Medicon Medical Sciences
Volume 1 Issue 1 August 2021
Review Article

Note on Pertinence of Graph Theory for Genetic Networks

V. Yegnanarayanan*

Department of Mathematics, Kalasalingam Academy for Research and Education, Deemed to be University, Krishnankoil-626126, India

***Corresponding Author:** V. Yegnanarayanan, Department of Mathematics, Kalasalingam Academy for Research and Education, Deemed to be University, Krishnankoil-626126, India.

Received: July 17, 2021; **Published:** August 01, 2021

Abstract

Data collection process has witnessed explosive growth that could even outsmart the measurable gains in the field of power of computation as speculated by Moore's Law. However, scientists spread over various domain are smearing victory through network models to denote their data. Graph- algorithms come in handy to investigate the topological structure and bring out the hidden relationships and attributes among variables. But the task of finding dense subgraphs, are still hard to negotiate from a computational viewpoint. So, it is stimulating to find new algorithmic methods in the probe of graphs derived from huge data sources that are complex. We provide a bird's eye view of application of graph theoretic tools crisply for the probe of complex networks based on gene expression signatures and to find exact and effective solutions to hard combinatorial tasks such as epigenetic biomarker detection. To make an efficient design which is algorithm based, a probe of two core components of fixed parameter tractability namely branching and kernelization is suggested for parallel imposition of vertex cover. We have also discussed the concept of Para clique and domination concept to better handle gene expression networks.

Keywords: Genes; Genetic Networks; Data Types; Graphs; Clique Domination

Introduction

Lately we have seen a tremendous development in the volume of raw data thrown for researchers. Access to such data gives wonderful chance for new findings, when the dataset's size of interest paves way for ever evolving challenges that are computational. Although rise in computing power still follows Moore's Law and becomes two times every two years, the gains are being nullified and, in several cases, lags behind by a corresponding increase in data size. It was predicted in 2013 that almost entire data available in the universe had been generated in the last two years alone [5]. Further, the probe of networks like biological, and communication networks has become an all present across the sciences domain. For instance, network replicas were used in the description of gene networks that are putative [1, 32], the probe of protein-protein interactions [26], watching the spread across social networks [19], and analyzing the organization of the human brain [30]. Pertinent to the comprehension of such networks are their structures that are topological. As such, the probe of complex networks is closely interconnected with graph theory and algorithms. In the probe of networks, we must deal with two difficulties. First, the non-decreasing rate of growth of data sets. The second, is that the problems to solve are difficult from a computational viewpoint. For instance, the problem of determining the largest fully connected subgraph also called clique belongs to class of NP-complete problems. Overcoming these challenges require the invention of novel algorithms such as fixed Parameter tractability (FPT) and the scalable solutions that are parallel to be developed. They are the two key components of FPT namely vertex cover-kernelization and branching.

Various Gene Expression Types

Modern DNA microarrays history goes back to Grunstein's colony centric hybridization method and Hogness tactics [11]. In the past four decades, microarray technologies assumed a pivotal role in biological research. They measure gene expression levels through chips with hybridization analysis meant to target and bond to a particular mRNA sequence.

