IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 24, No. 3, June, 2019, pp. 229-232
Journal homepage: http://iieta.org/journals/isi

ANOVAG3: A Hybrid Algorithm for Inferring Gene Regulatory Network Using Time Series Gene Expression Data

Shaimaa M. Elembaby[1*], Vidan F. Ghoneim[2], Manal Abdel-Wahed[1]

[1] Faculty of engineering, Cairo University, Giza 12613, Egypt
[2] Faculty of engineering, Helwan University, Cairo 11795, Egypt

Corresponding Author Email: eng_s_elembaby@yahoo.com

https://doi.org/10.18280/isi.240301

Received: 3 March 2019
Accepted: 24 May 2019

Keywords:
gene regulatory network, GENIE3, DREAM5, one-way analysis of variance, *tree-based ensemble method*

ABSTRACT

GENIE3 achieves best results in inferring Gene Regulatory Network (GRN) with DREAM4 challenge data. Whereas, correlation coefficient derived from two-way analysis of variance (ANOVA) records best result for DREAM5 challenge data. Here we try to improve results of GENIE3 on time series gene expression data by using one-way ANOVA along time axis as a prior step to GENIE3. GENIE3 takes long time with huge number of genes so one-way ANOVA finds significant genes before execution of GENIE3. Integration between one-way ANOVA and GENIE3 is a hybrid algorithm entitled ANOVAG3. ANOVAG3 is applied only on time series gene expressions and takes less running time than GENIE3 with huge data. ANOVAG3 is compared with other algorithms which infer GRN by Area Under the Receiver Operating Characteristic Curve (AUROC) using DREAM5 challenge networks. Although ANOVAG3 is not dependent on perturbation data or transcription factors, it records comparable results for networks 1 and 3 and records best results for network 4 (AUROC =0.5628) of DREAM5 challenge data. ANOVAG3 records better results in DREAM 5 networks 2, 3 and 4 (AUROC= 0.5190, 0.6458 and 0.5628) compared to GENIE3 and PLSNET considering large scale time series data employed in this work.

1. INTRODUCTION

Bayesian Networks and Boolean Networks are used to infer GRN [1]. Stochastic Master Equations methods and Differential Equation (DE) methods are used in GRN inference which can be divided into Qualitative, Nonlinear, Piecewise-linear and partial. [2] Probabilistic Boolean Network Models and Neural network models were produced as Methods for inference of GRN. Relevance and Bayesian networks are used to infer structure of GRN but Dynamic Bayesian and Ordinary Differential Equation (ODE) networks are used to infer the dynamics of GRN [3]. There are traditional and non- traditional model for GRN inference as models based on Evolutionary algorithms [4]. DREAM competitions data used in comparing algorithms of GRN inference [5].

TSNI Algorithm used time series gene expression to infer GRN using ordinary differential equations. it increases number of samples by interpolation and reduce data dimension by principal component analysis to solve problem that number of samples almost is less than number of genes. After that, it solves the equation by singular value decomposition [6].Time series gene expression data only is used here in this work, whereas other methods used Knockout, Knockdown and perturbation data[7-9].Other algorithms can deal with time series and perturbation data as GENIE3[10], ENNET [11], NIMEFI[12], PLSNET algorithm [13]. GENIE3 decomposes GRN inference problem of N genes into N different regression problems and solves them by tree-based ensemble method. ENNET algorithm combines Gradient Boosting with regression and use machine learning model to select subset of edges for building GRN. NIMEFI algorithm solves N sub problems by Ensemble Elastic Net or Support Vector Regression (E-SVR). PLSNET algorithm uses Partial least squares (PLS) regression as feature selection method to solve N sub problems [13]. TIGRESS applies least angle regression with stability selection to infer GRN [14], Several information theoretic methods as MRNET [15], ARACNE [16] and CLR [17] are dependent on the mutual information. iRafNet is improving of GENIE3 by adding heterogeneous data as gene knock down and protein-protein interaction to random forest algorithm [18]. RGPM also uses heterogeneous data to identify transcriptional regulators [19]. BIGENIE uses several biclustering methods to group date after that GENIE3 was applied to each group [20]. Non-linear correlation coefficients derived from two-way ANOVA between transcription factors TF and target genes TG are also used in inferring GRN [21]. Three-way ANOVA is used to improve the results of two-way ANOVA and to detect network's three genes motifs and interactions between them as cascade chain (CSC), dense overlapping regulon (DOR) or feed forward loop (FFL) [22]. ADANET algorithm converts problem of GRN inference to set of independent tasks and solves them with AdaBoost ensemble classifier and uses structure of models to discover relation between transcription factors and regulatory genes [23]. Some of previous algorithm collected and implemented in comparison study with DREAM4 data (10 genes and 100 genes) [24].

