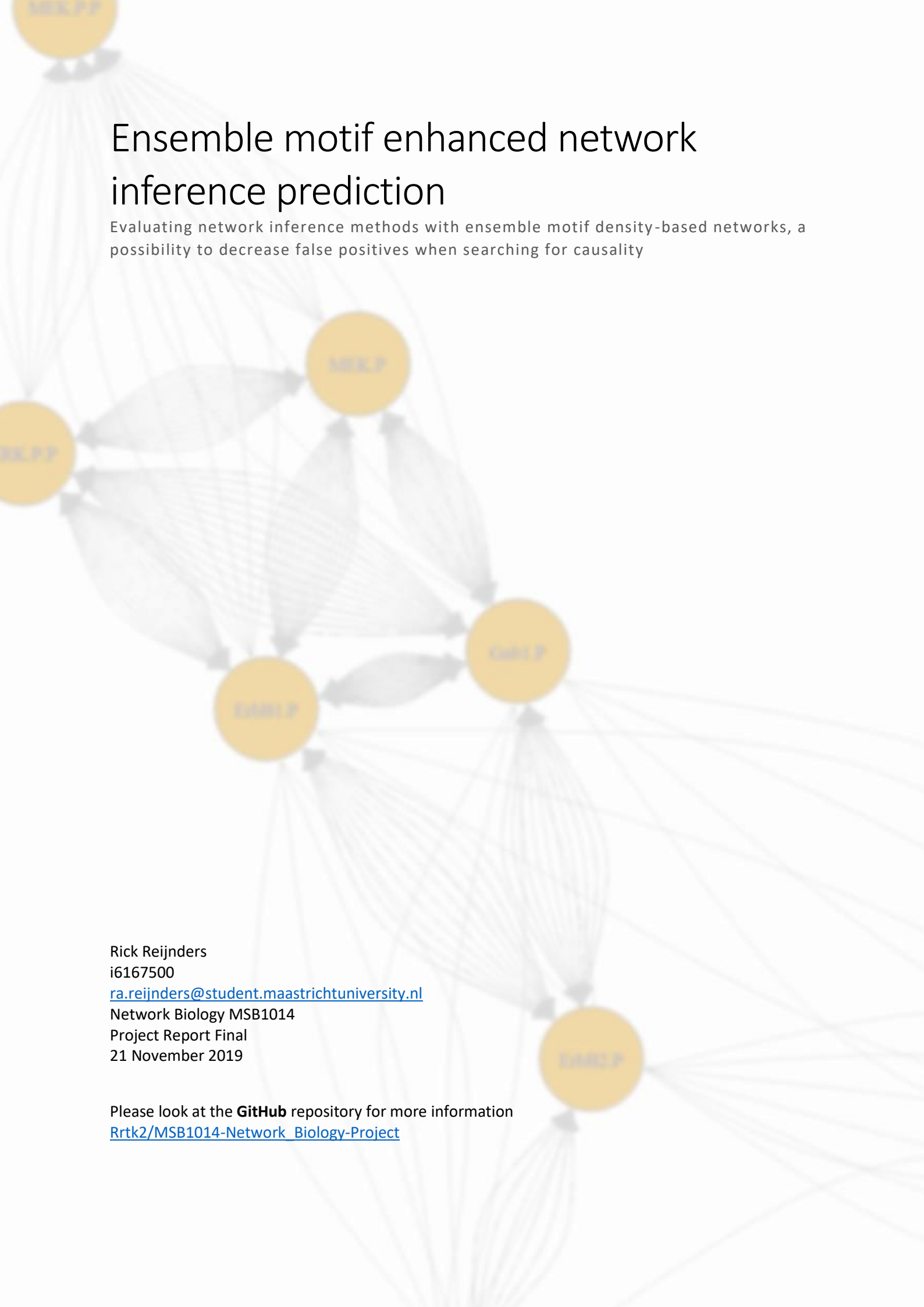
Ensemble motif enhanced network inference prediction

Evaluating network inference methods with ensemble motif density-based networks, a possibility to decrease false positives when searching for causality



Rick Reijnders
i6167500
ra.reijnders@student.maastrichtuniversity.nl
Network Biology MSB1014
Project Report Final
21 November 2019

Please look at the **GitHub** repository for more information
Rrtk2/MSB1014-Network_Biology-Project

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 cooperate in detail [1,2]. Methods to infer regulatory mechanisms from experimental and modelled datasets have been studied greatly and many methods have been developed. Network inference methods are widely used to improve understanding of complex regulatory networks. Literature-based information is often 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.

