Reviews and Comments on Paper 158

Submission details
Add comment
Add new review
Request review
Revise review 2
Edit note

Paper information

Paper: Benjamin A Logsdon and Jason G Mezey.
Cyclic Regulatory Network Reconstruction from Genetic Perturbations
Current decision: **REJECT** (reject)

Summary of received reviews and comments

Reviews superseded by other reviews are shown in the grey color in the table. All times are GMT.

	date	PC member	subreviewer	score	confidence
Review 1	Feb 2	Florian Markowetz	Xin Wang	**1**	2
Review 2	Feb 21	Josh Stuart		**1**	2
Review 3	Feb 22	Michael Brent	Brian Haynes	**1**	3
Comment 1	Feb 23	Eric Xing			
Comment 2	Mar 1	Josh Stuart			
Comment 3	Mar 1	Josh Stuart			
Review 4	Mar 8	Hanah Margalit		**-1**	2

Review 1

PC member: Florian Markowetz
Reviewer: Xin Wang
Overall rating: **1** (weak accept)
Confidence: **2** (medium)

Summary:
With the motivation of seeking the minimal class of perturbation architectures providing the maximum resolution for any given regulatory relationships, this paper derives three theorems and infers the sufficient set of independent perturbations theoretically. Furthermore, based on conclusion of Theorem 3, the authors designed a three-step algorithm—EXPLoRE to integrate gene expression phenotype with cis-eQTL genotype data to generate large regulatory networks. Both simulated and HapMap data were used to demonstrate the power of EXPLoRE.

Clarity of contributions:
In general, this paper is very well organised and written in pretty fluent English. The only concern is that it would be clearer if conclusion part can be separated from result in the Abstract.
The contributions of this paper were clearly claimed in "Results" of "Abstract", "Introduction" and the beginning of "Conclusion" parts.

Interest:
As a big topic, regulatory network identification is not novel. Recent years have seen many methods and their applications in different scenarios aiming at this question, especially probabilistic graphical models. However, the major innovation of this paper is that the authors focus on the issue of finding a sufficient set of independent perturbations providing maximum resolution for the identification of directed cyclic networks, avoiding the difficulty of identifying Markov blanket or separating set for each node in the directed graph. Step by step, this paper then derives a few theorems to support their ideas and based on the third theorem, they establish a novel algorithm to do regulatory network identification.

Method:
Both demonstrations based on simulation and biological data were presented in this paper.
Although the authors discuss the prediction power of EXPLoRE algorithm across 4 regulatory networks with different sizes, the privilege of this algorithm over other methods are not mentioned. It would be more convincing if more comparisons between EXPLoRE and other methods are discussed in terms of prediction accuracy, computational efficiency, etc.

Review:

The best expression regulatory networks predicted based on data from HapMap project were also presented. It was mentioned that "..., as determined by cross-validation, with $\lambda=0.7$". Would it be better if the authors give more details about how they did "cross-validation", and why they chose 0.7 as the threshold of λ? As for the interpretation of the prediction result, the authors said that "this inferred network should be interpreted with caution", but the advantage of using this algorithm in this particular biological question is not very convincing.

Big Problems:
1. In INTRODUCTION part, the review of probabilistic graphical model applications in regulatory network is far from
complete. Apparently, a lot of important models were missed, such as many methods about directed acyclic networks,
and Andreas Wagner's Bioinformatics paper.
2. Apparently, eQTL was introduced by Jansen and Nap in 2001. But in section 1 paragraph 2 sentence 5, the
authors cited Rockman's paper in 2008.
3. Prediction power of EXPLoRE algorithm was clearly demonstrated in simulation study; however, in HapMap data
analysis, the accuracy of prediction was not mentioned.
4. This paper didn't discuss whether or not the EXPLoRE algorithm can incorporate prior knowledge to improve
prediction accuracy.