In this work, we set a comparison between area under the receiver operating characteristic curve (AUROC) of DREAM5[25] of some algorithms (GENIE3, ENNET,

NIMEFI, PLSNET, TIGRESS, MRNET, ARACNE, CLR, iRafNet, RGBM, BiGENIE and correlation coefficient of two-way ANOVA) that were published and AUROC of ANOVAG3, GENIE3 and PLSNET which are applied on time series data extracted from DREAM5. Most of the published algorithms use all samples including time series, perturbation data and transcription factors.

2. MATERIALS AND METHODS

2.1 Data set

DREAM5 provides data of each network in three files: chip feature, gene expression and transcription factor (TF) file. Gene expression file represents data in a matrix form of microarray chip (rows or samples) versus gene expression of each gene (columns). Chip feature file has information for each microarray chip.

In this work, extracting time series data was done using time column of chip feature file. Table 1 represents networks of DREAM5; number of chips (samples) which were used in other algorithms, number of genes and number of chips (samples) of time series data used in this work. Bold numbers explain numbers of genes and samples which are used with ANOVAG3 algorithm.

Table 1. DREAM5 data description

Network	Organism	Number of samples	Number of used time series samples	Number of genes
Network1	In silico	805	463	1643
Network2	S. aureus	160	30	2810
Network3	E. coli	805	463	4511
Network4	S. cerevisiae	536	298	5950

2.2 One-way analysis of variance (ANOVA)

One-way ANOVA grouped by time is performed to each gene. It returns the p-value for the null hypothesis whether the means of groups related to time are equal or not. If P-value < 0.05, then the gene is significant and has clear pattern with time and so we take this gene in building GRN. One-way ANOVA explains the behavior of genes with time. The limit of P-value can be decreased to decrease the number of genes used with GRN inference algorithm; least P-values represents most significant genes.

2.3 GENIE3 method

GENIE3 records best results with DREAM4 challenge, it decomposes GRN problem with N genes into N regression trees problems solved by random forest or Extra trees. GENIE3 thus gets relations between each gene and the other genes. Rank aggregation is held to construct global GRN. GENIE3 is a popular algorithm which is used in GRN inference for steady state data [10].

2.4 ANOVAG3

ANOVAG3 is a hybrid algorithm integrating both one-way ANOVA and GENIE3. As most of real gene expression data is time series, ANOVAG3 is applied on time series data. One-way ANOVA presents behavior of genes with time and extract genes which have a significant effect in forming gene regulatory network (GRN). GENIE3 separates each gene as a target gene and uses other genes to predict this target gene by tree-based ensemble method. Relations between target gene and other genes are ranked. This process is repeated for every gene then rank aggregation is employed to construct a global GRN. The benefit of integrating both methods is that One-way ANOVA reduces the number of genes introduced to GENIE3 as illustrated in Figure 1.