Technology has improved to the platforms such as the Affymetrix Exon 1.0 ST array which can handle exon level resolution of expression with 1.4 million probe sets approximately consisting of over 5×10^5 individual probes. For an excellent literature on types of DNA microarrays and applications one can see [3]. When the Human Genome Project undertook its mission in 1990 to analyze epigenetic data and to map the whole sequence of DNA due to human genome, it gave the hope of changing our perception of biology. It was even said explicitly in 2001 in one episode of NOVA on PBS [20]. While witnessing a giant leap ahead in our basic knowhow of genetics, it is evident that since its consummation way back in 2003 that several mechanisms contribute in the true gene expression that has far reaching consequences which go beyond the genetic code's physical arrangement. Such findings have led to epigenetics a new branch of research. This was coined by Robin Holiday in 1990 [15]. Lately there is no consensus over whether to agree the concept of an epigenetic trait and deem it as heritable [29]. Several epigenetic variation mechanisms were found. Two major of them are DNA methylation and histone modification. The former happens upon adding methyl group at the cytosine ring's 5' position, converting it to 5-methylcytosine. This happens at CpG dinucleotides, even though non-CpG methylation has been witnessed to happen quite often in particular contexts like in embryonic stem cells and neural development [25]. It seems to be controlled by DNA methyltransferases that comprises DNMT1, DMNT3a, and DMNT3b. DNMT1 works to upkeep methylation patterns by copying them to the unmethylated daughter strands when DNA replication happens. DNMT3a and DMNT3b are considered to be cause for de novo methylation events. Mutations in the DNMT3b gene are responsible for ICF (Immunodeficiency, centromeric instability, facial anomalies) syndrome [13], when mutations to any of DNMT1, DNMT3a, or DNMT3b were determined to be embryonically lethal in mice [23, 22]. In humans, about seventy percent of CpG dinucleotides are methylated [31]. Observe that there is a heavy concentration of CpG content in genomic regions and the same can be found in the genes' promoter regions. Cytosines lying in CpG intense regions, called CpG islands, tend to be unmethylated except in the inactive X chromosome [36] and genes that are imprinted [27, 2]. Aberrant methylation patterns can be seen in many diseases. Specifically, these patterns assume a dual role in several types of cancer through either a pattern of global hypomethylation permitting aberrant overexpression and/or guaranteeing oncogenesis, along with CpG islands' hypermethylation in the sponsor regions of tumor suppressor genes, paving the way for silencing [21, 14, 7]. Histone acetylation happens if an acetyl group is introduced to the NH3+ group on Lysine. This process occurs at the N-terminal of histone tails. Note that Acetylation and deacetylation are usually catalyzed by either histone deacetylase (HDAC) or histone acetyltransferase (HAT) enzymes, respectively. Acetylation acts to convert the full positive charge of the histone tail to neutral, thereby weakening the concentration of the nucleosomal components and allowing the DNA more reachable to transcriptional agents. Hence hyperacetylation is in positive correlation with transcribed genes that are active. It is possible to methylate the histone tails' lysine and arginine residues, but it is observed mostly on the lysine residues of the H3 and H4 tails. Observe that Lysine can be mono-, di-, or trimethylated through methyl group by using a hydrogen of its NH3+ group. Similarly, Arginine can be mono- or demethylated as it has a free NH2 and NH2+ group. Arginine's demethylation can happen either as a single group, or through asymmetric methylation of every group. Even-though DNA methylation and histone modifications are well analyzed regarding epigenetic mechanisms, a large quanta of current research is also about post-translational changes such as ubiquitination, phosphorylation and somoylation. As of now, they can be tested on virtually any type of data that permits such a replica. The only constraint is that we should be able to point some entities that can be denoted by vertices and become able to calculate some familiar metric among them. Data can be either considered, or categorical.

Graphs and Algorithms

Langston Lab started using graph algorithms to analyze health related data [21]. To study massive networks, interactions can be prototyped as a graph so that graph algorithms come in handy to sharpen our comprehension of the relationships that are latent. Vertices can be used to denote entities of interest, whether they are locations in a transportation network, proteins in a protein-protein interaction network, or people in a social network or genes in a gene network. To create a prototype about the relationships among the actors pointed by the vertices, one must have certain concept regarding how to put edges among them. In certain scenario such as social networks, it is easy to link relevant vertices by introducing edges among them if the entity they denote are familiarity. Generally we require a mathematical notion regarding a similarity measure like correlation coefficient. In the case of gene expression networks, we normally construct graphs with vertices denoting individual genes. Karl Pearson correlation coefficients are then found for every pair of genes considered across the relevant measures for every lot.

An edge is then introduced linking the vertices If the correlation value goes above certain threshold value. There are a plenty of approaches that can be put in place to choose an apt threshold value to put in use. Mostly we go by techniques depending on experimentation. For instance, while constructing graphs for a paraclique probe like in [35, 4], one may develop a series of graphs that slowly witness reduction in the value of the threshold until an inflection point is faced in the number of paracliques generated. Alternative way is to involve the spectral method [24], which exploits the eigen structure of the associated adjacency matrix.

In a formal manner a graph G (V, E) comprises a vertex set V and an edge set E holding ordered pairs of vertices from V such that $(u, v) \in E$ points to an edge among u and v. if it includes no self-loops or multiple edges then it is called simple. A graph is called a weighted graph by allotting weight either to its vertices, its edges, or both. u and v are said to be adjacent if $(u, v) \in E$. The order of G (V, E) is its number of vertices, $|V|$. By degree of v we mean the number of vertices to which v is adjacent, or the number of edges with v as an endpoint. The neighborhood of a vertex v in G, denoted by (v) or N(v), when N(v), is the set of vertices that are adjacent to v. A subgraph of a graph (V, E) is a graph G' (V', E') such that $V' \subseteq V$ and $E' \subseteq E$, with the condition that if $(u', v') \in E'$, both u' and $v' \in V'$. A subset of vertices involved to identify a subgraph of G called an induced subgraph. An induced subgraph comprises the inducing vertices as its vertex set, and all the edges among those vertices that were present in G. The complement of a graph G, denoted G^c, is formed by omitting the existing edges, and adding those that were not previously present.

