A Heuristic Gene Regulatory Networks Model for Cardiac Function and Pathology

Armita Zarnegar[1], Peter Vamplew [1], Andrew Stranieri[1], Herbert F Jelinek[2]

[1] Centre for Informatics and Applied Optimization, Federation University, Ballarat, Australia
[2] Australian School of Advanced Medicine, Macquarie University, Sydney and School of Community Health, Charles Sturt University, Albury, Australia

Abstract

Genome-wide association studies (GWAS) and next-generation sequencing (NGS) has led to an increase in information about the human genome and cardiovascular disease.

Understanding the role of genes in cardiac function and pathology requires modeling gene interactions and identification of regulatory genes as part of a gene regulatory network (GRN). Feature selection and data reduction not sufficient and require domain knowledge to deal with large data.

We propose three novel innovations in constructing a GRN based on heuristics. A 2D Visualised Co-regulation function. Post-processing to identify gene-gene interactions. Finally a threshold algorithm is applied to identify the hub genes that provide the backbone of the GRN. The 2D Visualized Co-regulation function performed significantly better compared to the Pearson's correlation for measuring pairwise associations ($t=3.46$, $df=5$, $p=0.018$). The F-measure, improved from 0.11 to 0.12. The hub network provided a 60% improvement to that reported in the literature. The performance of the hub network was then also compared against ARACNe and performed significantly better ($p=0.024$).

We conclude that a heuristics approach in developing GRNs has potential to improve our understanding of gene regulation and interaction in diverse biological function and disease.

1. Introduction

The advent of more advanced and faster gene sequencing programs for genome-wide association studies (GWAS) and next-generation sequencing (NGS) has led to a plethora of information about the human genome and cardiovascular disease (CVD in particular that requires novel approaches for determining gene regulatory networks (GRN). Information on cardiac genes consists of several databases that are difficult to combine with traditional gene regulatory network (GRN) models as they address heart evolution, cardiac development, function, cardiac conduction systems and cardiac pathology separately and therefore are based on different heuristics. As an example, well above one hundred heart rhythm determinant genes that are sex-dependent have been identified [1], with many more associated with development, function and pathology. Restricting GRN studies to possible disease biomarkers disregards genes associated with controlling expression of downstream genes and interconnection of gene regulatory pathways. Thus co-occupancy of transcription factors located on chromatin has been used to identify cardiac enhancer genes with ChIP and high-throughput sequencing (ChIP-seq) identifying thousands of prospective cardiac regulatory sequences associated with gene enhancers [2]. To this complexity of large data has now been added miRNA studies that indicate their role in gene function and regulation as well as additional complex mechanisms associated with cardiac development including transcriptional and post-transcriptional mechanisms [3, 4]. Understanding these regulatory gene interactions that are the basis of gene expressions and functions, still remains a difficult task. Although several methods have been proposed to infer gene regulatory (interactions) networks from gene expression data, most current methods have limited accuracy due to the curse of dimensionality where the number of genes far exceed the number of observations [5]. Feature selection and data reduction attempt to address this but computational and statistical techniques are not sufficient and require domain knowledge [6, 7]. A great deal of knowledge is known heuristically about GRNs and the nature of gene interactions but few studies have incorporated heuristics into a GRN discovery process. Heuristics can aid the GRN discovery process by identifying heuristics related to the nature of gene interactions [8] and heuristics related to the structure of the gene network applying the greedy hill-climbing algorithm, simulated annealing or the K2 algorithm[9]. GRN are graphs with scale-free properties where most genes are connected to a small number of global transcription factor genes referred to as *Hubs* [10]. This paper proposes the use of *hub* genes as a meaningful

heuristic approach in GRN discovery.

2. Methods

The heuristic framework adopted here is based on first implementing a co-regulation function which measures pair-wise dependencies between genes. Results are then post-processed to reduce false positives and finally a *hub* network is designed to construct the backbone of the GRN.