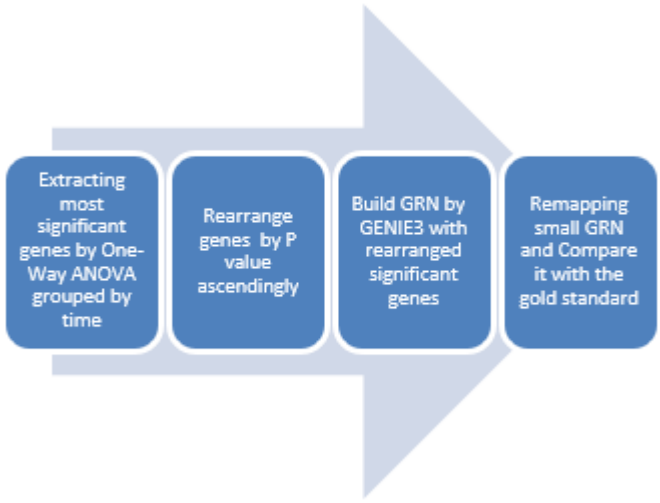

Figure 1. Steps of ANOVAG3

3. RESULTS

Time series samples of DREAM5 are used to infer GRN with PLSNET, GENIE3 and ANOVAG3 algorithms, Area under the Receiver Operating Characteristic Curve (AUROC) is used to compare the accuracies of all. The resultant AUROC of GENIE3[10], ENNET [11], NIMEFI[12], PLSNET[13], TIGRESS[14], MRNET [15], ARACNE [16], CLR [17], iRafNet [18], RGBM[19], BiGENIE[20], the score of non-parametric and nonlinear correlation coefficient derived from two way analysis of variance(ANOVA)[21], ADANET[23] are all shown in table 2. Most of the published algorithms record results of three networks out of four networks that were presented by DREAM5challenge except BiGENIE which were applied on E. coli network only (network3). PLSNET records best result with network one. correlation coefficient of two-way ANOVA records best result with network3 and ANOVAG3 records best result with network4.

4. DISCUSSION

Through all algorithms applied on DREAM5 data ANOVAG3 accomplished best result with network 4 (5950genes). The explanation of this results is that DREAM5 network 4 has the largest number of genes and one-way ANOVA is basically reducing number of genes as it plays the main role in extracting significant genes which has special response patterns with time. This makes ANOVAG3 perform well with large scale GRN. In network 4, 3, one-way ANOVA step takes less than one minute after that genes are

reduced from 5950 genes to 5620 genes in network 4 and from 4511 genes to 4252 genes in network 3 at P value= 0.05. Execution time of ANOVAG3 is less than execution time of GENIE3 by nearly an hour in network 3 and 4. This time is different from one situation to other and it is dependent on number of samples and computer's clock speed. ANOVAG3 records best results also for networks 2 (AUROC =0.5190), While none of the rest algorithms attained results with network 2. In network 2 ANOVAG3 reduce number of genes from 2810 to 1919. In network one, ANOVAG3 reduces 1643 genes to 1581 genes at P value= 0.05. ANOVAG3 reduces execution time by several minutes than GENIE3 in network 1 and 2 with time series data.

Table 2. Results of DREAM5

Algorithm	Net1	Net2	Net3	Net4
GENIE3[10]	0.814	---	0.618	0.517
ENNET [11]	0.867	---	0.642	0.532
NIMEFI [13]	0.817	---	0.625	0.518
PLSNET [13]	0.862	---	0.577	0.519
TIGRESS [14]	0.789	---	0.589	0.514
Naïve TIGRESS [14]	0.782	---	0.595	0.517
MRNET [11]	0.668	---	0.525	0.501
CLR [11]	0.773	---	0.590	0.516
ARACNE [11]	0.763	---	0.572	0.504
iRafNet [19]	0.813	---	0.641	0.523
RGBM [19]	0.846	---	0.633	0.546
BiGENIE [20]	---	---	0.642	---
Correlation coefficient of two-way ANOVA [21]	0.78	---	0.671	0.518
ADANET [11]	0.752	---	0.596	0.517
GENIE3 with time series	0.7363	0.4897	0.6352	0.525
PLSNET with time series	0.7154	0.4897	0.543	0.559
ANOVAG3 with time series (at P value=0.05)	0.7244	0.5190	0.6458	0.5628

