A Mixture-of-Experts Approach for Gene Regulatory Network Inference

Borong Shao

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Computer Science. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:

Author(s):

Borong Shao

E-mail: bosh12@student.bth.se

External advisor:

Prof. Veselka Boeva

Dept. Computer Systems & Technologies

Technical University of Sofia, Bulgaria

E-mail: vboeva@tu-plovdiv.bg

University advisor:

Dr. Niklas Lavesson

Dept. Computer Science & Engineering

Faculty of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

ABSTRACT

Context. Gene regulatory network (GRN) inference is an important and challenging problem in bioinformatics. A variety of machine learning algorithms have been applied to increase the GRN inference accuracy. Ensemble learning methods are shown to yield a higher inference accuracy than individual algorithms.

Objectives. We propose an ensemble GRN inference method, which is based on the principle of Mixture-of-Experts ensemble learning. The proposed method can quantitatively measure the accuracy of individual GRN inference algorithms at the network motifs level. Based on the accuracy of the individual algorithms at predicting different types of network motifs, weights are assigned to the individual algorithms so as to take advantages of their strengths and weaknesses. In this way, we can improve the accuracy of the ensemble prediction.

Methods. The research methodology is controlled experiment. The independent variable is *method*. It has eight groups: five individual algorithms, the generic average ranking method used in the DREAM5 challenge, the proposed ensemble method including four types of network motifs and five types of network motifs. The dependent variable is *GRN inference accuracy*, measured by the area under the precision-recall curve (AUPR). The experiment has training and testing phases. In the training phase, we analyze the accuracy of five individual algorithms at the network motifs level to decide their weights. In the testing phase, the weights are used to combine predictions from the five individual algorithms to generate ensemble predictions. We compare the accuracy of the eight method groups on Escherichia coli microarray dataset using AUPR.

Results. In the training phase, we obtain the AUPR values of the five individual algorithms at predicting each type of the network motifs. In the testing phase, we collect the AUPR values of the eight methods on predicting the GRN of the Escherichia coli microarray dataset. Each method group has a sample size of ten (ten AUPR values).

Conclusions. Statistical tests on the experiment results show that the proposed method yields a significantly higher accuracy than the generic average ranking method. In addition, a new type of network motif is found in GRN, the inclusion of which can increase the accuracy of the proposed method significantly.

Keywords: GRN inference, Ensemble learning,
Mixture-of-Experts, network motif analysis

Acknowledgements

First and foremost, I thank my supervisor Dr. Niklas Lavesson. He supported my thesis work from beginning to the end, with patience, by giving insightful academic opinions, and with kind attitude. The thesis would not be in shape without his help.

Secondly, I express my thanks to my external supervisor Prof. Veselka Boeva. She helped me by providing me many insightful suggestions.

Thirdly, I thank my friend Raja Muhammad Khurram Shahzad for his encouragement, discussions, and suggestions. I also thank Dr. Gupta Udatha for his advices.

Last but not least, I express my gratitude for my parents my father Mr. Jiafeng SHAO and my mother Mrs. Ping HE for giving me an opportunity to study abroad, encouraging me during hard times, and having trust on me.

Contents

Abstract	i
Acknowledgements	ii
1 Introduction	1
2 Background	3
2.1 Terms and Concepts	3
2.2 Related Work	6
3 Aims and Objectives	9
3.1 Aims	9
3.2 Objectives	9
3.2.1 Training phase	9
3.2.2 Testing phase	10
4 Research Questions	11
5 Method	12
5.1 Mixture-of-Experts architecture	12
5.2 Network motif analysis – decide weights of experts	14
5.3 Experiment	15
5.3.1 Datasets	15
5.3.2 Measurements	16
5.3.3 Experiment flow	16
5.3.4 Data collection	18
5.3.5 Statistical tests	18
5.4 Validity threats	19
5.4.1 Internal validity	19
5.4.2 External validity	20
5.4.3 Construct validity	20
5.4.4 Statistical validity	20

6	Results	22
6.1	Training phase	22
6.2	Testing phase	26
6.3	Results from statistical tests	27
7	Analysis	29
8	Discussion	31
9	Conclusions and Future Work	33
	References	35
	Appendix	41
	Appendix I	41
	Appendix II	44

Chapter 1

Introduction

Gene regulatory network (GRN) inference plays an important role in assisting biologists to investigate the gene regulatory mechanisms (Almudevar et al., 2006). High-throughput technologies, such as microarrays (Bubendorf, 2001), RNA-seq (RNA Sequencing) (Lalonde et al., 2011), ChIP-on-chip (a technique that combines chromatin immunoprecipitation with microarray technology) (Buck & Lieb, 2004) and genome-wide association studies produce large amount of gene expression data. These data enable computational biologists to infer transcriptional gene regulations on the genome-scale (Wu & Chan, 2011). GRN inference has been a long-standing challenge (Hache et al., 2009; Margolin & Califano, 2007; Marbach et al., 2010). More accurate GRN inference algorithms and methods are in demand.

Machine learning has gained an important role in analyzing biological data (Baldi & Brunak, 2001; Tarca et al., 2007; Larrañaga et al., 2006). For the GRN inference problem, applied machine learning methods include multiple linear regression (Honkela et al., 2010), Bayesian network analysis (Li et al., 2011), mutual information (Margolin et al., 2006; Faith et al., 2007), and Random Forests (Huynh-Thu et al., 2010), etc. Each method has its own underlying methodology which is not related to gene regulatory mechanisms. So the methods inevitably make assumptions when dealing with gene expression datasets, which cause different types of systematic inference bias (De Smet & Marchal, 2010). For example, mutual information based algorithms such as ARANCE (Margolin et al., 2006), CLR (Faith et al., 2007), MRNET (Meyer et al., 2007) and Relevance Networks (Butte & Kohane, 2000) systematically discriminate activating interactions and bias towards repressing interactions (Altay & Emmert-Streib, 2010).

Experiments have shown that the accuracy of GRN inference is still low (Marbach et al., 2010). GRN inference is an under-determined problem. The amount of independent microarray datasets is too insignificant to determine the correct network out of all possible networks. In order to deal with the under-determination and to reduce the complexity, researchers make certain assumptions regarding the GRN features. These assumptions simplify the gene regulatory mechanisms

and cause bias. For example, De Smet & Marchal (2010) states that module-based inference methods such as LeMoNe (Joshi et al., 2009) and DISTILLER (Lemmens et al., 2009) assume that the regulators who regulate the same target genes act combinatorially. This assumption makes it difficult to use the methods for identifying the mode of combinatorial interactions between the regulators.

In addition to improving individual GRN inference algorithms, ensemble learning has shown its advantages in terms of GRN inference accuracy. Marbach et al. (2010) show that the predictions from ensemble method are more accurate than the predictions from individual algorithms. Marbach et al. (2012) shows that the ensemble method is as accurate or more accurate than the top individual algorithms. Similarly, Kolde et al. (2012) propose a robust rank aggregation (RRA) method for gene list integration, which yields a higher accuracy than the individual gene lists. Consistently, ensemble methods have been shown to yield a higher accuracy than individual methods in other biological problems, such as finding homologous members of a protein family in databases (Alam et al., 2004), and cancer classification (Tan & Gilbert, 2003). To conclude, ensemble methods have a strong potential in biological data mining.

However, no ensemble approach has been developed at the network motif level to deal with the GRN inference problem. Although Marbach et al. (2010) suggests that algorithms have different accuracy on predicting the four types of network motifs, there have not been any quantitative comparisons among GRN inference algorithms at the network motif level.

In this study, we compare the prediction accuracy of five GRN inference algorithms at predicting network motifs, e.g., which algorithm is more accurate in predicting a certain type of network motif. Then, we assign weights to the five algorithms according to their accuracy. After that we use the Mixture-of-Experts architecture (Jacobs et al., 1991) to generate the ensemble prediction. Finally, we use precision-recall curves to compare the prediction accuracy among the proposed ensemble method, the generic average ranking used in the DREAM5 challenge, and the five individual algorithms. Results from the experiments show that the proposed ensemble method yields significantly higher accuracy than the generic average ranking method. In addition, the inclusion of a new type of network motif can increase the accuracy of the proposed ensemble method significantly.