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 times/states.

R version and packages

All processing of the dataset was performed in R version 3.6.1 (2019-07-05)[4], using packages 'GRENITS'[5], 'igraph'[6], 'corrr'[7], 'GENIE3'[8], 'parmigene'[9] and 'RCy3'[10]. The dataset was 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 the 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 is described below.

An Igraph graph object was loaded and the *graph.motifs* function counted all the occurring 3-node directed 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 nodes 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 generation. This random network generation was executed using the *random.graph.game* function, using parameters directed=TRUE, 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 risen to the power of *betaFactor* (set to 1.3). If the occurrence of the specific motif in the

inferred network is higher than the average amount in the random network and higher than the set threshold *hardTH* (3), it is considered important. The important motifs are known but the exact location in the inferred network is not. To extract all the possible motif locations from the inferred network, *subgraph_isomorphisms* is used. The result of this function is a list of all possible motif isomers in the inferred network, which can be merged to a graph. This created a subgraph of the inferred network, with link density representing the multiple occurrences of a specific motif between different nodes. 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.

Comparison with true graph

To compare the generated graphs with the true network, Cytoscape[11] was used to load the networks and merge the intersection of the graphs. The resultant graph is processed further as unconnected nodes were removed. The number of nodes and links reflect the correct prediction, when compared to the number of nodes and links the graphs started with, a prediction accuracy can be calculated. This was executed for the inferred and motif-based networks.

Results

True network

The inference methods were compared to the true network, the resultant networks and true network can be seen in figure 1. The true network (1.A) indicates the true causality which should be inferred using the inference methods, containing 20 nodes, of which 3 are unconnected and 74 directed links. Node sizes are linked to link count, scaled to min/max of the local graph. Figure 1.B visualizes the mutual information inferred network, which contains 20 connected nodes with 82 links. Figure 1.C contains the network of the Bayes inferred network, containing 14 connected nodes and 6 unconnected nodes, with a total of 16 links. Figure 1.D is the regression based inferred network, containing 16 connected nodes and 4 unconnected nodes, with a total of 42 links.

