Summary of Gene regulatory network inference using PLS-based methods

Authors of original paper:
Shun Guo
Quinshan Jiang
Lifei Chen
and Dongui Guo
12/28/2016

Author of this summary:
Sams Khan

February 21, 2017

Abstract

This paper is a summary of an existing paper by Shun Guo, Quinshan Jiang, Lifei Chen and Dongui Guo. In the abstract the authors divided up parts of the entire paper based on the main titles and briefly explained the structure of each section. They also included key words that are most important throughout the whole paper.

1 Background:

The paper goes in to describe how the problem they are presented with is very important in the field of Bioinformatics. It explains that because the amount of data obtained that can be corresponded to gene expression profiles of thousands of genes are so large, introducing different ways of being able to infer the GRN using reverse engineering is being proposed.

Many different types of statistical dependency methods have been applied to try to solve the problem, but it turns out there is always something lacking. Some methods that are applied that solve one part of the problem but still always have room for improvement, so researchers introduce other methods that improve the weakness, but still has a weakness elsewhere, and so on and so forth. Each method is always better than the preceding and manages to solve the previous methods weaknesses.

Other methods based on probabilistic graphical models have also been used to solve such issues, which comes with their own shortcomings, no feedback loops. A dynamic version of such methods was used to solve the issue, but it could only handle time-series expression data. Plus learning and implementing these methods are extremely difficult not only from a computational perspective but also from a theoretical one.

Other ensemble methods such as GENIE3, ENNET, TIGRESS and NIMEFI were used, but they also had their weaknesses. GENIE3 is not well understood theoretically, Tigress was the same, ENNET had extremely high computational costs when applied to high dimensional data, and NIMEFI connects GENIE3 and some other methods such as Elastic net under one framework, but it increases the model of uncertainties by a large number.

In this paper, the authors focus on introducing a new method based on PLS (Partial Lease Squares).The method uses a combination of PLS and a statistical technique to refine the predictions by taking into account the hub regulatory gene. This method has been used on 2 well known datasets and the benchmarks produced are very competitive to that of state-of-the-art methods used in the gene-reconstruction challenges.

2 Methods

2.1 problem and definition

They focused on directed topology of GRNs using gene expression data. The input data was the same as used against the previously mentioned methods, and they used a general framework for the GRN inference problem.

Most methods give us a ranking list from least to most confident, and then the end user can use this to explore different threshold levels. In this paper they decided to just focus on the ranking list itself.

2.2 Network inference with feature selection methods

For each gene that they wish to target they must determine the subset of the genes which directly influence the target gene. They define the regulators containing expression values in n experimental conditions finally as regression function.

2.3 GRN inference with PLS-based ensemble methods

Other PLS methods has been used by other authors to solve the problem at had such as TotalPLS and KernelPLS, only problem is that we don't know how many candidate regulatory genes are needed to provide a good model, which is why they used a PLS-based ensemble method.

2.4 Feature selection with PLS-based method

They created a function so that they can use w and u in such a way that Q=Yu can carry as much info as possible thus they coined this criterion function:

$$\begin{cases} max \quad J(u,v) = \frac{(v^T\Sigma_{XY}u)^2}{v^Tv\cdot u^Tu} \\ s.t.v^Tv = u^Tu = 1. \end{cases} \tag{1}$$

Since SIMPLS is slightly superior and computationally efficient than NIMPALS, they used.

Refining the inferred regulatory network

After they use the statistical technique to refine their inferred regulatory network, they make an adjacency matrix, where the strength of a gene is represented and the genes are scored based on how much they impact the target genes. Here is the adjacency matrix:

$$W(i,:) = W(i.:) * \sigma_i^2, \forall i \in \{1, 2, ..., p\} \tag{2}$$

The adjacency matrix has a row which contains relative scores that correspond to different target genes, so if a gene regulates many different target genes, the variance of the row is increased.

Parameter settings

Here the authors emphasize what type of configurations they used in order to carry out their experiment.

They also go on to explain why they used the configurations they did, how there were advantages and also disadvantages.

They go on to explain how they applied the parameters o the DREAM5 and DREAM4 datasets, and what their thoughts where when they applied this before running the test. They utilized many different specifications based on recommendations from previous methods such as GENIE3.

Computational complexity

In this section the authors try to give us a peek into how complex the computation behind their method is, they define different parts of their adjacency matrix in terms of variables. They then go on to describe how parts of these variables behave.

Results

Here the authors emphasize the challenge of being able to test these GRNs, they also talk about how previous tests had issues. They then go onto explain how their test works and how they would compare it to previously known regulations.

The Authors emphasize on which types of datasets they used to compare the results of their method. They also explain which ones they didn't use and why, they also show a table of the datasets.

The paper then talks about how some of the datasets they used were also used to test previous experiments of methods such as TIGRESS and NIMEFI. The authors evaluated the accuracy of the other methods by using the overall score metric and then tells us what the score metric means.

2.5 Performance evaluation

They compare the performance of their method with 5 very famous methods, they go on to discuss which implementations of the other methods they used and where they got them. They also discuss which parameters they used in those implementations.

2.6 *Performance on the DREAM4 multifactorial dataset*

They explain how the silico size 100 multifactorial challenge of DREAM4 worked, what it contained and how they used it. They display a chart of benchmarks of other methods on this challenge.

The paper then goes onto explain what their focus was when using this dataset. It also states and explains how certain scores were important to the paper and how the performance went.

2.7 *Influence of parameters*

Here they present us with multiple box-plots comparing the results of different methods on the DREAM4 dataset. They explain in detail how the configuration changes in specific methods influenced their performance and they compared it with other's using these box-plots.

2.8 *Performance on the DREAM5 datasets*

This paragraph goes on to explain how the DREAM5 dataset is structured and where they originated from. This also emphasizes the reason behind this challenge and where they were used.

In the next paragraph they explain a chart where they laid out the scores of their method compared to the other methods and go on to explain how to deduce the chart itself.

Here the authors talk more about the results of their own method and how other methods achieved their results. They talk about different scores and how they compare to the different networks in the challenge.

Since the amount of regulatory genes on DREAM5 datasets are so much bigger than DREAM4 it was harder for them to figure out the amount of candidate genes.They talk about what configuration they used and what they observed.

They explain how most of the networks performed poorly on the other 2 networks other than network 1, because assuming information about the other 2 are very vague. Whereas on network 1, they provide enough info to be able to reverse engineer everything.

2.9 *CPU time*

The Authors talk about which implementations they used of which method. Also, the specs of the machine they used, they also include which methods they didn't compare to and also which methods works best with which dataset.

3 Conclusions

Here they are minimally summarizing the entire paper and concluding what they have talked about to in this paper.

They highlight the main topic and reason behind why they went with the approach mentioned. Which is in case they have many regulatory target genes.

References

[1] Guo, S. et al., 2016. Gene regulatory network inference using PLS-based methods. BMC Bioinformatics, 17(1), p.545. Available at: http://dx.doi.org/10.1186/s12859-016-1398-6.