Small Problems:
1. In section 6 paragraph 3 sentence 1, space of the word "demonstrating" was not well aligned.
2. In section 5 paragraph 2 sentence 2,"Friedman et al. (2008)" should be "(Friedman et al., 2008)".
3. In section 5 paragraph 2 sentence 4, "described below" should be clearly specified.
4. Many of "cis-" throughout the manuscript were not italicized.
5. Why use "ANALYSES" in "SIMULATION ANALYSES", but "ANALYSIS" in "HAPMAP NETWORK ANALYSIS"?
6. Personally, it would be better not use "THE EXPLoRE ALGORITHM" as the title of section 5.
7. In Fig. 5., more explanation about figure in figure legend would be better. E.g. the meaning of "red circles",
arrows, and the names with black blocks.
8. The inter-paragraph space between 2nd and 3rd paragraphs of section 7 should be adjusted.
9. Space between Fig. 4 and Fig. 5 is too large.
10. In section 5 "Step 3", "... to which the are cis-eQTL. ..." may be "... to which they are cis-eQTL. ...".

PC only:
Time: Feb 2, 13:25

Review 2

PC member: Josh Stuart
Overall rating: **1** (weak accept)
Confidence: **2** (medium)

Review:

The authors develop theory for identifying the minimal combination of perturbations to use in order to narrow down a list of potential network models that explain a set of expression observations. They also have developed a method to infer a regulatory network using eQTL data.

Major Comments

The biggest shortcoming of the work is that the authors do not compare their method to any competing methods for inferring regulatory interactions. What are the 1) naive methods in this area and 2) the state-of-the-art methods? The authors should provide a brief literature review in the introduction and introduce at least one leading method they can compare EXPLoRE's performance against. If its absolutely true that this represents the first such algorithm to connect cis-eQTLs to infer regulatory nets (as claimed by the authors in the Discussion), then they should at least implement a naive approach as a baseline comparison.

The main results are not convincing. First, while much theory was elaborated upon, I'm still left wondering what is the limiting set of perturbations needed to elucidate a regulatory network? It is not clear how the theoretical results shed light on this even thought the authors claim it is one of their main results. Second, the main networks output by the EXPLoRE method are speculative. The authors should bring an independent validation to support the network identified in Figure 5. Can they support any of the interactions predicted?

Great detail has been given for the derivation of both the theory and development of the authors' EXPLoRE algorithm. To fully evaluate the exposition, one needs a background in linear algebra and some modern multivariate statistics (e.g. knowledge of the lasso and related techniques). I could see no fundamental flaws in the arguments, but this may be due to my own cursory knowledge of the statistical theory on which much of the methods rely. I have to admit that I had trouble following the theoretical part of the methods section. I think this mainly has to do with the fact that much of it lacks motivation or a high level description of the goals of what the authors are trying to do. I understand that they want to show how to reduce a set of equivalent models by determining the number of perturbations needed. However, I have no idea how Section 3 accomplishes this with the three Theorems and lemma provided. My guess is that these methods need to be rewritten so that a general bioinformatics audience can follow the arguments. I found myself lost in many places because I lack some esoteric knowledge about what

particular matrices are called.

Minor Comments

Page 2. Section 2. Para 2. The arguments in this paragraph are rather vacuous. They justify the use of a linear model by effectively arguming that if a linear model holds for the biological system then their linear model is appropriate. This is just silly.

Page 3. Sec 3. It is difficult to follow this section because there is not enough motivation. Also, definitions are not provide for some terms such as "model matrix" or "full precision matrix." Perhaps I don't have the background to readily bring these concepts to mind. The authors need to explain in English what exactly the relationship between cappa-gamma and cappa-sigma means. Just telling me that it is a system of second order polynomials is useless. I need more motivation to understand why they wish to draw a link between a graph that encodes the network structure and the precision matrix (of course, it would help to know what a "precision matrix" is in this context too!).

Page 3. Sec 3. The authors need to motivate why they provide a definition of equivalence here. Why is it necessary? Does it need to be proven or has that been done in Pearl 2000?

