203, 2010.

[8] J. J. Jay, J. D. Eblen, Y. Zhang, M. Benson, A. D. Perkins, A. M. Saxton, B. H. Voy, E. J. Chesler, and M. A. Langston, "A Systematic Comparison of Genome Scale Clustering Algorithms," *BMC Bioinformatics,* vol. 13 no. Suppl 10, pp. S7, 2012.

[9] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society," *Nature,* vol. 435, no. 7043, pp. 814-818, 2005.

[10] E. J. Chesler, and M. A. Langston, "Combinatorial Genetic Regulatory Network Analysis Tools for High Throughput Transcriptomic Data," *Systems Biology and Regulatory Genomics*, E. Eskin, ed., pp. 150–165: Springer, 2006.

[11] R. D. Hagan, C. A. Phillips, B. J. Rhodes, and M. A. Langston, "Compound Analytics: Templates for Integrating Graph Algorithms and Machine Learning," in Proceedings, Workshop at the Intersection of Graph Algorithms and Machine Learning, Orlando, Florida, 2017, pp. 1550-1556.

[12] R. D. Hagan, C. A. Phillips, M. A. Langston, and B. J. Rhodes, "Multiscale Graph Theoretical Tools Reveal Subtle Patterns in Big Geospatial Data."

[13] R. D. Hagan, C. A. Phillips, M. A. Langston, and B. J. Rhodes, "Classification and Anomaly Detection in Traffic Patterns of New York City Taxis: A Case Study in Compound Analytics," in Proceedings, Workshop at the Intersection of Graph Algorithms and Machine Learning, Vancouver, British Columbia, Canada, 2018.

[14] P. D. Juarez, D. B. Hood, G. L. Rogers, S. H. Baktash, A. M. Saxton, P. Matthews-Juarez, W. Im, M. P. Cifuentes, C. A. Phillips, M. Y. Lichtveld, and M. A. Langston, "A novel approach to

analyzing lung cancer mortality disparities: Using the exposome and a graph theoretical toolchain," *Environmental Disease,* vol. 2, pp. 33-44, 2017.

[15] M. A. Langston, R. S. Levine, B. J. Kilbourne, G. L. Rogers, A. D. Kershenbaum, S. H. Baktash, S. S. Coughlin, A. M. Saxton, V. A. Agboto, D. B. Hood, M. Y. Litchveld, T. J. Oyana, P. Matthews-Juarez, and P. D. Juarez, "Scalable Combinatorial Tools for Health Disparities Research," *International Journal of Environmental Research and Public Health,* vol. 11, no. 10, pp. 10419-10443, 2014.

[16] M. A. Langston, A. D. Perkins, A. M. Saxton, J. A. Scharff, and B. H. Voy, "Innovative computational methods for transcriptomic data analysis: A case study in the use of FPT for practical algorithm design and implementation," *The Computer Journal,* vol. 51, pp. 26-38, 2008.

[17] A. Schoenrock, B. Samanfar, S. Pitre1, M. Hooshyar, K. Jin, C. A. Phillips, H. Wang, S. Phanse, K. Omidi, Y. Gui2, M. Alamgir, A. Wong, F. Barrenäs, M. Babu, M. Benson, M. A. Langston, J. R. Green, F. Dehne, and A. Golshani, "Efficient Prediction of Human Protein-Protein Interactions at a Global Scale," *BMC Bioinformatics,* vol. 15, no. 383, pp. DOI: 10.1186/s12859-014-0383-1, 2014.

[18] D. Macartney-Coxson, M. C. Benton, R. Blick, R. S. Stubbs, R. D. Hagan, and M. A. Langston, "Genome-Wide DNAMethylation Analysis Reveals Loci that Distinguish Different Types of Adipose Tissue in Obese Individuals," *Clinical Epigenetics,* vol. 9, no. 48, pp. DOI 10.1186/s13148-017-0344-4, 2017.