In this chapter, we provide explanations of the terms and concepts that are relevant to this study. The terms and concepts come from the fields of machine learning, bioinformatics and computational biology. They are explained in the context of this study. After that, we introduce the related work including individual GRN inference algorithms and ensemble methods.

2.1 Terms and Concepts

Machine learning concerns developing methods and building systems to find useful information from data. Mitchell (1997) provides a formal definition of machine learning: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E ”. In bioinformatics, machine learning methods have been applied to a wide variety of biological problems. For example, machine learning methods can be used to identify different types of cancer (Cho & Won, 2003; Jaskowiak et al., 2013) and to predict genes that share the same regulatory mechanisms or functions (Heyer et al., 1999). There are three types of machine learning: supervised learning, unsupervised learning and reinforcement learning (Mitchell, 1997). The first two types of machine learning are commonly applied to solve problems in bioinformatics, such as GRN inference (Maetschke et al., 2013).

A machine learning process includes at least training and testing. The data used in training is referred to as the training set and the data used in testing is referred to as the testing set. The training set is used to train the learning algorithms, e.g., linear regression, to build models, e.g., linear regression equations. The testing set is used to assess how well the models fit the data, e.g., how accurate can the models make predictions on the same type of data that is not used for training. Ensemble learning refers to the procedure of training multiple machine learning algorithms and integrating their output (Dietterich, 2000). The individual algorithms used in the ensemble learning method give the same type of output, which can be integrated by many methods, such as averaging and ma-

majority voting (Dietterich, 2000).

Both individual learning algorithms and ensemble learning methods have been applied in gene regulatory network (GRN) inference. They aim to predict the network of gene regulations from the microarray datasets. In detail, genes are DNA molecules which control the biological traits and biochemical processes that comprise life. They interact with each other to realize the precise regulation of life activities. One gene can directly or indirectly regulate other genes. The gene who regulates is the regulator and the gene being regulated is the target gene. Genes can be regulators and targets at the same time. All the interactions between genes constitute the gene regulatory network (GRN). Microarray technology can measure the expression levels of thousands of genes simultaneously using DNA chips. Many microarray measurements (chips) constitute the microarray datasets, which are used to infer the GRN. Table 2.1 shows the microarray dataset:

Table 2.1: An example of Microarray dataset

Measurements (chips)	Gene 1	Gene 2	Gene 3	...	Gene m
1	0.3245	0.4815	0.7854	..	0.0132
2	0.1021	1.0942	0.3464	..	0.2521
...	...	...	...	..	...
n	0.2042	0.0773	0.9976	..	0.9986

From the machine learning point of view, each row (measurement) of the dataset is an instance and each column (gene) is an attribute or feature. The goal of GRN inference is to identify the network connections among these attributes. GRN is made up of nodes and directed edges, which represent genes and the direct regulations among genes respectively.

GRN has characteristic features: sparseness, scale-freeness, enriched network motifs and modularity (Hecker et al., 2009). In this study, we focus on exploring the feature of enriched network motifs. Shen-Orr et al. (2002); Milo et al. (2002) state that these four types of motifs: Fan-out, Fan-in, Cascade and Feed-Forward Loop are the building blocks of GRN. They appear significantly more frequently in GRN than in random networks. The four types of network motifs are shown by Figure 2.1.

In GRN inference research, accuracy is the most important evaluation standard. It is evaluated against the experimentally validated gene interactions (gold standards), such as RegulonDB (Gama-Castro et al., 2011) for *Escherichia coli*, ChIP-chip and evolutionarily conserved binding motifs (MacIsaac et al., 2006)

Figure 2.1: Four types of GRN motifs – the building blocks of GRN

for *S.cerevisiae*. Precision and recall are the commonly applied measurements. Precision is the fraction of retrieved instances that are relevant (Mitchell, 1997). In the context of GRN inference, it corresponds to the ratio of correctly predicted edges to all predicted edges. Recall is the fraction of relevant instances that are retrieved (Mitchell, 1997). In the context of GRN inference, it corresponds to the ratio of correctly predicted edges to all edges in the true network.

Precision and recall are not isolated. Increasing one of them often causes the decrease of the other. For example, if we take only the top 100 edges from a GRN prediction, the precision is high because the edges ranked on top of the list have higher confidence values; but the recall is low because even if all the 100 edges are correct, there are large amount of edges not covered. Similarly, if we only increase recall, e.g., to cover all possible edges between the nodes, the recall will be equal to 1 but the precision will be low. As both precision and recall are important, we need to model the trade-off between them.

In the GRN inference problem, both precision and recall are important. The aim of GRN inference is to have more correctly predicted edges (high recall) using a shorter list of predictions (high precision). To address both aspects when assessing GRN inference algorithms, we need to have a measurement, such as the F-score or the area under the precision-recall curve (AUPR), that can trade off precision and recall. F-score is calculated from one precision value and one recall value only, given the number of edges included. Since we have more than 10,000 edges in each GRN prediction, we will generate more than 10,000 F-score values with the inclusion of more edges. So in this case, the F-score is dependent on the number of edges included. Therefore, in this study, AUPR is the more suitable measurement which is not in dependence of the number of edges and can address both precision and recall. In addition, AUPR is regarded suitable for reflecting the information retrieval ability of machine learning algorithms (Han et al., 2012).

2.2 Related Work

To achieve more accurate GRN ensemble learning, it is necessary to gain some insight about individual GRN inference algorithms. They fall into several categories, each characterized by its underlying methodology. De Smet & Marchal (2010) categorizes methods by how they deal with the under-determination problem. The categories are direct (Faith et al., 2007) versus module-based (Joshi et al., 2009) methods, supervised (Mordelet & Vert, 2008; Ernst et al., 2008) versus unsupervised (Faith et al., 2007) methods, integrative (Sabatti & James, 2006) versus non-integrative (Joshi et al., 2009) methods. Heyer et al. (1999) categorizes methods according to their network model architectures. The categories are information theory models (Margolin et al., 2006), Boolean networks (Bornholdt, 2008), differential and difference equations (Yeung et al., 2002), and Bayesian networks (Vignes et al., 2011). In the DREAM5 challenge¹, 35 individual GRN inference algorithms and 6 commonly used tools are compared in terms of accuracy (Marbach et al., 2012). The included algorithms are divided into six categories based on the descriptions supplied by participating teams. In addition, the methods within each category are sorted by accuracy.

One would like to know which algorithm is the most accurate. However, with the increasing number of GRN inference methods, it becomes common to observe that one algorithm has higher accuracy than several other algorithms on certain datasets. De Smet & Marchal (2010) points out that the assumptions and constraints made by different algorithms already determine the type of interactions that can be found. Marbach et al. (2010) supports this claim by experiments showing that inference algorithms are prone to different types of systematic prediction errors. For example, when predicting the Feed-Forward Loop network motif, regression based methods perform apparently less reliably than mutual information based methods. This implies that the prediction accuracy of single algorithm is dependent on the network structure. Therefore, it is impossible to decide a single most accurate algorithm. The no free lunch theorems of supervised learning (Wolpert & Macready, 1997) also corroborates this phenomenon. This theorem states that any single learner lacks inherent superiority and learning is impossible without a learning bias.

Therefore, to improve the inference accuracy further, one possible way is to integrate the merits of several algorithms using ensemble learning. That is, if each individual algorithm is specialized to solve one part of the problem; and we take several algorithms that specialize in the different parts of the problem; then the problem should be better solved. In this GRN inference problem, algorithms

¹DREAM is an acronym for Dialogue on Reverse Engineering Assessments and Methods. In the DREAM challenge, participating algorithms or methods are assessed under the same experimental settings

from different categories do have complementarity in terms of network prediction accuracy (Marbach et al., 2012). For example, CLR (Faith et al., 2007) and LeMoNe (Joshi et al., 2009) algorithms have the complementarity on the types of inferred interactions. CLR can identify more regulators with fewer known target genes while LeMoNe can identify more global regulators (De Smet & Marchal, 2010). Marbach et al. (2010, 2012) have demonstrated that ensemble prediction is as accurate or more accurate than the best individual algorithm. The following advantages of ensemble learning on GRN inference are also pointed out:

- Robust to poor inference methods (Marbach et al., 2012). The ensemble method assigns more weight to the output of accurate individual algorithms. Inaccurate algorithms have less effect on the ensemble prediction accuracy.
- Robust to diverse network structures (Marbach et al., 2010). When we infer new networks, the structure of the network is not known beforehand. So it is impossible to tell which algorithm can give a more accurate predictions. Ensemble methods can reduce the prediction bias caused by the network structure.
- The performance of the ensemble learning method can be improved by increasing the diversity of the underlying inference methods (Marbach et al., 2012).