2.1. 2D Visualised co-regulation

To allow for multiple two-gene interactions a 2D Visualised Co-regulation function is proposed based on a frequentist approach by constructing a matrix comprising discretized expression levels for one gene along rows and for the other genes along columns. The matrix provides normalized data on how often two genes have appeared or being expressed, in one sample as either high in both samples (HH), low in both samples (LL), high-low (HL) or low-high (LH). These interactions are related to the possible up or down or dual regulation by genes in the network.

2.2. Post-processing

Post-processing is required to identify when a gene directly up or down- regulates another gene and also for indirect relationships when the interaction is mediated through a chain of intermediary genes [11]. The post-processing step eliminates false positive interactions by looking for the absence of the reverse interaction that is HL is a reverse interaction to HH. To eliminate these false positives we apply a simple rule:

$$threshold\ low < \frac{number\ of\ LL}{number\ of\ HH} < threshold\ up$$

The user-defined lower threshold was set at 10% and the upper threshold at 60%.

2.3. Identification of hubs

The final step was to use *hubs* to construct the backbone on which the GRN can be built [12]. The first layer of the GRN was built by first calculating the co-regulation measure for all genes in the expression data. This is followed by the post-processing and background correction steps and finally selecting highly connected genes based on the previously mentioned threshold of the hub's connectivity. This resulted in a network of hubs, which formed the backbone structure of our target network. The second layer of the GRN is made up of genes that are the most strongly connected to the hub nodes using a weights function and applied a background correction by first normalizing the correlation values between each gene and any other genes and then filtered out those which were less than 0.5 starting at the top of the list of genes interacting with this gene. The GRN is a scale free made up of three categories of genes based on their degree of connectivity, which was set at 15. In our heuristic selection procedure if the degree of the node was less than 15 (identified as being the average hub degree in the biological network literature), the node was selected exactly according to its connectivity degree otherwise it was selected according to:

$$n = \begin{cases} degree & , \quad degree < 15 \\ 15 + (degree - 15) * 0.3, & degree \geq 15 \end{cases}$$

2.4. Dataset

To compare the performance of our proposed system with other systems, we assembled a gold standard network with a known level of biological and experimental noise, and regulatory relationships with SynTReN, which uses known network parts to build a simulated network [13]

2.5. Statistics

F-measure was applied to indicate accuracy of the heuristics. The F-measure is defined by the harmonic mean of positive predictive power (PPV) and specificity (S):

$$F = \frac{2PPV}{PPV+S}$$

3. Results

(1)The 2D Visualized Co-regulation function performed significantly better compared to the Pearson's correlation for measuring pairwise associations (t=3.46, df=5, p=0.018). The F-measure, defined as the harmonic mean of the positive predictive value and sensitivity improved from 0.11 to 0.12. Applying the heuristic post-processing improved the F-measure further to 0.19. The *hub* network provided a 60% improvement to that from the literature [14, 15]. The performance of the *hub* network (F=0.17) was compared against ARACNe (0.11) and performed significantly better (p=0.024) [16].

4. Discussion

We developed a novel technique for Gene Regulatory Network discovery that integrates heuristic information into the discovery process. The heuristic model is conceptually simple and computationally efficient. Our

current results using the co-regulation, post-processing and *hub* modelling heuristic demonstrates that this model has the potential to derive an accurate GRN architecture. Integration of data from different sources has the potential for GRNs to be more accurate and relevant to biology and medicine. Future research needs to extend to applying the heuristics to functional gene sets in combination with our Hub Network [17]. Related gene sets to a *hub* can be retrieved and information about other genes in the gene set can be used to build a GRN more effectively. Performance of the 2D Visualised Co-regulation function could also be further improved by sophisticated discretization methods such as reported by Fayyad and Irani's MDL [18]. The heuristic methods described here may be valuable for building dynamic regulatory networks in cardiac cell development and function that also contain some uncertainty [19, 20]lengths.

