Advanced Studies in Biology, Vol. 9, 2017, no. 1, 1 - 7
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/asb.2017.61143

NSGA-II for Biological Graph Compression

A. N. Zakirov and J. A. Brown

Innopolis University, Innopolis, Russia

Abstract

Examinations of a common biological reference organism, (E. coli), demonstrate that NSGA-II is able to provide a series of compressions at various ratios, allows a biologist to examine the organism's connective networks with a measure of certainty of connectiveness. This is due to a novel method of scoring the similarity of the compressed network to the origional during the graph's creation based on the number of false links added to the graph during the compression method.

Keywords: bioinformatics, genetic algorithm, graph compression

1 Introduction

Nowadays, graphs form the foundation of many real-world datasets: computer networks, social networks, biological networks. Development of technologies leads to larger and larger graphs. Some graphs contain millions or even billions of nodes and edges. Therefore, storing and processing their information has too high a cost. The interest in graph data is increasing [4], [18], [21] and many algorithms have been proposed for graph compression. Feder and Motwani [7] consider transforming a graph into a smaller one (in terms of the number of vertices and edges) that preserves certain properties of the original graph, such as connectivity. Adler and Mitzenmacher [3] and Suel and Yuan [19] consider losslessly compressing the Web Graph for efficient search engine storage and retrieval. In [6] there is proposed an query preserving graph compression. Work [8] a Web graph compression algorithm which can be seen as engineering of the Boldi and Vigna (2004) method is presented. Graph compression method based on representing communities with compact data structures is proposed

in [8]. Many of the ideas are similar to those proposed in [16] and [14]. They are based on hierarchical, agglomerative clustering and different methods for efficient implementation.
Comparison of various kinds of biological data is one of the main problems in bioinformatics and systems biology. The difference between biological graphs and, for example, web graphs is in weights of connections, which such graphs as E. coli regulatory network has. In [20] the weighted graph compression problem is proposed, and some initial solutions are provided, which can be used for processing biological and social graphs. Due to increased interests in systems biology, extensive studies have recently been done on comparison of biological networks. In [17] clustering based method for metabolic networks compression in presented. In [12], [15] data compression methods have been applied to comparison of large sequence data and protein structure data [13], [22] In [9] CompressEdge and CompressVertices methods for comparing large biological networks are proposed. [11] introduces a genetic algorithm for graphs (social and biological) compression that is based on the similarity of nodes, also, genetic algorithm approach was utilized in order to develop a compressed graph for a single compression ratio on a number of biological and non-biological graphs. This paper study the application of the Non-dominated Sorting Genetic Algorithm (NSGA-II) [5] for biological graph compression. We target a good compression ratio as well as keeping as much natural biological information in compressed graph as possible.

2 Materials and Methods

The NSGA-II, see [5], is a Multiple Objective Optimization (MOO) algorithm and is an instance of an Evolutionary Algorithm from the field of Evolutionary Computation. NSGA-II is an extension of the Genetic Algorithm for multiple objective function optimization. The objective of the NSGA algorithm is to improve the adaptive fit of a population of candidate solutions to a Pareto front constrained by a set of objective functions. The algorithm uses an evolutionary process with surrogates for evolutionary operators including selection, genetic crossover, and genetic mutation. The population is sorted into a hierarchy of sub-populations based on the ordering of Pareto dominance. Similarity between members of each sub-group is evaluated on the Pareto front, and the resulting groups and similarity measures are used to promote a diverse front of non-dominated solutions.

2.1 Objectives and fitness functions

On problem of biological graph compression we have an objectives of compressing the graph and still save enough information about its origin. The fitness

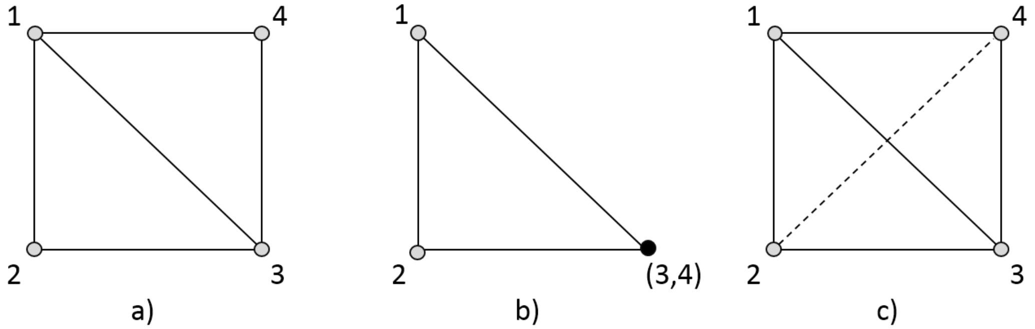

Figure 1: Example of fake link appearance after compressing original graph a) by merging nodes 3 and 4 on b). Fake link 2-4 on c) appears after decompression