More work is needed to adapt ensemble learning to the GRN inference problem, so as to take advantage of the existing algorithms. Initial studies are carried out in the DREAM challenges (Marbach et al., 2010, 2012). In ensemble prediction, the confidence value of each edge is equal to the average rank of the edge among methods. This generic average ranking method gives an as accurate or more accurate prediction as the top individual algorithms. Fioretto & Pontelli (2013) uses constraint programming to combine the output of four GRN inference algorithms. The experiments are conducted on GP-DREAM platform², which shows that constraint programming outperforms generic average ranking. At the same time, we notice that individual algorithms that are based on ensemble learning, e.g., GENIE3 (Huynh-Thi et al., 2010), LockING (Fouchet et al., 2013) yield a higher accuracy than other individual algorithms. However, they have lower accuracy compared with the ensemble method which combines the output of different algorithms.

In this study, a novel GRN inference method based on ensemble learning is introduced. To increase the diversity and quality of the five individual algorithms, we select them from five different categories. We adopt the categorization provided

²“GP-DREAM Network Inference Tools - GenePattern.” 2014. Accessed May 14. <http://dream.broadinstitute.org/gp/pages/index.jsf>.

in (Marbach et al., 2012), where methods are ranked by accuracy in each category. After considering both diversity and prediction quality, ANOVA (Küffner et al., 2012), CLR (Faith et al., 2007), Correlation (Marbach et al., 2012), GENIE3 (Huynh-Thu et al., 2010) and TIGRESS (Haury et al., 2012) are chosen as the five individual algorithms.

An ensemble learning method consists of a set of models and a method to combine them (Brown, 2010). Because each model has its limitations, the combining method should manage the strengths and weaknesses of them and lead to the best overall decision making. In this context, the five individual algorithms predict five gene regulatory networks (models) and a method is needed to combine them. We analyze the strengths and weaknesses of the five algorithms by assessing how well they can predict different types of network motifs. Each edge in the network prediction is associated with a confidence value. As we break down the original network prediction to different motif types, there are no confidence values associated with the motifs. Thus it is difficult to use numerical combining methods such as averaging, and linear combination. It is also difficult to use majority voting as the number of motifs is too large to treat them one by one. Instead, we can take the prediction of each type of network motif as one subproblem and use the divide and conquer approach to solve each subproblem. To implement this idea, the Mixture-of-Experts ensemble learning method is applied.

Chapter 3

Aims and Objectives

3.1 Aims

In this study, we analyze the error profiles of the five GRN inference algorithms. Then we adapt the Mixture-of-Experts architecture at the network motif level to form the ensemble prediction. After that we compare the accuracy of the proposed ensemble method with generic average ranking method and the five individual algorithms by conducting experiments.

3.2 Objectives

According to the research aims, the study can be divided into two phases: training and testing. In the training phase, we decide the weights of each algorithm on predicting each type of the network motifs using training set. In the testing phase, we use the weights decided in the training phase to generate the ensemble prediction and compare it with the predictions from other methods. To be specific, we list the objectives below:

3.2.1 Training phase

- Firstly, we run the five individual algorithms on the training set to predict gene regulatory networks.
- We take each of the predicted networks and the true network, identify different types of network motifs, and group the network motifs by type.
- Compare each of the network motif types in the predicted networks with the same motif type from the true network. Draw the precision-recall curve and compute the AUPR for each combination of algorithms and network motifs.
- For each type of the network motifs, assign weights to the five algorithms according to their AUPR values. A higher weight is assigned to the algorithms with high AUPR.

3.2.2 Testing phase

- Run the five individual algorithms on the testing set. Identify each type of the network motifs from the predicted networks.
- Generate the ensemble prediction using Mixture-of-Experts architecture based on the weights decided in training phase. Generate the ensemble prediction using generic average ranking.
- Compare the GRN inference accuracy among the five individual algorithms, generic average ranking and the proposed ensemble method using AUPR.

Chapter 4

Research Questions

The generic average ranking method used in the DREAM5 challenge is shown to outperform the included individual GRN inference algorithms. We would like to know if the proposed ensemble method can yield a higher accuracy than the generic average ranking method. So we state the following hypothesis:

Hypothesis 1: The proposed ensemble method is as accurate as the generic average ranking method on the testing set.

As the proposed ensemble method is implemented at the network motifs level, the inclusion or exclusion of certain types of network motifs affects the prediction accuracy. We carry out a preliminary study to identify the network motifs in the *in silico* and *in vivo* networks using a motif detection tool. The result shows that there are more than four types of network motifs that appear significantly more frequently in GRN than in random networks, as listed in Table 4.1.

Table 4.1: Results from a preliminary study on identifying network motif types

Motif ID	Motif	p-value
6	Fan-out	0
12	Cascade	1
36	Fan-in	1
38	Feed-Forward Loop	0
46	Motif-46	0
166	Motif-166	0

We would like to know whether the inclusion of more types of network motifs can improve the accuracy of the proposed ensemble method. Since Motif-46 has higher frequency than Motif-166, we state the following hypothesis:

Hypothesis 2: The inclusion of Motif-46 in the network motif analysis does not affect the prediction accuracy of the proposed ensemble method.

The research methodology in this study is controlled experiment. The independent variable is *method* and the dependent variable is *GRN inference accuracy*. The *method* variable has eight levels (groups): the five individual algorithms, the generic average ranking method, the proposed ensemble method using four types of the network motifs, and the proposed ensemble method using five types of the network motifs. The *GRN inference accuracy* is measured by AUPR. Therefore, the experiment conforms to the single-factor (algorithm) multiple-group experiment design.

The proposed ensemble learning method is based on the Mixture-of-Experts architecture. The generic average ranking method is the same as used in the DREAM5 challenge. The experiment consists of two phases: training and testing. In the training phase, the goal is to decide the weights of individual algorithms at the network motif level. In the testing phase, the assigned weights are used to make the ensemble prediction. Then we compare the prediction accuracy (AUPR) among the eight groups using statistical tests.

We divide this chapter into four sections. The first section concerns the adaptation of the Mixture-of-Experts architecture to the GRN inference problem. The second section is about the network motif analysis method. Then in the third section, we focus on introducing the experiment settings, including datasets, measurements, experiment flow, data collection and statistical tests. In the last section, we discuss the validity threats of this study.

5.1 Mixture-of-Experts architecture

The Mixture-of-Experts architecture is widely used to create a combination of learning models (Jacobs et al., 1991). The principle is that, for a given input space, different models are specialized to particular parts of the input. Then the output of different models are integrated. The architecture is shown in Figure 5.1.

Figure 5.1: Mixture-of-Experts architecture

In the Mixture-of-Experts ensemble method, the component experts and the Gating network receive the same input. The output of the experts are their individual learning models. The output of the Gating network is a linear combiner with weights for different experts. In detail, it follows the steps below to give the ensemble prediction:

1. Divide the input space into subproblems;
2. In each subproblem, evaluate and compare the problem solving abilities among the experts. Increase weights for the strong experts and decrease weights for the weak experts;
3. Obtain the output for each subproblem;
4. Integrate the outputs from all subproblems and generate the ensemble output.

From these steps, the Mixture-of-Experts learning architecture can take advantage of the strength of different algorithms (experts). In the case of GRN inference problem, individual algorithms have the complementarity in predicting the four types of network motifs. The prediction of each network motif type can be considered as one subproblem of the entire GRN prediction. Therefore, the GRN inference problem is suitable to apply the Mixture-of-Experts architecture. To

be specific, in the GRN inference problem, the input space is divided into four or five subproblems: prediction of Fan-in motif, Fan-out motif, Cascade motif, Feed-Forward Loop motif and Motif-46 (the fifth subproblem). Each of the five individual algorithms is an expert which has different abilities in solving the four or five subproblems. Depending on the algorithms' prediction accuracy on the subproblems, the gating network assign weight to each combination of algorithms and subproblems. The overall ensemble prediction is given by integrating the output of algorithms according to the weights.

