Author's response to reviews

Title: Supervised Learning Methods in Modeling of CD4+ T Cell Heterogeneity

Authors:

Pinyi Lu (lpy0526@vt.edu)
Vida Abedi (abedi@vbi.vt.edu)
Yongguo Mei (mygzxy@gmail.com)
Raquel Hontecillas (rmagarzo@vbi.vt.edu)
Stefan Hoops (shoops@vbi.vt.edu)
Adria Carbo (adria.carbo10@gmail.com)
Josep Bassaganya-Riera (jbassaga@vt.edu)

Version: 2
Date: 17 May 2015

Author's response to reviews: see over
Dear Editors,

We would like to thank the reviewers and the editorial board for taking the time to review our manuscript (212402241154702) entitled “Supervised Learning Methods in Modeling of CD4+ T Cell Heterogeneity”. The review was very helpful to us, and the recommended changes have allowed us to provide better documentation and clarity. We have considered the reviewers' comments.

Our point-by-point response is as follows:

Reviewers' comments:

----------------------------------------

Reviewer 1:

Some details about the data and analysis were unclear to me from the description:

1. Typically, Random Forest has one output variable. How were the multiple output cytokines handled in RF?

Response:
For each output cytokine, we built a Random Forest model, respectively. Since there are totally five outputs, IL17, RORgt INFγ, Tbet, and FOXP3, five Random Forest models were created. This detail was added to the manuscript.

-------------------------------------------------------------------------------------

2. When output cytokines are predicted, does it not matter what the baseline levels are for prediction? Is it presumed they are near zero because of cell type?

Response:
Since all the training data are normalized to the range of [0, 1], all the output cytokines are expected in that range as well.

-------------------------------------------------------------------------------------

3. What time point is chosen for determining steady-state levels?

Response: We used COPASI to perform the parameter scan for creating the entire dataset. COPASI has function of Steady-State analysis. There is no time associated with a steady state. In theory this state is only reached at T that tends to infinity. COPASI estimates the distance from the steady-state by inverting the Jacobian of the system. Once this is sufficiently small, a steady-state is found

-------------------------------------------------------------------------------------

4. In Table 1, it appears that four simulated instances were used to train and test the learners with different combinations of initial cytokine levels. This seems like a small number. Is this a problem when doing the bootstrap bagging in Random Forests?

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Response: We chose INFγ, IL12, IL6, and TGFβ as inputs based on the experimental discovery, which shows that those four cytokines are most important for CD4+ T Cell Heterogeneity. With current size of the dataset, we have no problem building models using Random Forests algorithms.

5. In real data, what is a realistic number of training instances? How independent are they?

Response: In this study, the training instances are assumed to be independent. But based on the limited number of input parameters, all possible combinations were considered to generate the training dataset. There is very limited experimental data available for building the training dataset. Thus, we tried to use simulated data to train the models and use the models for prediction.

6. The noise added to the simulated data seems low. Are these noise levels realistic?

Response: We have added higher level noise in this study. Since there are many sources of noise, the real noise levels may vary in different environment. Here, we used two levels of artificial noise to shows our models have ability to deal slight variability. Furthermore, we applied a relatively low noise level because we have no indication of any species in our model to have a low copy number and therefore be subject to higher levels of fluctuation and noise. The text was clarified.

7. Some of the paper suggests that the supervised method can predict cell differentiation. Perhaps it can indirectly? I think a more accurate characterization might be that the method predicts the levels of cytokines that help characterize some cell types. I think some clarification of this distinction would be helpful.

Response: We agree with that. We have applied the predicted levels of cytokines to help cell type classification, which is a more accurate characterization. The results were published in the following paper:

Minor: Table 1 column headings say Training and Testing, but do you mean input and output cytokines?

Response: Table 1 column headings have been changed from Training and Testing to Input and Output.
Spelling: “as oppose to” should be “as opposed to.”

Response:
“as oppose to” has been replaced by “as opposed to”.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Word choice: “which approved that our models are capable.” Instead of “approved,” a more appropriate word might be “demonstrated.”

Response:
“approved” has been replaced by “demonstrated”.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Reviewer 2:

The paper is clear, well written and interesting.

I do have amjoe comment: How is cross validation done? it is not clear to me and I believe that the authors should better clarify this issue.

Response: We have performed 10-fold cross validations for all models. The results are shown in Table 7.