function is based on the notion of similarity of original and compressed graphs. The similarity definition is related to fake links, that appear after compression and then decompression of the graph (fig 1). So the similarity is

$$S = 1 - \frac{F_r}{F_T}$$

where F_r is number of fake links that appears after decompression for tested graph and F_T is total number of all possible fake links in the graph. F_T is calculated from the original graph and its maximum compression in a one node, which gives as full graph after decompression. Amount of links in this full graph is $n(n-1)/2$, where n is a number of nodes in graph. Amount of links in original graph is known. The difference of this two link number will be F_T. We are targeting minimizing the fake links while maximizing the compression. Minimizing the fake links equals to maximizing the similarity S. The compression ratio is

$$C = \frac{N_c}{N_o}$$

where N_c is number of nodes after compression and N_o is original number of nodes.

2.2 Dataset

Examining the dataset exampled for the compression of biological networks. Different biological networks are available at [2]. In this work, we are targeting the processing of the gene regulatory network of Escherichia coli (E. coli) [1]. This network is a relatively small network, it was cleaned of all duplicate links (nodes that indicate both activation and inhibition) and all unknown links. Our final graph consist of 1123 nodes and 2108 edges.

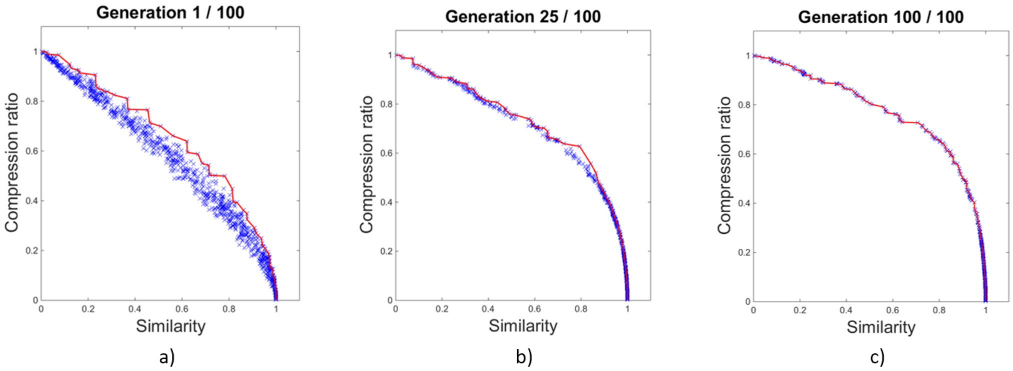

Figure 2: Results of algorithm after 1 (a), 25 (b) and 100 (c) populations

3 Results and Discussion

As evolution progresses we see a relatively steady increase in the relative fitness of the compression. As we can see on figure 2 (a) , first population is far away from the optimal condition. Nevertheless, after 25 generations (b) we can see how Pareto fronts began to appear and move towards optimum. The last results - 100 generations - is shown on (c). We run our tests with mutation rate 0.25 and crossover rate 0.9. This parameters have shown best results in [11]. Nevertheless, influence of mutation and crossover rates is field for further researches.

The results show that NSGA-II forms a thigh front to optimum similarity until a compression ratio of about 0.3. The Similarity degrades with increasing speed until at about 0.8 we have nearly half of the links in the graph being false. By developing not just a single good graph at a target compression ratio, such as in [11], but a series of graphs at many ratios, a biologist can select the highest amount of compression with minimal loss of information contained in the graph.

4 Conclusion

Taking into consideration biological background of graph, NSGA-II, as multi-objective genetic algorithm, provide more natural view of space, than the algorithms focusing only compression ratio as a measurement. We proposed similarity of original and compressed graph as a second objective. Further testing on other biological datasets is required in order to demonstrate the generality of the method.

Acknowledgements. This study was funded by RFBR according to the research project No. 16-31-60090.