5.2 Network motif analysis – decide weights of experts

It is necessary to understand the input, output and evaluation methods of the GRN inference problem. The input is the microarray dataset. Each row represents one measurement/chip (one instance) and each column represents the expression values of one gene (one attribute). The output is the GRN in the format of a ranked list of directed edges between genes. GRN inference algorithms are required to predict the GRN from the microarray dataset. The performance of algorithms are evaluated against the known network (gold standard). Figure 5.2 below shows the GRN inference problem:

Figure 5.2: The input, datasets, output and evaluation method of GRN inference

The network motif analysis is performed on the predictions of the individual algorithms as well as the gold standard network. To analyze the accuracy of one algorithm on predicting the network motifs, one has to identify all four or five types of network motifs in the networks and separate them by type. However, the number of possible network motifs increases exponentially with the network size. To deal with the complexity, a fast and accurate motif detecting algorithm is needed in this study. Wong et al. (2011) compares the execution time of several state-of-the-art motif detection algorithms and tools: MODA (Omidi et al.,

2009), MFINDER (Kashtan et al., 2004), Grochow (Grochow & Kellis, 2007), and FANMOD (Wernicke & Rasche, 2006). Zuba (2009) compares the accuracy of two most important network motif detection tools: MFINDER and FANMOD. As a result, FANMOD is proved to be unbiased in sampling, have shorter runtime, and able to detect the four types of gene network motifs. So we use FANMOD to detect network motifs in this study. After detecting the network motifs, we write shell and awk code to group the network motifs. Then for each network motif type, we compare the motif predictions from the five individual algorithms with the motifs from the gold standard network. In this way, for each type of network motif, we obtain the precision-recall curves of the five individual algorithms. The area under precision-recall curve (AUPR) is used to evaluate the accuracy of algorithms on predicting each type of network motif. Figure 5.3 shows the process of network motif analysis.

Figure 5.3: Network motif analysis process

5.3 Experiment

As mentioned above, the experiment conforms to the single-factor multiple-group experiment design. The factor is *algorithm*, which has eight groups: ANOVA, CLR, Correlation, GENIE3, TIGRESS, generic average ranking, the proposed ensemble method with four network motif types, and the proposed ensemble method with five network motif types. In this section, we introduce the experiment settings including datasets, measurements, experiment flow, data collection, and statistical tests.

5.3.1 Datasets

We choose two well-recognized datasets in GRN inference research domain for the training and testing respectively. The training set is the *in silico* dataset used in the DREAM5 challenge. It is generated by extracting network modules from

in vivo gene regulatory networks, endowing with detailed dynamic models and adding experimental noises (Schaffter et al., 2011). Since the dataset is simulated, the true GRN network is known for evaluating predictions.

The testing set comes from the Many Microbe Microarrays database (Faith et al., 2008). It contains more than one thousand microarrays collected from different experiments on *Escherichia coli*. The dataset is normalized to make the data generated from different laboratories comparable. RegulonDB version 7.0 (Gama-Castro et al., 2011) is chosen as the standard to validate the network predictions. It contains the experimentally confirmed gene interactions of *Escherichia coli*, which is considered in the DREAM5 challenge as one of the three gold standards for the performance evaluation.

5.3.2 Measurements

In the training phase, AUPR is used to evaluate the accuracy of each algorithm on predicting each type of network motifs. In the testing phase, AUPR is used to compare the prediction accuracy among the eight methods.

5.3.3 Experiment flow

The experiment consists of training and testing phases, as shown in Figure 5.4.

In the training phase, we run the five algorithms on the training set using the GP-DREAM platform. The predictions of the five algorithms serve as the input to the network motif analysis. For each network motif type, we calculate the AUPR values for all the five algorithms, which are used to assign weights to the algorithms. As this study is the first attempt on the network motif level ensemble learning, we try two simple weight assignment methods. The first method is to assign weight 1 to the top two learning algorithms of each network motif type and assign weight 0 to the bottom three algorithms. That is, we only take the top two performers for predicting each type of network motif. The second method is to assign weights to all the five algorithms in proportion to their AUPR values for each motif type. We compare the accuracy of ensemble predictions using the two weight assignment methods. The method which has higher accuracy is used in testing phase.

In the testing phase, the five individual algorithms make predictions on the testing set using the GP-DREAM platform. These predictions are used to generate the ensemble prediction. In detail, each network prediction is analyzed in FANMOD and grouped by motif types. Then, according to the weights of each algorithm decided in the training phase, the ensemble prediction of each motif type is generated. After that, all four or five integrated motif predictions make up the

Figure 5.4: Experiment flow

ensemble prediction.

For example, suppose that the first weight assignment method is more accurate than the second, and GENIE3 and ANOVA algorithms are the top two algorithms in predicting the Fan-in motif. Then in the testing phase, we identify the Fan-in motifs from the predictions of GENIE3 and ANOVA algorithms and combine them to a list of non-duplicated edges. The edges in the list are ranked by their confidence values from high to low. If an edge is predicted by both GENIE3 and ANOVA algorithms, we take the higher confidence value between the two predictions. In the same way, we can generate the ensemble prediction for each other type of network motifs.

After obtaining the ensemble predictions for all four or five motif types, we combine them to one single list of non-duplicated edges as the ensemble prediction. The edges in the list are ranked by their confidence values from high to low. If an edge is included in more than one motif types, we take the highest confidence value among all predictions. Finally, the predictions from the eight groups are evaluated against the RegulonDB gold standard.

5.3.4 Data collection

Throughout the experiment, the following data are collected:

1. Predictions of the five individual algorithms on the training set
2. Network motif analysis results of 1.
3. Precision-recall curves for each combination of the five individual algorithms and the network motif types on the training set
4. AUPR values of the two weight assignment methods on the training set using both four and five network motif types.
5. Predictions of the five individual algorithms on the testing set
6. Network motif analysis results of 4.
7. Precision-recall curves of the eight method groups on the training set and the testing set.

5.3.5 Statistical tests

In this study, we need to perform statistical tests on multiple groups of data. Normality test is carried out first on each sample. If all samples are distributed normally, ANOVA test is used to identify if there exist significant differences among samples. If so, Post-hoc analysis is applied to determine the significant differences between each two samples. If not all samples are distributed normally, Kruskal-Wallis test and Mann-Whitney U test are used to determine the significant differences among samples. We provide detailed explanations below on how we perform these statistical tests in this study.

Shapiro-Wilks normality test is used to test the normality of samples (Devore, 2011). The null hypothesis is: The sample comes from a normally distributed population. The alternative hypothesis is: The sample does not come from a normally distributed population. The significance level is 0.05. If the p-value of the Shapiro-Wilks test on a sample is less than 0.05, we reject the null hypothesis.

The ANOVA (analysis of variance) test is applied to determine the differences among several normally distributed and independent samples (Devore, 2011). In this experiment, we have only a single dependent variable (*GRN inference accuracy*) and one single factor (*algorithm*). So one-way ANOVA can be applied if samples are normally distributed and independent. The null hypothesis is: the means of all samples are equal. The alternative hypothesis is: at least one of the means is different. The level of significance is 0.05. Firstly, we calculate the (x,y) degrees of freedom of the sample: x denotes the degrees of freedom between

groups and y denotes the degrees of freedom within groups. Then we can obtain the critical F-ratio from the table of critical values for the F distribution using (x,y) degrees of freedom. After that we perform the ANOVA test and get the value of F. If the value of F is equal to or larger than the critical F-value ($p < 0.05$), then we reject the null hypothesis.

If the null hypothesis of the ANOVA test is rejected, we want to know among the means of samples, which of these are significantly different from the others. Post-hoc analysis (Tukey's HSD test) can perform this analysis (Devore, 2011). For every two groups (levels), the null hypothesis is: the means of the two groups are equal. The alternative hypothesis is: the means of the two groups are different. The level of significance is 0.05. If the p-value between two groups (levels) is less than 0.05, we reject the null hypothesis.