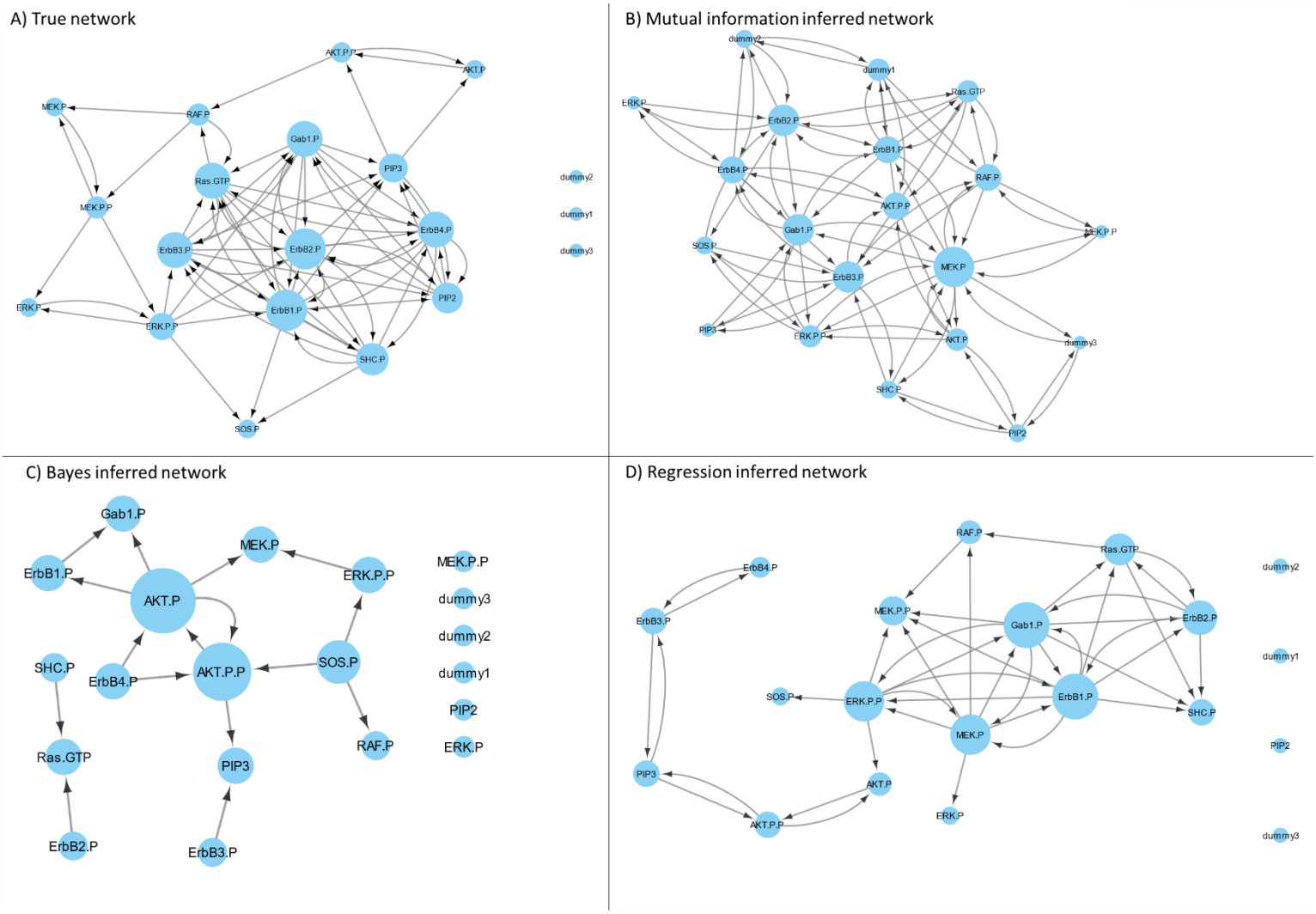


Figure 1. Core networks used in this project. A) True network, which data is based on and other methods should infer. B) Mutual information inferred network, indicative of a bidirectional connected network without unconnected nodes. C) Bayes inferred network, chain-like regulation with some unconnected nodes. D) Regression inferred network, visually similar to the true network containing directed and bidirected links, some unconnected nodes. Node sizes are dependent on the linkcount, scaling to relative min and max of the current graph.

Motif detection

3-node directed motif detection on each inferred network leads to diverse important motifs. The top 2 most occurring motifs can be seen in figure 2. The mutual information motifs (2.A) seem to depend on bidirectional links and occurs in high numbers compared to other methods. Bayes indicates regulation-like motifs, starting in one node and passing through other nodes, as seen in figure 2.B. In figure 2.C regression motifs have bidirectional links and directed with relative high number motif occurrences.