References

[1] Avi Maayan: E.coli gene regulatory network dataset.
http://research.mssm.edu

[2] Avi Maayan: Network datasets for download.
http://research.mssm.edu/maayan/datasets/qualitative_networks.shtml

[3] M. Adler, M. Mitzenmacher, Towards compressing web graphs, *In Proceedings of the Data Compression Conference,* (2001), 203-212.
https://doi.org/10.1109/dcc.2001.917151

[4] R. Agrawal, T. Imielinski, and A. Swami, Mining association rules between sets of items in large databases, *SIGMOD Rec.,* **22** (1993), 207-216.
https://doi.org/10.1145/170036.170072

[5] K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, *IEEE Transactions on Evolutionary Computation,* **6** (2002), 182-197.
https://doi.org/10.1109/4235.996017

[6] W. Fan, J. Li, X. Wang and Y. Wu, Query preserving graph compression, *In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data,* (2012), 157-168.
https://doi.org/10.1145/2213836.2213855

[7] T. Feder, R. Motwani, Clique partitions, graph compression and speeding-up algorithms, *In Proceedings of the Twenty-third Annual ACM Symposium on Theory of Computing,* (1991), 123-133.
https://doi.org/10.1145/103418.103424

[8] S. Grabowski, W. Bieniecki, Tight and simple web graph compression for forward and reverse neighbour queries, *Discrete Appl. Math.,* **163** (2014), 298-306. https://doi.org/10.1016/j.dam.2013.05.028

[9] M. Hayashida, T. Akutsu, Comparing biological networks via graph compression, *BMC Systems Biology,* **4** (2010).
https://doi.org/10.1186/1752-0509-4-s2-s13

[10] C. Hernandez, G. Navarro, Compressed representations for web and social graphs, *Knowl. Inf. Syst.,* **40** (2014), 279-313.
https://doi.org/10.1007/s10115-013-0648-4

[11] T. K. Collins, J. A. Brown, S. K. Houghten and Q. Qu, Evolving graph compression using similarity measures for bioinformatics applications, *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology,* (2016).

[12] A. Kocsor, A. Kertesz-Farkas, L. Kajan and S. Pongor, Application of compression-based distance measures to protein sequence classification: A methodological study, *Bioinformatics,* **22** (2005), 407-412. https://doi.org/10.1093/bioinformatics/bti806

[13] N. Krasnogor, D. A. Pelta, Measuring the similarity of protein structures by means of the universal similarity metric, *Bioinformatics,* **20** (2004), 1015-1021. https://doi.org/10.1093/bioinformatics/bth031

[14] K. LeFevre, E. Terzi, Grass: Graph structure summarization, *Proceedings of the 2010 SIAM International Conference on Data Mining,* (2010), 454-465. https://doi.org/10.1137/1.9781611972801.40

[15] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney and H. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics,* **17** (2001), 149-154. https://doi.org/10.1093/bioinformatics/17.2.149

[16] S. Navlakha, R. Rastogi and N. Shrivastava, Graph summarization with bounded error, *In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data,* (2008), 419-432. https://doi.org/10.1145/1376616.1376661

[17] H. Ogata, W. Fujibuchi, S. Goto and M. Kanehisa, A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters, *Nucleic Acids Research,* **28** (2000), 4021-4028. https://doi.org/10.1093/nar/28.20.4021

[18] Q. Qu, J. Qiu, C. Sun and Y. Wang, Graph-based knowledge representation model and pattern retrieval, *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery,* **5** (2008), 541-545. https://doi.org/10.1109/fskd.2008.7

[19] T. Suel, J. Yuan, Compressing the graph structure of the web, *In Proceedings of the IEEE Data Compression Conference,* (2001), 213-222. https://doi.org/10.1109/dcc.2001.917152

[20] H. Toivonen, F. Zhou, A. Hartikainen and A. Hinkka, Compression of weighted graphs, *In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* (2011), 965-973. https://doi.org/10.1145/2020408.2020566

[21] X. Yan, M. R. Mehan, Y. Huang, M. S. Waterman, P. S. Yu and X. J. Zhou, A graph-based approach to systematically reconstruct human transcriptional regulatory modules, *Bioinformatics,* **23** (2007), i577-i586. https://doi.org/10.1093/bioinformatics/btm227

[22] Y. Zhao, M. Hayashida and T. Akutsu, Integer programming-based method for grammar-based tree compression and its application to pattern extraction of glycan tree structures, *BMC Bioinformatics,* **11** (2010). https://doi.org/10.1186/1471-2105-11-s11-s4

Received: November 30, 2016; Published: December 15, 2016