If the Shapiro-Wilks normality test shows that not all samples are normally distributed, we use Kruskal-Wallis one-way analysis of variance by ranks (Devore, 2011) instead of ANOVA test. The Kruskal-Wallis test is a non-parametric test to compare multiple independent samples and determine differences among them. The null hypothesis is: all samples have the same median. The alternative hypothesis is: not all samples have the same median. The level of significance is 0.05. If the p-value given by this test is less than 0.05, we reject the null hypothesis. If the Kruskal-Wallis test shows that there are significant differences among samples, Mann-Whitney U test can identify where the differences occur (Devore, 2011).

5.4 Validity threats

In this section, we discuss the internal validity, external validity, construct validity and statistical validity of this study. We identify the validity threats and analyze how much they can influence the validity of the conclusions.

5.4.1 Internal validity

Internal validity concerns to which extent the causal conclusion is valid. In other words, to which extent does the systematic error (bias) influence the conclusions. Firstly, the network structure influences the accuracy of individual algorithms as well as ensemble method. However, it affects the latter much less than the former. The aim of the ensemble learning is to minimize the effects of the network structure. Although the prediction accuracy of individual algorithms is dependent on the network structure, the error profile at the network motif level is intrinsic to each algorithm itself (Marbach et al., 2010). Therefore, we can almost avoid the impact of the network structure on testing the hypotheses.

Secondly, in the training phase, only the first 10,000 edges of the prediction from each algorithm are included for the network motif analysis. The computational power of experimental computer does not allow to include all edges (100,000) within the limited project duration. As the edges are ranked from high confidence value to low confidence value, the first 10,000 edges are comparatively more accurate and less noisy.

Thirdly, the FANMOD network motifs detection tool may bring bias. Although it is relatively accurate, it has a trade-off between speed up and the cost of missing potential motifs (Zuba, 2009). As it is impossible to use full enumeration in large gene networks, the small portion of undetected motifs is acceptable given a much faster speed. In this study, the FANMOD tool is used to identify network motifs from the networks that have the same number of edges. So the effect of FANMOD tool is neglected.

5.4.2 External validity

The external validity concerns how well the conclusions of the study can generalize to other situations. In the experiments we can use known networks to evaluate the ensemble predictions. In external cases, we need to use network inference methods to infer unknown networks or networks with very limited knowledge. Still, the proposed strategy can provide a more accurate ensemble prediction, if all individual algorithms are properly trained and assigned weights accordingly. The weights should not be dependent on the network structure. Instead, they should reflect the learning abilities of the individual algorithms.

5.4.3 Construct validity

Construct validity concerns if the measurements and tests can fulfill the aims of the study. In this study, the measurements (mainly AUPR) can reflect the accuracy of the methods. Thus they are capable for testing the hypotheses. The datasets for training and testing are commonly used in this research area. However, the sampling of individual algorithms are not repeated due to the limited project duration.

5.4.4 Statistical validity

Statistical validity concerns the correctness of statistical tests. In this study, we carry out statistical tests logically. We make sure that the statistical tests match the features of the samples. However, there are limitations in the validation datasets. Although RegulonDB is considered the gold standard for validation, it

does not contain all the gene interactions. So some correctly predicted interactions may be considered incorrect because they are not present in the RegulonDB. This can affect the accuracy of statistical tests. It is a common statistical threat in GRN inference research. More genetic high-throughput experiments will supplement the RegulonDB dataset in the future. So this validity threat is not discussed in this study.

6.1 Training phase

In this phase, we aim to analyze the accuracy of the five individual algorithms on predicting the network motifs. We use FANMOD tool to analyze the network predictions of the five individual algorithms. As discussed above, we include another type of motif (Motif-46) that appears significantly more frequently than in random network. So we take into account all five types of network motifs as five subproblems in the Mixture-of-Experts architecture. To summarize, the motifs involved in this study are listed in Figure 6.1. The adjacency matrix is used to describe the edges in the motif, direction is from row to column (Wernicke & Rasche, 2006). The motif ID is the decimal representation of the adjacency matrix.

Motif ID	Motif	Adjacency Matrix	Motif structure
6	Fan-out	0 0 0 0 0 0 1 1 0
12	Cascade	0 0 0 0 0 1 1 0 0
36	Fan-in	0 0 0 1 0 0 1 0 0
38	Feed-Forward Loop	0 0 0 1 0 0 1 1 0
46	Motif-46	0 0 0 1 0 1 1 1 0

Figure 6.1: Illustration of the five types of network motifs involved in this study

For each type of network motif, the motifs predicted by the five individual algorithms are evaluated against the motifs from the true network. We draw the

precision-recall curves and calculate the AUPR values for every combination of the algorithms and the network motifs. The AUPR values are listed in Table 6.1. The "-" in the table means "no correct prediction". The precision-recall curves for each type of network motif are listed in Appendix I.

Table 6.1: AUPR values of the algorithms on predicting network motifs

AUPR	Motif 6	Motif 12	Motif 36	Motif 38	Motif 46
ANOVA	0.0068	2.1377e-04	3.4001e-04	3.0672e-04	4.8232e-04
CLR	0.0124	-	5.9734e-04	-	4.5265e-04
Correlation	0.0101	-	3.0354e-04	-	1.5649e-04
GENIE3	0.0192	3.5030e-05	0.0016	3.8372e-04	0.0011
TIGRESS	0.0149	5.8089e-05	9.6379e-04	3.4272e-04	5.4388e-04

Based on Table 6.1, we compare the accuracy of the two weight assignment methods. In both methods, the weights are assigned to the edges that constitute the network motifs, not the motifs themselves. In the first methods, the top two performers of each motif type are assigned weight 1, with rest three performers assigned weight 0. The weights of algorithms are shown in Table 6.2.

Table 6.2: The weights of algorithms using the first weight assignment method

AUPR	Motif 6	Motif 12	Motif 36	Motif 38	Motif 46
ANOVA	0	1	0	0	0
CLR	0	-	0	-	0
Correlation	0	-	0	-	0
GENIE3	1	0	1	1	1
TIGRESS	1	1	1	1	1

In the second method, we assign the weights in proportion to the AUPR values. Then we scale the weights for each type of network motif to the range from 0.5 to 1. We choose this range instead of the range from 0 to 1. The reason is that the algorithms with lower accuracy also have certain amount of correct predictions with high confidence values. If we assign very small weights to these algorithms, the correctly predicted edges can not be included. The weights of the algorithms are shown in Table 6.3.

We compare the accuracy of the ensemble predictions generated by using the two weight assignment methods including both four and five network motif types. The precision-recall curves and AUPR values of the two methods are listed in Figure

Table 6.3: The weights of algorithms using the second weight assignment method

Algorithms	Motif 6	Motif 12	Motif 36	Motif 38	Motif 46
ANOVA	0.50000	0.99936	0.51385	0.49818	0.67284
CLR	0.66000	-	0.61282	-	0.65712
GENIE3	1.00000	0.50008	0.99846	0.99818	1.00000
Correlation	0.54500		0.49982	-	0.50026
TIGRESS	0.78500	0.56449	0.75377	0.73195	0.70544

6.2 and Table 6.4. *Ensemble4* and *Ensemble5* denote the ensemble predictions with four and five network motif types using the first weight assignment method. *Ensemble4w* and *Ensemble5w* denote the ensemble predictions with four and five network motif types using the second weight assignment method.

Figure 6.2: The precision-recall curves of the two weight assignment methods

Apparently, the first weight assignment method gives more accurate predictions than the second method. So we adopt the first method in the following experiments. Before going to the testing phase, we compare the AUPR values of the eight methods on the training set. The precision-recall curves and AUPR values

Table 6.4: The AUPR values of the two weight assignment methods

Methods	Ensemble4w	Ensemble5w	Ensemble4	Ensemble5
AUPR	0.2388	0.2408	0.2890	0.2921

are shown in Figure 6.3 and Table 6.5.