[19] C. E. Nestor, F. Barrenäs, H. Wang, A. Lentini, H. Zhang, S. Bruhn, R. Jörnsten, M. A. Langston, G. L. Rogers, M. Gustafsson, and M. Benson, "DNA Methylation Changes Separate Allergic Patients from Healthy Controls and May Reflect Altered CD4+ T-cell Population Structure," *PLoS Genetics,* vol. 10, pp. e1004059, 2014.

[20] J. D. Eblen, I. C. Gerling, A. M. Saxton, J. Wu, J. R. Snoddy, and M. A. Langston, "Graph Algorithms for Integrated Biological Analysis, with Applications to Type 1 Diabetes Data," *Clustering Challenges in Biological Networks*, W. A. Chaovalitwongse, ed., pp. 207-222: World Scientific, 2009.

[21] S. Bruhn, F. Barrenäs, R. Mobini, B. A. Andersson, S. Chavali, B. S. Egan, E. Hovig, G. K. Sandve, M. A. Langston, G. Rogers, H. Wang, and M. Benson, "Increased Expression of IRF4 and ETS1 in CD4+ Cells from Patients with Intermittent Allergic Rhinitis," *Allergy,* vol. 67, pp. 33-40, 2012.

[22] L. S. Gittner, B. J. Kilbourne, R. Vadapalli, H. M. K. Khan, and M. A. Langston, "A Multifactorial Obesity Model Developed from Nationwide Public Health Exposome Data and Modern Computational Analyses," *Obesity Research & Clinical Practice,* vol. 11, no. 5, pp. 522-533, 2017.

[23] O. M. Peck-Palmer, G. Clermont, G. L. Rogers, S. Yende, D. C. Angus, and M. A. Langston, "Graph Theoretical Analysis of Genome-Scale Data: Examination of Gene Activation Occurring in the Setting of Community-Acquired Pneumonia," *Shock: Injury, Inflammation, and Sepsis: Laboratory and Clinical Approaches,* vol. 50, pp. 53-59, 2018.

[24] B. H. Voy, J. A. Scharff, A. D. Perkins, A. M. Saxton, B. Borate, E. J. Chesler, L. K. Branstetter, and M. A. Langston, "Extracting Gene Networks for Low Dose Radiation using Graph Theoretical Algorithms," *PLoS Computational Biology,* vol. 2, no. 7, pp. e89, 2006.

[25] R. D. Hagan, M. A. Langston, and K. Wang, "Lower bounds on paraclique density," *Discrete Applied Mathematics,* vol. 204, pp. 208-212, 5/11/, 2016.

[26] M. R. Garey, and D. S. Johnson, *Computers and intractability: a guide to the theory of NP-completeness*: W. H. Freeman and Company, 1979.

[27] J. D. Eblen, C. A. Phillips, G. L. Rogers, and M. A. Langston, "The maximum clique enumeration problem: algorithms, applications, and implementations," *BMC Bioinformatics,* vol. 13 Suppl 10, pp. S5, Jun 25, 2012.

[28] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, and G. O. Consortium, "Gene Ontology: tool for the unification of biology," *Nature Genetics,* vol. 25, no. 1, pp. 25-29, May, 2000.

[29] G. Antonazzo, H. Attrill, N. H. Brown, S. J. Marygold, P. McQuilton, L. Ponting, G. H. Millburn, A. Rey, R. Stefancsik, S. Tweedie, M. Harris, J. Hayles, S. G. Oliver, K. Rutherford, and V. Wood, *Expansion of the Gene Ontology knowledgebase and resources*, 2017.

[30] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, no. 43, pp. 15545-15550, Oct 25, 2005.

[31] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research,* vol. 30, no. 1, pp. 207-210, Jan 1, 2002.

[32] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. G. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets-update," *Nucleic Acids Research,* vol. 41, no. D1, pp. D991-D995, Jan, 2013.

[33] NCBI. "Gene Expression Omnibus," https://www.ncbi.nlm.nih.gov/geo/.

[34] O. R. N. Laboratory. "CADES – Compute and Data Environment for Science "; https://www.olcf.ornl.gov/olcf-resources/rd-project/cades-compute-and-data-environment-for-science/.

[35] D. W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki, "DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists," *Nucleic Acids Research,* vol. 35, pp. W169-W175, Jul, 2007.