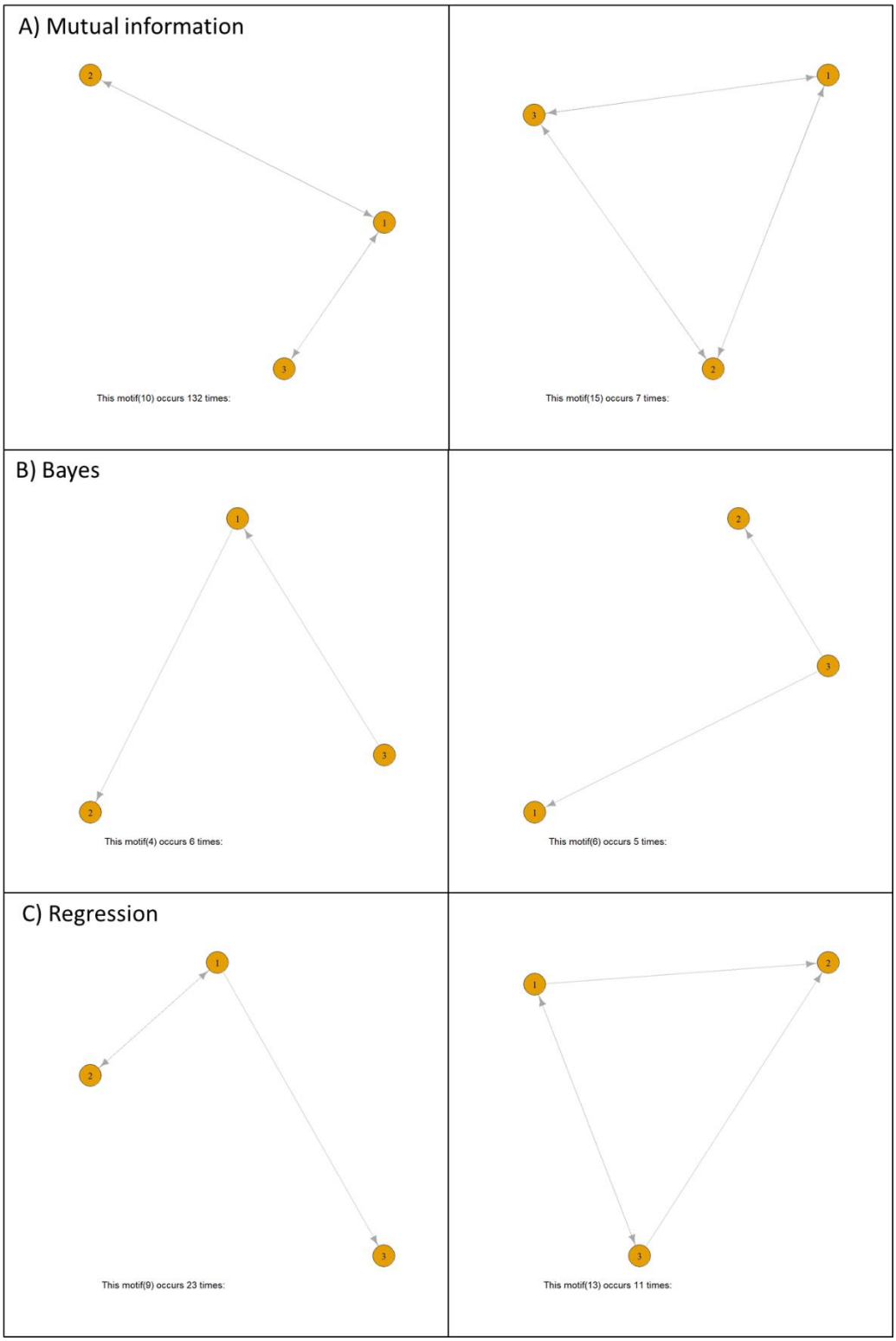


Figure 2. Top 2 occurring motifs per inference method. Each inference method with some of the most occurring motifs. Direction indicated by arrow.

Motif graph generation

By assessing the location of the most *important* motifs in the inferred networks, it is possible to recreate a subset of the inferred graph when motifs are combined. In figure 3, it can be observed how the density of the motif occurrences is reflected by the density of the links. These density-based networks are simplified, which results in the networks that can be compared to the true graph. Figures 3.A and 3.C dominate the density graphs due to high occurrence of motifs. When simplified, it can be seen these graphs are very similar to the inferred networks (figure 1.B and 1.D). Figure 3.B resembles a small subgraph of the inferred network as seen in figure 1.C.