PC only:

Time: Feb 21, 04:54

Review 3

PC member: Michael Brent
Reviewer: Brian Haynes
Overall rating: **1** (weak accept)
Confidence: **3** (high)

In this paper the authors address the problem of identifying a regulatory network consisting of interactions of two types: 1) gene to gene and 2) loci to gene. Genotype data in conjunction with gene expression data is applied to this problem of network identification. The problem of network identification is formalized as a linear structural equation model, wherein, a matrix, Lambda, of gene to gene interactions and a matrix B of genotype to
gene interactions encodes the network structure. The assumptions made by the authors here is that the perturbations do not move the
system out of the range of linear behavior, and that regulatory

Review:

effects are additive and do not exhibit combinatoric logic.

The authors proceed to rigourously prove that if there exists at least one independent genotypic perturbation per expression phenotype, then there exists a unique equivalence class barring the direction of cycles in the network. The SEM is solved by first deriving the undirected conditional dependence structure, and then resolving the directionality of this undirected network using the perturbation data and LASSO regression. Two strong assumptions are made here, namely: B is a diagonal matrix, meaning
each perturbation has a direct effect on a single gene, and secondly, each gene is influenced by at least one perturbation. Both of these assumptions may not be realistic in reality, but the authors do acknowledge this.

The method is evaluated on artificially generated sets of data and HapMap data, but unfortunately no comparisons are made to alternative methods that make use of genotype data. The method appears to do reasonably well, but the improvements yielded by restricting the equivalence classes using the perturbation data are not overwhelmingly evident from the figures. A better comparison
could have been made between the accuracy at Step 2 of the method
versus Step 3 (Step 3 being the novel component). For large sample sizes, the novel component of the method appears to improve
accuracy, but for small sample sizes it appears to have a negative effect, which is not addressed by the authors. Comparison could have been more easily done if both stages were shown in the same plot or, if area under the curve statistics had been given. Additionally, RoC statistics are not as informative as Precision Recall measures for evaluating accuracy on sparse networks (see Margolin et. al 2006). The evaluation on the HapMap data was interesting but difficult to interpret considering no strong conclusions were drawn, and no known interactions were used to validate the method's performance.

PC only:

Time: Feb 22, 15:13

Comment 1

By: Eric Xing

Comment:

All give a boardlinae score. Any strong advocates? Please discuss!

eric

Time:	Feb 23, 22:16

Comment 2

By:	Josh Stuart
Comment:	Eric, Even though we all gave borderline accepts, when you put together our reviews it looks like all 3 of us agree on some points. We agree that this paper has potential and that the theoretical results are interesting. However, all three of us agree that there are significant shortcomings with how the authors evaluated the approach. Brian and I want to see a comparison to competing methods; Florian wants to see evaluation on HapMap (real) data and not just simulations. This amount of work is probably too much for the authors to revise for ISMB. So, although I gave a borderline accept previously, I could not lean toward borderline reject. --Josh
Time:	Mar 1, 20:54

Comment 3

By:	Josh Stuart
Comment:	Boy, talk about a confusing final sentence in my last comment. I meant to say "... I could *now* lean toward a borderline reject." --Josh
Time:	Mar 1, 23:03

Review 4

PC member:	Hanah Margalit
Overall rating:	**-1** (weak reject)
Confidence:	**2** (medium)
Review:	Summary of PC discussion: All three reviewers ranked the paper as boderline. They all agreed that this paper has potential and that the theoretical results are interesting. However, all three reviewers agreed that there are significant shortcomings with the way the authors evaluated the approach. A comparison to competing methods is missing, and evaluation on HapMap (real) data and not just simulations is highly

recommended. The paper needs a substantial revision before it can be considered for ISMB.

PC only:

Time: Mar 8, 15:31

Add comment

Please type your comments in the text area below. You comments will only be visible to PC members having access to this paper. It will not be sent to the authors.

Add Comment