Review for

"An empirical Bayes approach to estimating dynamic models of co-regulated gene expression"

Jiaxin Hu

This work proposes a new similarity metric, Bayesian lead-lag R^2 (LLR^2), based on the temporal expressions of two genes. The LLR^2 describes the goodness of fit of two gene dynamic ODE models (1) (2) that share a common time-dependent expression regulation factor $p(t)$. Bayesian approaches are adopted to combine the known gene interaction information in the ODE model fitting to obtain a more biologically-meaningful similarity metric.

The paper is well-organized and the writing is clear. Incorporating the prior biological knowledge and data information in the dynamic similarity via Bayesian approach is also an interesting idea. Though, there are some methodology concerns make me on the fence of this paper.

Comments:

1. (Motivation) The motivation of Bayesian LLR^2 is unclear for me. Based on Section 2.2, Bayesian LLR^2 is motivated to address the occurrence of false positive gene pairs with spuriously high LLR^2. The main reason for the spuriously high LLR^2 is the high correlation between the response ($m_A(t)$) and the covariates independent with gene B (the third and fourth columns in X). Because the R^2 of model (4) not only represents the variation explained by the co-regulated factors between A and B but also includes the variation explained by marginal factors independent with other genes. This issue always exists if marginal factors $\int m_A(s)ds, t, 1$ are in the model, no matter how we estimate the coefficients. Hence, the false positvity issue is not a convincing motivation to propose Bayesian approach.

   Instead of Bayesian LLR^2, I think the R^2 of sub-model (18), LLR^2_{other}, in the Section 3.5 is a more direct way to solve false positivity issue. However, LLR^2_{other} does not gain much attention in main analysis and is not applied in main data analyses.

2. (Prior) The setting of prior mean β_0 in page 9 seems unreasonable for me. Specifically, when $W_{ij} = 1$, the setting $c_1 = c_2 = 1, c_3 = 0$ seems inappropriate. In model (4), we have $c_2 = c_1\kappa_B$ and $c_3 = -\kappa_A$. One would expect κ_A and κ_B are at the same scale since they have same interpretation in marginal models (1) and (2). Then, when we assume $c_3 = 0$, the parameter c_2 should also be close to 0 despite $c_1 = 1$. Hence, the statement in page 8-9, "$c_2 = 1$, since c_2 represents a parameter that is proportional to c_1", seems incorrect. Consequently, author may re-consider the construction of prior distribution.

   In addition, there is no sensitive analysis for the incorrect prior knowledge. For example, if the gene interaction in the previous database is not applicable in current data, will the method still give satisfactory result? Also, one may concern that whether the new gene

interaction discoveries will be dominant or covered by the previous known pathways due to the informative Bayesian prior. More discussion on the self-correction ability of proposed method with wrong or non-applicable prior knowledge will be helpful.

3. (Computation) The computation results are not reported. Bayesian approach is known to be slower than other method in computation. It would be helpful to report the time and computation resource required by the method.

References