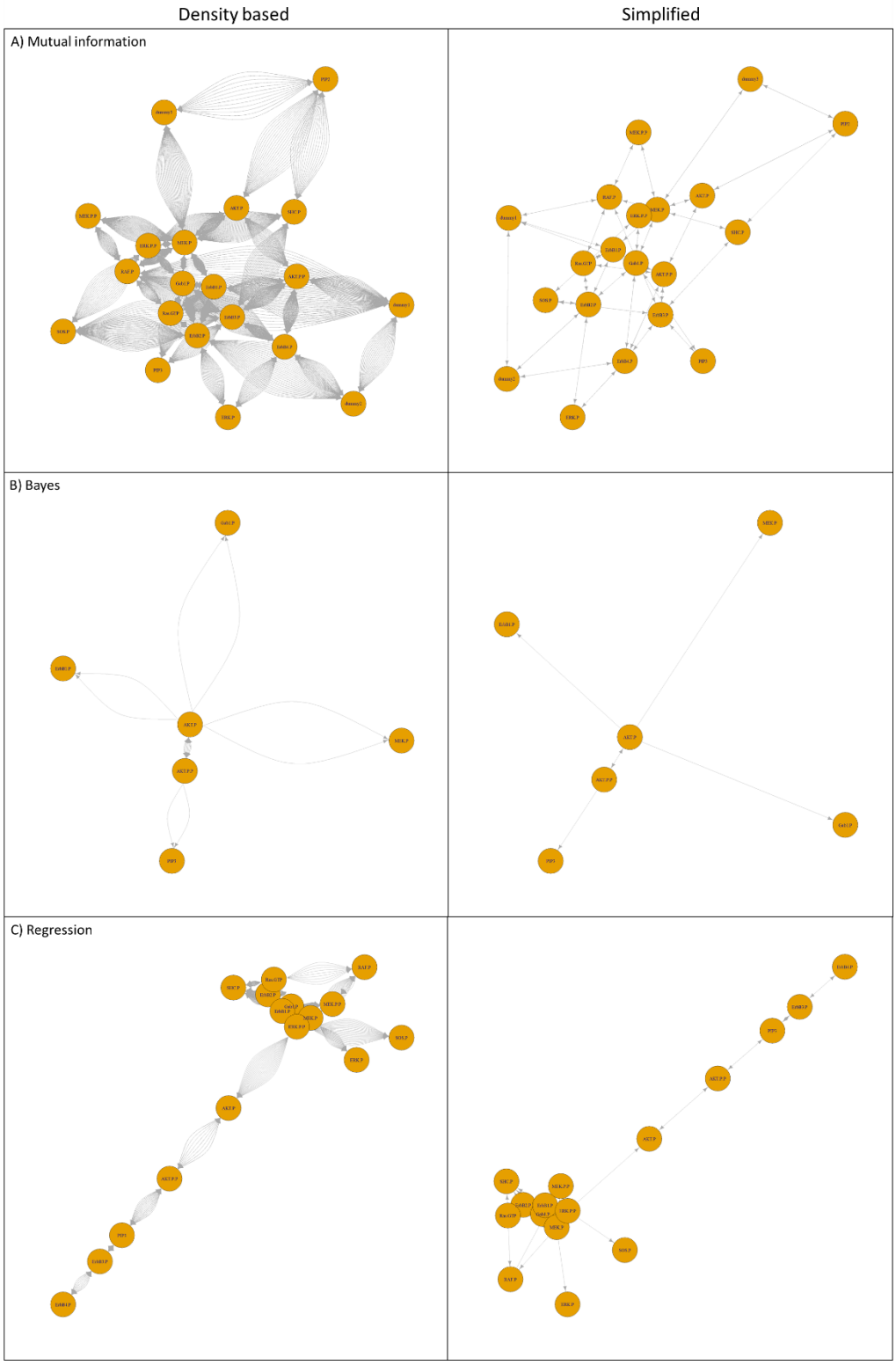


Figure 3. Density-based graphs and simplified versions. Density-based is made by overlapping all important motifs. Simplified does not assess the density, only if a link and direction is present. Direction indicated by arrow.

Comparison

To compare all results, table 1 was created containing the number of nodes and links together with the prediction accuracy. It can be seen for the mutual information and regression methods that accuracy does not decrease when using the motif networks. The motif networks based on Bayes inference contain less nodes and links. This resulted in a lower node accuracy, however, the link accuracy decreased slightly.

Table 1. Accuracies of correctly predicted graph features

Description	Mututal information		Bayes		Regression	
	Inferred	Motifs	Inferred	Motifs	Inferred	Motifs
Nodes before	20	20	14	6	16	16
Links before	82	82	16	6	42	42
Nodes correct	15	15	9	2	14	14
Links correct	26	26	6	2	23	23
Nodes accuracy	75%	75%	64%	33%	88%	88%
Links accuracy	32%	32%	38%	33%	55%	55%

