Motif enhanced ensemble network inference prediction

Rick Reijnders

i6167500

[ra.reijnders@student.maastrichtuniversity.nl](mailto:ra.reijnders@student.maastrichtuniversity.nl)

Network Biology MSB1014

Project Report Draft

15-10-2019

## Introduction

With the era of high-throughput data, the amount of gene expression, microRNA expression, and methylation data has increased tremendously while the understanding of these regulatory mechanisms is lagging behind. (Epi)genetic dysregulation can lead to altered phenotypes such as cancer, therefore, understanding the inference of these regulatory mechanisms is essential to treat or cure altered phenotypes. Little is known on how these regulating layers work together in detail [1,2]. Methods to infer regulatory mechanisms from experimental and modelled datasets have been studied greatly, and many methods have been developed. However, each method has its own strengths and weaknesses. By combining these methods in an ensemble fashion, the strengths can be combined which creates a better initial prediction of the inference in the network. To further improve the prediction, statistically significant overrepresented topological patterns (network motifs) are used as an additional filter to extract biologically-derived significant patterns found in the inferred networks.

Network inference methods are widely used to improve understanding of complex regulatory networks. Often literature-based information is used to create the initial networks, such as pathways or protein-protein interaction networks. By combining network inference and literature-based information the inferred networks possibly give new insight to the understanding of regulatory networks. By incorporating network motifs, topological significance is included in the analysis. This results in a combined effort of information driven (all genes in model are known), network inference (inference based on *in sillico* data) and topological information (network motifs) to predict the true inference of the model.

## Aim

This project aims to estimate the improvement of network inference of regulatory networks by using network motifs.

## Methods

### Obtain data

Data was obtained from a dedicated network inference challenge, the HPN-DREAM breast cancer network inference challenge (<https://www.synapse.org/HPN_DREAM_Network_Challenge>) (synaps id syn1720047). This dataset was generated using *in sillico* simulations of a model described in Chen et al. [3]. This means that the inferred networks can always be compared to the correct, original model. The dataset consists of 20 genes as rows and 360 columns as time/state.

### R version and packages

All processing of the dataset was performed in R version 3.6.1 (2019-07-05), using packages ‘GRENITS’, ‘igraph’, ‘corrr’, ‘GENIE3’, ‘parmigene’ and ‘RCy3’. The datasetwas imported, as was the true network in graph format.

### Inference methods

The Bayesian network inference method from the GRENITS package was used to infer the dataset based on Bayesian networks. *LinearNet* created Monte Carlo Marcov Chains (MCMCs) based on the inputted data. The result was analyzed using *analyse.output*, which created a network probability matrix. All values of this matrix lower than the set threshold of 0.08 (removing noise) was set to 0. This created a weighted adjacency matrix which was plotted using the *graph\_from\_adjacency\_matrix* function from the Igraph package, converting it into an Igraph graph object.

Regression-based trees inference from the GENIE3 package was used to infer the dataset based on the GENIE3 algorithm. The function GENIE3 was used to obtain a weighted matrix from the dataset, based on ensembles of regression trees. *getLinkList* was used to convert the weighted matrix to a list of links, using a threshold of 0.1 (removing noise). Afterward, *graph\_from\_data\_frame* was used to convert the list to an Igraph graph object.

The parmigene package contains multiple algorithms for mutual-information-based inference. The function *knnmi.all* computes the mutual information based on all pairs of rows of the matrix, which was used in the function *aracne.m*. This function uses the ARACANE algorithm to reconstruct gene interactions. Afterward, *graph\_from\_adjacency\_matrix* was used to convert the list to an Igraph graph object.

### Motif detection

After obtaining three Igraph objects, these were analyzed using a custom-made function, called *fGetMotifsFromGraph*. This function used the Igraph package as many required functions were available, however, needed intensive knitting to produce the desired result. The *fGetMotifsFromGraph* function works as followed. An Igraph graph object was loaded and the *graph.motifs* function, using a motif size of 3, counted all the occurring 3-node motifs based on the input graph. For every motif which occurred more than 0, an image of the current motif, and amount of found occurrences, was created using *graph.isocreate*. To locate the motif in the given graph, first, the unconnected odes of the inferred network were removed. Based on the resultant graph, metrics such as amount nodes and links were used as input for random network creation. This random network creation was performed using the *random.graph.game* function, using parameters directed=T, type = “gnm”. For every motif, 1000 random networks were generated with matching number of nodes and links. For every random network the amount of currently selected motif was calculated. This amount is saved and summed for the 1000 random networks, afterward, divided by 1000 to get the average occurrence of the specific selected motif. This number is multiplied by the *betaFactor* (set to 1.3), if the occurrence of the specific motif in the inferred network is higher than the random network, it is considered important. At this point, the important motifs are known but the exact location is not. To extract all the possible motif locations from the inferred network, *subgraph\_isomorphisms* is used using the motif and inferred graph as input. The result is a list of all possible motif isomers, which can be merged to a graph. This created a subgraph of the inferred network with edge density representing the occurrence of a motif multiple times. By merging the graphs from all different motifs, an inferred motif-density-based graph was formed. This graph was further simplified using the *simplify* function, which creates a graph that can be directly compared to the true graph.

Load methods into Cytoscape? & compare with true net mismatch n stuff?

By comparing the ensemble network with the original network, the overlap and mismatch can be determined which indicates the performance of prediction.

## Results

## Discussion

Improvements:

t-test

increase random nets

reflect the original network better

## References

[1] Albert R. (2007). Network inference, analysis, and modeling in systems biology. The Plant cell, 19(11), 3327–3338. doi:10.1105/tpc.107.054700

[2] Ahnert, S. E., & Fink, T. M. (2016). Form and function in gene regulatory networks: the structure of network motifs determines fundamental properties of their dynamical state space. Journal of the Royal Society, Interface, 13(120), 20160179. doi:10.1098/rsif.2016.0179

[3] W. W. Chen, B. Schoeberl, P. J. Jasper, M. Niepel, U. B. Nielsen, D. A. Lauffenburger, and P. K. Sorger, ‘Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data.,’ Mol. Syst. Biol., vol. 5, no. 239, p. 239, Jan. 2009.

[4] Guo, S., Jiang, Q., Chen, L., & Guo, D. (2016). Gene regulatory network inference using PLS-based methods. BMC bioinformatics, 17(1), 545. doi:10.1186/s12859-016-1398-6

[5] Wong, E., Baur, B., Quader, S., & Huang, C. H. (2012). Biological network motif detection: principles and practice. Briefings in bioinformatics, 13(2), 202–215. doi:10.1093/bib/bbr033