ANOVAG3 can also be used with determined number of genes as it deals with arranged list of genes. For example, instead of number of genes extracted at P valve <0.05 we can construct GRN from the most 1000 significant genes. As the execution time of one-way ANOVA is short (it takes less than one minute with DREAM5 data), and the execution time of GENIE3 considering thousands of genes may take several hours. Wherefore, reducing the number of genes by one-way ANOVA as a prior step to GENIE3 can reduce execution time, hopefully raise the results. And is essential with problems including huge number of genes to simplify it by constructing sparse matrix of GRN in shorter time.

The highlighted part of table 2 represents our results after the implementation of PLSNET, GENIE3 and ANOVAG3 with time series data only of DREAM5. GENIE3 records best results in network 1 (0.7363) but ANOVAG3 record best results in network 2,3 and 4 (0.5190, 0.6458 and 0.5628).

5. CONCLUSION

Although reducing number of samples using time series samples only without perturbations, ANOVAG3 records best result with network 4 compared to previously published work and yet to two popular inferring algorithms that were implemented in this work; PLSNET, GENIE3.

ANOVAG3 deals with time series data only and has a different job not as correlation coefficient of two-way ANOVA. two-way ANOVA determines the similarities or associations between transcription factors and target genes, where determining transcription factors may be difficult in real data. One-way ANOVA in ANOVAG3 extract significant genes which have a special behavior with time.

One-way ANOVA can be used as a prior step to several GRN inference algorithms to reduce number of genes in large scale problems and thus reduce execution time. And hopefully raise the results as accomplished in this work.

Execution time of ANOVAG3 as GENIE3 depends on number of genes, number of samples and computer's clock speed but execution time of ANOVAG3 is often less than execution time of GENIE3.

REFERENCES

[1] Jong, H.D. (2002). Modeling and simulation of genetic regulatory systems: A literature review. Journal of Computational Biology, 9(1). https://doi.org/10.1089/10665270252833208

[2] Lee, W.P., Tzau, W.S. (2009). computational methods for discovering gene networks from expression data. Briefing in Bioinformatics, 10(4): 408-423. https://doi.org/10.1093/bib/bbp028

[3] Sima, C., Hua, J., Jung, S. (2009). Inference of gene regulatory network using time-series data: A survey. Current Genomics, 10: 416-429. https://doi.org/10.2174/138920209789177610

[4] Panse, C., Kshirsagar, M. (2013). Survey on modeling methods applicable to gene regulatory network. International Journal on Bioinformatics & Biosciences (IJBB), 3(3): 13-23. https://doi.org/10.5121/ijbb.2013.3302

[5] Linde, J., Schulze, S., Henkel, S.G., Guthke, R. (2015). Data –and knowledge-based modeling of gene regulatory network: An update. EXCLI Journal, 14: 346-378. https://doi.org/10.17179/excli2015-168

[6] Bansal, M., Gatta, G.D., di Bernardo, D. (2006). Inference of gene regulatory network and compound mode of action time course gene expression profiles. Bioinformatics, 22(7): 815-822. https://doi.org/10.1093/bioinformatics/btl003

[7] Pinna, A., Soranzo, N., De La Fuente, A. (2010). From Knockout to networks: establishing direct cause-effect relationships through graph analysis. PLoS ONE, 5(10): e12912. https://doi.org/10.1371/journal.pone.0012912

[8] Kalmet, S., Flassing, R.J., Sundmacher, K. (2010). TRANSWESD: Inferring celluler networks with transitive reduction. Bioinformatics, 26(17): 2160-2168. https://doi.org/10.1093/bioinformatics/btq342