References

[1] Iacobas DA, Iacobas S, Thomas N, Spray DC. Sex-dependent gene regulatory networks of the heart rhythm. Functional and Integrative Genomics 2010 2010/03//;10(1):73-86.

[2] He A, Kong SW, Ma Q, Pu WT. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proceedings of the National Academy of Science USA 2011 April 5, 2011;108(14):5632-7.

[3] Giudice J, Xia Z, Wang ET, Scavuzzo MA, Ward AJ, Kalsotra A, et al. Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. Nature Communications 2014 04/22/online;5.

[4] Rustagi Y, Jaiswal HK, Rawal K, Kundu GC, Rani V. Comparative characterization of cardiac development specific microRNAs: fetal regulators for future. PLoS ONE 2015;10(10):e0139359.

[5] Wang S-L, Li X-L, Fang J. Finding minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification. BMC Bioinformatics 2012;13(1):1-26.

[6] Lo K, Raftery A, Dombek K, Zhu J, Schadt E, Bumgarner R, et al. Integrating external biological knowledge in the construction of regulatory networks from time series expression data BMC Systems Biology 2012;6(101).

[7] Lee W, Tzou W. Computational Methods for Discovering Gene Networks from Expression Data. Briefings in Bioinformatics 2009;10(4):408 - 23.

[8] Fioravanti F, Helmer-Citterich M, Nardelli E. Modeling gene regulatory network motifs using state charts. BMC Bioinformatics 2012;13(Suppl 4).

[9] Numata K, Imoto S, Miyano S, editors. A Structure Learning Algorithm for Inference of Gene Networks from Microarray Gene Expression Data Using Bayesian Networks. Bioinformatics and Bioengineering, 2007 BIBE 2007 Proceedings of the 7th IEEE International Conference on; 2007 14-17 Oct. 2007.

[10] Lu X, Jain V, Finn P, Perkins D. Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. Molecular Systems Biology 2007;3(98).

[11] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics 2003;34(2):166 - 76.

[12] Isalan M, Lemerle C, Michalodimitrakis K, Horn C, Beltrao P, Raineri E, et al. Evolvability and hierarchy in rewired bacterial gene networks. Nature 2008;452:840-6.

[13] Leemput K, Van Den Bulcke T, Dhollander T, De Moor B, Marchal K, Van Remortel P. Exploring the Operational Characteristics of Inference Algorithms for Transcriptional Networks by Means of Synthetic Data. Artificial Life 2008;14(1):49-63.

[14] Martinez-Anonio A. Operation of the gene regulatory network in *Escherichia coli* In: Babu MM, editor. Bacterial gene regulation and transcriptional networks: Horizon Scientific Press; 2013.

[15] Ma H, Kumar B, Ditges U, Gunzer F, Buer J, Zeng A. An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. Nucleic Acids Research 2004;32(22):6643-9.

[16] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 2006;7 Suppl 1:S7.

[17] Langfelder P, Mischel P, Horvarth S. When Is Hub Gene Selection Better than Standard Meta-Analysis. PLoS ONE 2013;8(4).

[18] Fayyad U, Irani K, editors. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. International Joint Conference on Artificial Intelligence 1993. 9.

[19] Gong W, Koyano-Nakagawa N, Li T, Garry DJ. Inferring dynamic gene regulatory networks in cardiac differentiation through the integration of multi-dimensional data. BMC Bioinformatics 2015;16(1):1-16.

[20] Grieb M, Burkovski A, Sträng JE, Kraus JM, Groß A, Palm GN, et al. Predicting Variabilities in Cardiac Gene Expression with a Boolean Network Incorporating Uncertainty. PLoS ONE 2015;10(7):e0131832

Address for correspondence.

Herbert Jelinek.
School of Community Health
Charles Sturt University
Thurgoona Drive
Albury, 2460
Australlia.
hjelinek@csu.edu.au.