Clique Task and the Paraclique Algorithm

A clique of a graph is a wholly connected subgraph. It is denoted by K_n. The process of finding cliques has variety of applications, right from locating putative gene pathways [33] to finding trading patterns that are collusive in the stock market [34]. The largest clique problem is the most probed in graph theory. The decision task, namely deciding whether a graph includes a clique of a given size, is one among original 21 NP-complete problems given in Karp's first work [18]. Actually aiming for tracking cliques proves too complicated. Presence of noise in real data leads to missing edges. As the loss of even one edge decimates a clique, such noise produces higher number of false negatives when probing the data for signal. To nullify the effects of noise, the paraclique algorithm was suggested in [4]. The fundamental notion is that we begin with a largest clique and develop it to a paraclique by repeatedly adding in vertices that are adjacent to all but certain permissible number of vertices available at the given stage. The number of edges permitted to be missed at each step is controlled by a constant named the glom term. The paraclique algorithm were found to be highly effective and found to outsmart other clustering methods with reference to ontological enrichment and density [16]. Even though the algorithm has focused on its application, theoretical work can be seen in [4, 12].

Graph Domination concepts

A vertex-dominating set is a subset S of the V in G(V, E) such that for every v in V, either v is in S or at least one of its neighbors is in S. Specifically, consider a Red-Blue Dominating Set, in which the vertices are colored either red or blue and we search for the smallest set of red vertices that dominate all the blue vertices or the other way. Recently, fixed parameter tractability-FPT has been considered as an effective way to design and implement practical algorithms for finding solutions to tasks deemed intractable. The origins of FPT trace goes back at least to the basic work of Fellows and Langston in the area of polynomial time computability [9, 8, 10]. Concurrently, Robertson and Seymour established the graph minor theorem, pointing that undirected graphs are well-quasi-ordered under the graph minor relationship [28, 6]. One aim of FPT is to give a more fine-grained way to classify the difficulty than that considered in classical complexity theory. The main notion is that if there is some parameter k in the task that, when held fixed, will allow the task to be solved in time with any super-exponential dependence being only in k, then the problem is in the class FPT.

There are two variations to FPT namely kernelization and branching. Former points to a process of decreasing the task size of a kernel to depend only on the fixed parameter k. It is established that a task being kernelizable is equivalent to it being FPT. Once the task is reduced to a compute-intensive kernel, it only remains to probe the reduced search space that gets emanated. Several branching techniques attempt investigate this space as effectively as possible. Krnelization is sought to reduce the impact of the original task size in algorithm runtimes and improved branching is sought to reduce the exponential dependence on the size of k.

Conclusion and Scope for further Research

We have analyzed both the application and efficiency of graph-theoretical algorithms to the probe of complex networks. We mentioned the need of exact solvers for two NP- complete problems namely dominating set and clique, and their advantage in finding solution to hard biological problems. This warrants new methods and means to development of algorithms, as well as its adoption. on a variety of data that is available in public. We also touched upon the interplay among two vital parts of FPT adoption. The findings on methylation biomarker could form the basis for further probes about epigenetic data. One can improve the present state of the adoption process of dominating set. It is really promising that growth of structural measures could guess scalability and performance of various clique finders and this could lead to design a target solution to choose the best way depending on the input graph. Growth of the parallel graph algorithms area for tasks like vertex cover demand huge memory access.

Acknowledgement

Yegnanarayanan gratefully acknowledges National Board of Higher Mathematics -NBHM, Department of Atomic Energy, Mumbai, Government of India for its financial support with Ref No. NBHM/RP 8/Feb 2019/Fresh with their letter dated 18th Jan 2021.