[9] Pinna, A., Heise, S., Flassing, R.J., De La Fuente, A., Kalmet, S. (2013). Reconstruction of large-scale regulatory networks based on perturbation graphs and transitive reduction: Improved methods and their evaluation. BMC System Biology, 7: 73. https://doi.org/10.1186/1752-0509-7-73

[10] Irrthum, A., Wehenkel, L., Geurts, P. (2010). Inferring regulatory networks from expression data using treebased methods. PLoS One, 5(9): e12776. https://doi.org/10.1371/journal.pone.0012776

[11] Sławek, J., Arodź, T. (2013). ENNET: inferring large gene regulatory networks from expression data using gradient boosting. BMC Syst Biol, 7(1): 106. https://doi.org/10.1186/1752-0509-7-106

[12] Ruyssinck, J., Huynh-Thu, V.A., Geurts, P., Dhaene, T., Demeester, P., Saeys, Y. (2014). NIMEFI: gene regulatory network inference using multiple ensemble feature importance algorithms. PLoS One, 9(3): e92709. https://doi.org/10.1371/journal.pone.0092709

[13] Guo, S., Jiang, Q., Chen, L., Guo, D. (2016). Gene regulatory network inference using PLS-based methods. BMC Bioinformatics, 17: 545. https://doi.org/10.1186/s12859-016-1398-6

[14] Haury, A.C., Mordelet, F., Vera-Licona, P., VertJP. (2012). Tigress: trustful inference of gene regulation using stability selection. BMC Systems Biology, 6(1): 1. https://doi.org/10.1186/1752-0509-6-145

[15] Meyer, P.E., Kontos, K., Lafitte, F., Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol., pp. 79879. https://doi.org/10.1155/2007/79879

[16] Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D., Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics, 7(Suppl 1): S7. https://doi.org/10.1186/1471-2105-7-S1-S7

[17] Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol., 5: e8. https://doi.org/10.1371/journal.pbio.0050008

[18] Petralia, F., Wang, P., Yang, J., Tu, Z. (2015). Integrative random forest for gene regulatory network inference. Bioinformatics, 31(12): i197–i205. https://doi.org/10.1093/bioinformatics/btv268

[19] Mall, R., Cerulo, L., Kunji, K., Bensmail, H., Sabedot, T.S., Noushmehr, H., Iavarone, A., Ceccarelli, M. (2018). RGBM: Regularized gradient boosting machines for the identification of transcriptional regulators of discrete glioma subtypes. Nucleic Acids Research, 46(7): 7e39. https://doi.org/10.1093/nar/gky015

[20] Cannoodt, R., Ruyssinck, J., De Preter, K., Dhaene, T., Saeys, Y. (2013). Network inference by integrating biclustering and feature selection. BeNeLux Bioinformatics Conference – Brussels, December, Abstract ID: 014.

[21] Küffner, R., Petri, T., Tavakkolkhah, P., Windhager, L., Zimmer, R. (2012). Inferring gene regulatory networks by ANOVA. Bioinformatics, 28(10): 1376–1382. https://doi.org/10.1093/bioinformatics/bts143

[22] Tavakkolkhah, P., Zimmer, R., Küffner, R. (2018). Detection of network motifs using three-way ANOVA. PLOSONE. https://doi.org/10.1371/journal.pone.0201382

[23] Sławek, J., Arodz, T. (2012). ADANET: Inferring gene regulatory networks using ensemble classifiers. Conference: Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, At Orlando, Florida, pp. 434–441. https://doi.org/10.1145/2382936.2382992

[24] Elembaby, S.M., Ghoneim, V.F., Abdel Wahed, M. (2018). Comparing gene regulatory inferring algorithms with different perspective. Instrumentation, Mesure, Metrologie, 17(4): 653-661. https://doi.org/10.3166/i2m.17.653-661

[25] https://www.synapse.org/#!Synapse:syn2787209/wiki/70351, accessed on July 12, 2019.