Figure 6.3: The precision-recall curves of the eight methods on *in silico* datasetTable 6.5: AUPR of the eight methods on *in silico* dataset

Methods	ANOVA	CLR	Correlation	GENIE3	TIGRESS	Ensemble4	Ensemble5	Avgrank
AUPR	0.2066	0.1813	0.1558	0.2534	0.2665	0.2890	0.2921	0.2864

In Figure 6.3 and Table 6.5, the *Ensemble4* and *Ensemble5* have the same meanings as introduced above. The *Avgrank* represents the generic average ranking method. These three labels are also used in the following tables and figures. The results in Figure 6.3 and Table 6.5 suggest that:

- Both the generic average ranking method and the proposed method have higher accuracy than the individual algorithms on *in silico* dataset.
- The proposed ensemble method is more accurate than the generic average ranking method on *in silico* dataset

- The proposed ensemble method with five network motif types is more accurate than the proposed ensemble method with four network motif types on *in silico* dataset.

6.2 Testing phase

In this phase, we aim to compare the prediction accuracy of the eight methods on the testing set using precision-recall curves. Using the proposed ensemble method and the first 10,000 edges of the five individual predictions on the testing set, we generate the ensemble prediction in the form of a ranked list of edges. The list includes 19,049 non-duplicated edges ranked from high confidence value to low confidence value. As stated above, for one single edge in the ensemble prediction, its confidence value is equal to the highest confidence value of this edge among all individual predictions.

The ensemble prediction has more edges than the involved individual predictions. To compare the eight methods equally, we take the first 19,000 edges from the five individual predictions, the generic average ranking prediction, the prediction of the proposed ensemble method with four motif types, and the prediction of the proposed ensemble method with five motif types. Then we take ten samples from each method: the first 10,000 edges, 11,000 edges, ..., 19,000 edges. After that, we calculate the AUPR for each sample, as shown by Table 6.6.

Table 6.6: AUPR values of the eight methods (sample size = 10)

Edges	ANOVA	CLR	Correlation	GENIE3	TIGRESS	Ensemble4	Ensemble5	Avgrank
10000	0.0081	0.0129	0.0048	0.0121	0.0089	0.0081	0.0093	0.0073
11000	0.0082	0.0129	0.0048	0.0121	0.0090	0.0082	0.0093	0.0074
12000	0.0083	0.0130	0.0048	0.0122	0.0090	0.0083	0.0095	0.0075
13000	0.0084	0.0130	0.0049	0.0122	0.0091	0.0084	0.0095	0.0076
14000	0.0085	0.0130	0.0049	0.0123	0.0092	0.0085	0.0097	0.0077
15000	0.0086	0.0131	0.0050	0.0124	0.0092	0.0086	0.0098	0.0078
16000	0.0087	0.0131	0.0050	0.0125	0.0093	0.0087	0.0098	0.0079
17000	0.0088	0.0131	0.0050	0.0125	0.0093	0.0087	0.0099	0.0079
18000	0.0088	0.0132	0.0050	0.0126	0.0093	0.0088	0.0100	0.0079
19000	0.0089	0.0132	0.0051	0.0126	0.0094	0.0088	0.0101	0.0080

After that we carry out statistical tests on Table 6.6. The data in the table has a single dependent variable (*GRN inference accuracy (AUPR)*) and one single factor (*method*). The factor has eight groups (levels). The size of each sample is ten and the total sample size is eighty. We perform the statistical tests in the order stated in section 5.3. The precision-recall curves of the eight methods on the testing set are listed in Appendix II.

6.3 Results from statistical tests

Firstly, we carry out Shapiro-Wilks normality test on each sample in Table 6.6. The results are shown in Table 6.7.

Table 6.7: Results of Shapiro-Wilks normality test on the eight samples

Methods	ANOVA	CLR	Correlation	GENIE3	TIGRESS	Ensemble4	Ensemble5	Avgrank
p-value	0.6697	0.2582	0.1105	0.2063	0.4872	0.4025	0.5394	0.3798

From Table 6.7 we conclude that all samples are distributed normally ($p > 0.05$). Therefore, we use ANOVA test to determine if significant differences exist among samples. The degrees of freedom is (7,72). We perform the ANOVA test and the result is shown in Table 6.8.

Table 6.8: ANOVA test result on the eight methods

Source of variation	df	SS	MS	F	Pr(>F)
Between-groups	7	0.000464	6.63E-05	1458	<2e-16
Within-groups	72	3.3E-06	5.00E-08		
Total	80	0.000467			

Table 6.8 shows that the F value of the ANOVA test is significantly larger than the critical F value. Therefore, we conclude that there are significant differences among samples. Then we carry out the Tukey's HSD test to make all pairwise comparisons. A box plot is also generated to help identify the differences. The results of the Tukey's HSD test and the box plot are presented in Table 6.9 and Figure 6.4. In Table 6.9, the *difference* column gives the differences in the observed means. The *lower* and *upper* columns give the upper and lower end point of the interval. The *p-value* has been adjusted for the multiple comparisons.

Table 6.9: Tukey's HSD test result among the eight methods

Pairwise methods	difference	lower	upper	p-value adjusted
avgrank-ANOVA	-0.00083	-0.001127574	-0.000532426	0.0000000
CLR-ANOVA	0.00452	0.004222426	0.004817574	0.0000000
Correlation-ANOVA	-0.0036	-0.003897574	-0.003302426	0.0000000
ensemble4-ANOVA	-0.00002	-0.000317574	0.000277574	0.9999990
ensemble5-ANOVA	0.00116	0.000862426	0.001457574	0.0000000
GENIE3-ANOVA	0.00382	0.003522426	0.004117574	0.0000000
TIGRESS-ANOVA	0.00064	0.000342426	0.000937574	0.0000001
CLR-avgrank	0.00535	0.005052426	0.005647574	0.0000000
Correlation-avgrank	-0.00277	-0.003067574	-0.002472426	0.0000000
ensemble4-avgrank	0.00081	0.000512426	0.001107574	0.0000000
ensemble5-avgrank	0.00199	0.001692426	0.002287574	0.0000000
GENIE3-avgrank	0.00465	0.004352426	0.004947574	0.0000000
TIGRESS-avgrank	0.00147	0.001172426	0.001767574	0.0000000
Correlation-CLR	-0.00812	-0.008417574	-0.007822426	0.0000000
ensemble4-CLR	-0.00454	-0.004837574	-0.004242426	0.0000000
ensemble5-CLR	-0.00336	-0.003657574	-0.003062426	0.0000000
GENIE3-CLR	-0.0007	-0.000997574	-0.000402426	0.0000000
TIGRESS-CLR	-0.00388	-0.004177574	-0.003582426	0.0000000
ensemble4-Correlation	0.00358	0.003282426	0.003877574	0.0000000
ensemble5-Correlation	0.00476	0.004462426	0.005057574	0.0000000
GENIE3-Correlation	0.00742	0.007122426	0.007717574	0.0000000
TIGRESS-Correlation	0.00424	0.003942426	0.004537574	0.0000000
ensemble5-ensemble4	0.00118	0.000882426	0.001477574	0.0000000
GENIE3-ensemble4	0.00384	0.003542426	0.004137574	0.0000000
TIGRESS-ensemble4	0.00066	0.000362426	0.000957574	0.0000000
GENIE3-ensemble5	0.00266	0.002362426	0.002957574	0.0000000
TIGRESS-ensemble5	-0.00052	-0.000817574	-0.000222426	0.0000176
TIGRESS-GENIE3	-0.00318	-0.003477574	-0.002882426	0.0000000

Figure 6.4: Box plot of AUPR values among the eight methods

From the results of statistical tests, we can draw conclusions on the two hypotheses of this study. This chapter presents the conclusions and the causal analysis of the conclusions.

For hypothesis 1, the proposed ensemble method has significantly higher AUPR than the generic average ranking method. Both ensemble method with four motif types and five motif types have yielded a higher accuracy than the generic average ranking method. The generic average ranking method calculates the confidence value of each edge as the average rank among the included algorithms without differentiating the prediction accuracy of them. So weak algorithms influence the overall accuracy. In addition, the generic average ranking method treats all edges equally without identifying edges that belong to different types of network motifs. So further improvement of the accuracy is limited. In our method, we break down the entire network into four or five network motif types and analyze the accuracy of the individual algorithms on predicting each type of the network motifs. Then we select two most accurate algorithms for each motif type. And finally we integrate the predictions of every type of motif to constitute the ensemble prediction.