References

1. Akutsu., et al. "Identification of genetic networks from a small number of gene expression patterns under the Boolean network model". Pacific symposium on biocomputing1999: 17-28.
2. Barlow DP. "Gametic imprinting in mammals". Science 270(5242) 1995: 1610-1613.
3. Bumgarner R. "Overview of DNA microarrays: types, applications, and their future". Current protocols in molecular biology 22(21) 2013: 22.1.
4. Chesler EJ and Langston MA. "Combinatorial genetic regulatory network analysis tools for high throughput transcriptomic data". Systems Biology and Regulatory Genomics Springer 2005: 150-165.
5. Big Data, for better or worse: 90% of world's data generated over last two years.
6. Downey RG and Fellows MR. "Parameterized Complexity".
7. Ehrlich M. "DNA methylation in cancer: too much, but also too little". Oncogene 21(35) 2002: 5400-5413.
8. Fellows MR and Langston MA. "Nonconstructive Advances in Polynomial-Time Complexity". Information Processing Letters, 26(3) 1987:157-162.
9. Fellows MR and Langston MA. "Nonconstructive Tools for Proving Polynomial- Time Decidability". Journal of the ACM 35(3) 1988:727-739.
10. Fellows MR and Langston MA. "On well-partial-order theory and its application to combinatorial problems of VLSI design". SIAM Journal on Discrete Mathematics 5(1) 1992:117-126.
11. Grunstein M and Hogness DS. Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. Proceedings of the National Academy of Sciences 72(10) 1975:3961-3965.
12. Hagan RD., et al. Lower bounds on paraclique density. Discrete Applied Mathematics 204 2016:208-212.
13. Hansen RS., et al. "The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome". Proceedings of the National Academy of Sciences 96(25) 1999: 14412-14417.
14. Herman JG and Baylin SB. "Gene silencing in cancer in association with promoter hypermethylation". New England Journal of Medicine 349(21) 2003:2042- 2054.
15. Holliday R. "DNA methylation and epigenetic inheritance". Philosophical Transactions of the Royal Society of London B: Biological Sciences 326(1235) 1990:329-338.
16. Jay J., et al "systematic comparison of genome-scale clustering algorithms". BMC Bioinformatics 13(10) 2012: S7.
17. Jones PA and Baylin SB. "The fundamental role of epigenetic events in cancer". Nature reviews genetics 3(6) 2002:415-428.
18. Karp R. "Reducibility among combinatorial problems". Complexity of computer computations, 1972: 85-103.
19. Kempe D., et al. "Maximizing the spread of influence through a social network". Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM 2003: 137-146.

20. Krulwich R and Lander E. "Cracking the Code of Life. Public Broadcasting Service 2001.
21. Langston MA., et al. "Scalable combinatorial tools for health disparities research". International journal of environmental research and public health 11(10) 2014:10419-10443.
22. Li E., et al. "Targeted mutation of the DNA methyltransferase gene results in embryonic lethality". Cell 69(6) 1992: 915-926.
23. Okano M., et al. "DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development". Cell 99(3) 1999: 247-257.
24. Perkins AD and Langston MA. "Threshold Selection in Gene Co-Expression Networks Using Spectral Graph Theory Techniques". BMC Bioinformatics 10 2009.
25. Ramsahoye BH., et al, "R. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a". Proceedings of the National Academy of Sciences 97(10) 2000: 5237-5242.
26. Rhodes DR., et al. "Probabilistic model of the human protein-protein interaction network". Nature biotechnology 23(8) 2005: 951-959.
27. Razin A and Cedar H. "DNA methylation and genomic imprinting". Cell 77(4) 1994: 473-476.
28. Robertson N and Seymour PD. "Graph Minors. XX. Wagner's conjecture". Journal of Combinatorial Theory, Series B 92(2) 2004: 325-357.
29. Russo VE., et al. "Epigenetic mechanisms of gene regulation". Cold Spring Harbor Laboratory Press 69(2) 1996.
30. Sporns O. "Contributions and challenges for network models in cognitive neuroscience". Nat Neurosci 17(5) 2014: 652-660.
31. Strichman-Almashanu LZ., et al. "A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes". Genome research 12(4) 2002: 543-554.
32. Tamada Y., et al. "Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection". Bioinformatics 19(2) 2003: II227-236.
33. Voy BH., et al. "Extracting gene networks for low dose radiation using graph theoretical algorithms". PLoS Computational Biology 2(7) 2006: e89.
34. Wang J., et al. "Detecting potential collusive cliques in futures markets based on trading behaviors from real data". Neurocomputing 92(1) 2012: 44-53.
35. Wolen AR., et al. "Genetic dissection of acute ethanol responsive gene networks in prefrontal cortex: functional and mechanistic implications". Plos one 7(4) 2012: e33575.
36. Yen PH., et al. "Differential methylation of hypoxanthine phosphoribosyl transferase genes on active and inactive human X chromosomes". Proceedings of the National Academy of Sciences 81(6) 1984: 1759-1763.

Volume 1 Issue 1 August 2021