Discussion

Inferred networks

Different inference methods produce vastly different results, leading to different motif detection as can be seen in figure 1. When comparing the results, it can be seen that the true graph (1.A) and the regression inferred network (1.D) look similar, containing a complex cluster of nodes which are densely interconnected and a 'tail' of less dense interactions. Dummy nodes were unconnected which reflects the correct situation. The mutual information network (1.B) is fully connected, including the dummy nodes on the link of the network. This network seems to be randomly connected to everything increasing the number of links greatly. However, the number of links is the closest to the real number of links in the true network. Despite introducing many links, several could be correct by chance. The Bayes inferred network (1.C) has the least amount of links, but this is due to the method as self-loops are not possible. It excluded several nodes, including the dummy and 3 other nodes. In conclusion, the methods to infer the true causality of the network is diverse and parameters such as thresholds will have a considerable effect on the result. If the true network is unknown, methods how to approximate these parameters as best as possible should be considered.

Motif detection

Based on the inferred networks, motifs were determined to be important if occurred more than average in random networks with similar features. However, if the inferred networks are biased due to the method or parameter settings, the possibility for certain motifs to occur is biased as well. This effect can be observed in figure 2. As the mutual information inferred network contains many bidirectional links, it is likely that the most occurring motifs will involve the bidirectional links. The most occurring motif in figure 2.A confirms this assumption, as it is found 132 times; indicative of the inference method bias. These style of motifs, bidirectional links, can indicate forms of co-regulation. The Bayes inference method does not contain self-loops. Thus, these motifs will likely resemble chain-like motifs; one regulating the next node or regulating multiple nodes. The motifs in figure 2.B resemble this effect, as the first motif indicates a chain of regulation and the second motif indicates regulation of multiple nodes, like a transcription factor. Regression-based trees inference can account for nonlinear effects, thus this method is able to deal with self-loops, as well as chain like regulation. This is clearly reflected in figure 2.C, in which the two motifs indicate co-regulation and node regulation. To conclude on motif detection, it is very sensitive to the input, the inferred network. A great understanding of each inference method is required to determine how motifs can aid in each situation and should be accounted for when combining motifs from multiple inference methods.

Motif graph generation

The density-based graphs indicate 'hotspots' for high occurring motifs in the given location of the graph. This can be used as an additional filter, only selecting the most occurring links, thus the most important nodes which is assumed to be more important than the rest of the inferred network. However, due to time constrains the density-based graphs were simplified to obtain a similar network for comparison. As the inference method bias occurs in the motif detection, it is bound to occur in the density-based networks as well. Figure 3.A displays this effect strongly, as all nodes are connected with many links, simplifying to the original inferred network. In figure 3.B this effect is les prominent, as the motif detection resulted in low numbers. Figure 3.C indicates a good chance to filter on the density of the links, as there is a high packed cluster and less packed tail. All simplified networks were compared to the true network and indicates that motif filtering did not have a direct benefit. However, by varying the parameters and using density as threshold, the prediction accuracy might improve. To conclude on motif graph generation, it is interesting to see that the simplified networks made from several

motifs can capture the same information as the fully inferred network in most cases. By including density-based cutoffs, a better prediction can be possible removing the false positives.

User function fGetMotifsFromGraph

This function was constructed in short time span and many improvements are possible. Instead of observing the motif occurrence above average than random network occurrence, use a one sample t test. Increasing the number of random networks will increase significance, as well as increasing feature similarity to the inferred networks. Many optimizations are possible such as memory usage, visual plots, functionality, documentation and structure.

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] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[5] Morrissey E (2019). GRENITS: Gene Regulatory Network Inference Using Time Series. R package version 1.36.0.

[6] Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org

[7] Simon Jackson, Jorge Cimentada and Edgar Ruiz (NA). corrr: Correlations in R. R package version 0.3.1.9000. https://github.com/drsimonj/corrr

[8] Huynh-Thu et al. (2010) Inferring regulatory networks from expression data using tree-based methods. PLoS ONE 5(9): e12776.

[9] Gabriele Sales and Chiara Romualdi (2012). parmigene: Parallel Mutual Information estimation for Gene Network reconstruction. R package version 1.0.2. https://CRAN.R-project.org/package=parmigene

[10] Ono K, Muetze T, Kolishovski G, Shannon P, Demchak, B. CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API [version 1; referees: 2 approved]. F1000Research 2015, 4:478.

[11] Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.