By conducting the network motif analysis, we do find that certain algorithms have better accuracy to predict certain types of network motifs, or not able to predict certain types of network motifs, as shown in Table 6.1. For example, CLR and Correlation algorithms are not able to predict Cascade and Feed-Forward Loop motifs; ANOVA method has the highest accuracy in predicting the Cascade motif; GENIE3 has the highest accuracy in predicting the Fan-in motif.

For hypothesis 2, the ensemble predictions generated by integrating five types of network motifs have significantly higher AUPR than the ensemble predictions generated by integrating four types of network motifs. One possible reason is that the input space is divided into more subproblems so each subproblem is more specialized and easier to solve. The other possible reason is that the fifth type of network motif (Motif-46) has significant biological meaning. So when the prediction of Motif-46 is considered as a subproblem along with the other four

subproblems in the Mixture-of-Experts architecture, the AUPR value of the ensemble prediction increases significantly.

However, we are not sure how the weighting mechanisms influence the accuracy of the ensemble predictions. To analyze this, we need to carry out more experiments using different weighting mechanisms and algorithms in order to draw further conclusions.

GRN inference is considered an important and challenging problem in bioinformatics. This study proposes a novel ensemble learning method to improve GRN inference accuracy. The aim is to develop an ensemble method that can take more advantage of the complementarity among the existing algorithms. The ensemble learning method investigates the complementarity at the network motif level and integrate individual predictions at the network motif level. The experiments have shown that the proposed ensemble method is more accurate than the generic average ranking method used in the the DREAM5 challenge. This conclusion is also consistent with other studies in bioinformatics where ensemble methods yield higher accuracy than the individual methods, e.g., predicting gene functions (Guan et al., 2008), predicting species distributions (Araújo & New, 2007), and predicting RNA secondary structures (Ding et al., 2005).

From the results of the experiments, we notice that the proposed ensemble method ranks the first on the *in silico* dataset and the third on the *Escherichia coli* dataset. GENIE3 and CLR algorithms have higher AUPR than the proposed ensemble method on the *Escherichia coli* dataset. We have two possible interpretations for this. Firstly, the *in silico* dataset used in the DREAM5 challenge is generated by GeneNetWeaver software (Schaffter et al., 2011). It generates network structures by extracting modules from known biological interaction networks and adding experimental noises (Marbach et al., 2009). This method is shown to be biologically plausible. But it is difficult for *in silico* data to reflect the complexity of the real gene expression data (Hecker et al., 2009). So the differences between the *in silico* dataset and the real dataset affect the rank of the proposed ensemble method. Secondly, not all algorithms have complementarity with each other. For example, GENIE3 algorithm uses ensemble learning method (random forest), so it has relatively less bias and higher accuracy. It is chosen to predict four types of network motifs in the testing phase. However, the algorithms with low prediction accuracy are not chosen to predict any network motif types, e.g., Correlation method. So the features of individual algorithms may also influence the accuracy of the ensemble prediction.

It is worth noting that including the fifth network motif in the proposed ensemble method increases the prediction accuracy significantly. So it suggests that the Motif-46 found in this study has substantial biological meaning, although the frequency of motif 46 is lower than the other four types of motifs in the network. In addition, the network structure of Motif-46 is common in cell and gene regulatory pathways. For example, in the well-known JAK-STAT cell signaling pathway, two genes can regulate one common gene while interacting with each other.

This study is conducted under time and computational power limitations. Only the first 10,000 edges from each individual algorithm’s prediction are taken as the input to the ensemble method. On one hand, the number of edges is enough to compare the proposed ensemble method with the generic average ranking method by taking ten samples. On the other hand, it makes the comparisons between the ensemble methods and the individual algorithms inaccurate. According to the results from the DREAM5 challenge, out of all 35 individual inference methods, the generic average ranking method ranks the first on the *in silico* dataset, and the third on the *Escherichia coli* dataset. We select five most accurate algorithms from each category in the DREAM5 challenge. It turns out that the proposed ensemble method ranks the first on the *in silico* dataset and the third on the *Escherichia coli* dataset. In addition, on both datasets the proposed method has significantly higher AUPR than the generic average ranking method. So the experiment results can suggest that the proposed ensemble method is among the top performers in GRN inference.

To conclude, this study shows how Mixture-of-Experts architecture can be applied to the GRN inference problem. What’s more, the architecture can be adapted to solve other biological problems as well. The key is to divide the whole problem into characteristic subproblems; find the “experts” (methods) for each subproblem; and integrate the output from all experts. Therefore, the Mixture-of-Experts architecture has strong potential to integrate the strengths of the individual methods and make more accurate predictions. With the exponential growth of biological data, more and more algorithms are proposed with their strengths and weaknesses. The complexity of biological problems and the large amount of proposed methods indicate the increasing need of the ensemble learning.

Chapter 9

Conclusions and Future Work

This study has shown that the Mixture-of-Experts architecture has strong potential in improving GRN inference accuracy. It can investigate the strengths and weaknesses of individual algorithms at the network motif level. Then it generates the ensemble prediction from the network motif level integration.

We compare the prediction accuracy among the proposed ensemble method, the generic average ranking method used in the DREAM5 challenge and the five individual algorithms by conducting experiments. Results show that the proposed ensemble learning method is significantly more accurate than the generic average ranking under our experimental settings. In addition, a new type of network motif (Motif-46) is found which can significantly increase the prediction accuracy of the proposed ensemble learning method.

Above all, this study provides a potential learning strategy that can be applied to a variety of biological problems. This strategy can take advantage of the existing algorithms or methods in certain problem context and give a more comprehensive solution to the problem.

The future work has two folds: conducting this study further and extending this study to other problems in bioinformatics. Firstly, the experiments in this study can be further improved as follows:

- We will involve more computational power. So in the network motif analysis step, we can include more edges from the individual predictions so as to perform more accurate comparisons among the individual algorithms.
- Analyze how different datasets, weighting mechanisms, and algorithm combinations influence the prediction accuracy of the ensemble method. After that, find a method to adjust the weights of the individual algorithms more precisely to maximize the prediction accuracy.
- Include other GRN inference algorithms that use different underlying models, such as Bayesian networks. In our experiments, we find that algorithms based on Bayesian networks do not have high overall accuracy, but they

are more accurate than other individual algorithms on predicting Cascade motifs.

Secondly, the principle of the Mixture-of-Experts architecture can be applied to other biological problems. One possible way of doing this is to build a Decision Support System for these problems. The Decision Support System should include state-of-the-art algorithms and representative training and testing datasets for every included problem. Then we can decide the weights of the algorithms based on certain criteria, such as accuracy. The development of this Decision Support System can have two main benefits:

- It can provide a platform for researchers to compare the state-of-the-art algorithms, find the shortcomings of the algorithms, come up with new solutions, and evaluate the new solutions.
- It can provide biologists a much more convenient way to apply computational approach to accelerate their research. For example, if they need to solve one practical problem, e.g., cancer classification, they can refer to the "classification category" in the system, choose corresponding algorithms, upload their data, and obtain an ensemble prediction, solution or analysis results.

References

- Alam, I., Dress, A., Rehmsmeier, M., & Fuellen, G. (2004). Comparative homology agreement search: an effective combination of homology-search methods. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(38), 13814–13819.
- Almudevar, A., Klebanov, L. B., Qiu, X., Salzman, P., & Yakovlev, A. Y. (2006). Utility of correlation measures in analysis of gene expression. *NeuroRx: the journal of the American Society for Experimental NeuroTherapeutics*, *3*(3), 384–395.
- Altay, G., & Emmert-Streib, F. (2010). Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics*, *26*(14), 1738–1744.
- Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in ecology & evolution*, *22*(1), 42–47.
- Baldi, P., & Brunak, S. (2001). *Bioinformatics: The machine learning approach*. MIT Press.
- Bornholdt, S. (2008). Boolean network models of cellular regulation: prospects and limitations. *Journal of The Royal Society Interface*, *5*(1), S85–S94.
- Brown, G. (2010). Ensemble learning. In *Encyclopedia of machine learning* (p. 312–320). Springer.
- Bubendorf, L. (2001). High-throughput microarray technologies: From genomics to clinics. *European Urology*, *40*(2), 231–238.
- Buck, M. J., & Lieb, J. D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, *83*(3), 349–360.
- Butte, A. J., & Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Pac symp biocomput* (Vol. 5, pp. 418–429).

- Cho, S.-B., & Won, H.-H. (2003). Machine learning in DNA microarray analysis for cancer classification. In (pp. 189–198). Darlinghurst, Australia, Australia: Australian Computer Society, Inc.
- De Smet, R., & Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8(10), 717–729.
- Devore, J. (2011). *Probability and statistics for engineering and the sciences*. Cengage Learning.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems* (pp. 1–15). Springer.
- Ding, Y. E., Chan, C. Y., & Lawrence, C. E. (2005). RNA secondary structure prediction by centroids in a boltzmann weighted ensemble. *Rna*, 11(8), 1157–1166.
- Ernst, J., Beg, Q. K., Kay, K. A., Balázsi, G., Oltvai, Z. N., & Bar-Joseph, Z. (2008). A semi-supervised method for predicting transcription factor–gene interactions in escherichia coli. *PLoS Computational Biology*, 4(3), 1–14.
- Faith, J. J., Driscoll, M. E., Fusaro, V. A., Cosgrove, E. J., Hayete, B., Juhn, F. S., et al. (2008). Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic acids research*, 36(Database issue), D866–870.
- Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., et al. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1), e8.
- Fioretto, F., & Pontelli, E. (2013). Constraint programming in community-based gene regulatory network inference. In A. Gupta & T. A. Henzinger (Eds.), *Computational methods in systems biology* (pp. 135–149). Springer Berlin Heidelberg.
- Fouchet, A., Delosme, J.-M., & D’Alché-Buc, F. (2013). Gene regulatory network inference using ensembles of local multiple kernel models.
- Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muñiz-Rascado, L., Solano-Lira, H., et al. (2011). RegulonDB version 7.0: transcriptional regulation of escherichia coli k-12 integrated within genetic sensory response units (sensor units). *Nucleic acids research*, 39(Database issue), D98–105.

- Grochow, J. A., & Kellis, M. (2007). Network motif discovery using subgraph enumeration and symmetry-breaking. In *Proceedings of the 11th annual international conference on research in computational molecular biology* (pp. 92–106). Berlin, Heidelberg: Springer-Verlag.
- Guan, Y., Myers, C. L., Hess, D. C., Barutcuoglu, Z., Caudy, A., & Troyanskaya, O. G. (2008). Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome biology*, 9(1), S3.
- Hache, H., Lehrach, H., & Herwig, R. (2009). Reverse engineering of gene regulatory networks: A comparative study. *EURASIP J. Bioinformatics and Systems Biology*, 2009, 8:1–8:12.
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining : concepts and techniques*. Waltham, Mass.: Elsevier/Morgan Kaufmann.
- Haury, A.-C., Mordelet, F., Vera-Licona, P., & Vert, J.-P. (2012). TIGRESS: trustful inference of gene REgulation using stability selection. *BMC Systems Biology*, 6(1), 145.
- Hecker, M., Lambeck, S., Toepfer, S., Someren, E. van, & Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models-a review. *Bio Systems*, 96(1), 86–103.
- Heyer, L. J., Kruglyak, S., & Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. *Genome research*, 9(11), 1106–1115.
- Honkela, A., Girardot, C., Gustafson, E. H., Liu, Y.-H., Furlong, E. E. M., Lawrence, N. D., et al. (2010). Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences*.
- Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9).
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1), 79–87.
- Jaskowiak, P. A., Campello, R. J., & Costa Filho, I. G. (2013). Proximity measures for clustering gene expression microarray data: a validation methodology and a comparative analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 10(4), 845–857.

- Joshi, A., Smet, R. D., Marchal, K., Peer, Y. V. d., & Michoel, T. (2009). Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics*, 25(4), 490–496.
- Kashtan, N., Itzkovitz, S., Milo, R., & Alon, U. (2004). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11), 1746–1758.
- Kolde, R., Laur, S., Adler, P., & Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4), 573–580.
- Küffner, R., Petri, T., Tavakkolkhah, P., Windhager, L., & Zimmer, R. (2012). Inferring gene regulatory networks by ANOVA. *Bioinformatics*, bts143.
- Lalonde, E., Ha, K. C. H., Wang, Z., Bemmo, A., Kleinman, C. L., Kwan, T., et al. (2011). RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome research*, 21(4), 545–554.
- Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., et al. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1), 86–112.
- Lemmens, K., Bie, T. D., Dhollander, T., Keersmaecker, S. C. D., Thijs, I. M., Schoofs, G., et al. (2009). DISTILLER: a data integration framework to reveal condition dependency of complex regulons in escherichia coli. *Genome Biology*, 10(3), 27–37.
- Li, Z., Li, P., Krishnan, A., & Liu, J. (2011). Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis. *Bioinformatics*, btr454.
- MacIsaac, K. D., Wang, T., Gordon, D. B., Gifford, D. K., Stormo, G. D., & Fraenkel, E. (2006). An improved map of conserved regulatory sites for saccharomyces cerevisiae. *BMC Bioinformatics*, 7(1), 113.
- Maetschke, S. R., Madhamshettiwar, P. B., Davis, M. J., & Ragan, M. A. (2013). Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in Bioinformatics*, bbt034.
- Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8), 796–804.
- Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., & Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*, 107(14), 6286–6291.

- Marbach, D., Schaffter, T., Mattiussi, C., & Floreano, D. (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16(2), 229–239.
- Margolin, A. A., & Califano, A. (2007). Theory and limitations of genetic network inference from microarray data. *Annals of the New York Academy of Sciences*, 1115, 51–72.
- Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(1), S7.
- Meyer, P. E., Kontos, K., Lafitte, F., & Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology*, 2007.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, 298(5594), 824–827.
- Mitchell, T. M. (1997). *Machine learning* (1st ed.). New York, NY, USA: McGraw-Hill, Inc.
- Mordelet, F., & Vert, J.-P. (2008). SIRENE: supervised inference of regulatory networks. *Bioinformatics*, 24(16), i76–i82.
- Omidi, S., Schreiber, F., & Masoudi-Nejad, A. (2009). MODA: an efficient algorithm for network motif discovery in biological networks. *Genes & genetic systems*, 84(5), 385–395.
- Sabatti, C., & James, G. M. (2006). Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22(6), 739–746.
- Schaffter, T., Marbach, D., & Floreano, D. (2011). GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16), 2263–2270.
- Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1), 64–68.
- Tan, A. C., & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification. *Applied bioinformatics*, 2(3), S75–83.

- Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R., & Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS Computational Biology*, 3(6), e116.
- Vignes, M., Vandel, J., Allouche, D., Ramadan-Alban, N., Cierco-Ayrolles, C., Schiex, T., et al. (2011). Gene regulatory network reconstruction using bayesian networks, the dantzig selector, the lasso and their meta-analysis. *PLoS ONE*, 6(12).
- Wernicke, S., & Rasche, F. (2006). FANMOD: a tool for fast network motif detection. *Bioinformatics*, 22(9), 1152–1153.
- Wolpert, D., & Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82.
- Wong, E., Baur, B., Quader, S., & Huang, C.-H. (2011). Biological network motif detection: principles and practice. *Briefings in Bioinformatics*, bbr033.
- Wu, M., & Chan, C. (2011). Learning transcriptional regulation on a genome scale: a theoretical analysis based on gene expression data. *Briefings in Bioinformatics*, bbr029.
- Yeung, M. K. S., Tegnér, J., & Collins, J. J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences*, 99(9), 6163–6168.
- Zuba, M. (2009). A comparative study of network motif detection tools.

Appendix I

Figure 1: P-R curves of Fan-out motif

Figure 2: P-R curves of Cascade motif

Figure 3: P-R curves of Fan-in motif

Figure 4: P-R curves of FFL motif

Figure 5: P-R curves of Motif-46

Appendix II

Figure 6: P-R curves with 10000 edges

Figure 7: P-R curves with 11000 edges

Figure 8: P-R curves with 12000 edges

Figure 9: P-R curves with 13000 edges

Figure 10: P-R curves with 14000 edges

Figure 11: P-R curves with 15000 edges

Figure 12: P-R curves with 16000 edges

Figure 13: P-R curves with 17000 edges

Figure 14: P-R curves with 18000 edges

Figure 15: P-R curves with 19000 edges