The Malaria DREAM challenge: team TPOT's sub-challenge 2 write-up

This manuscript (permalink) was automatically generated from trang1618/plasmodium-falciparum@bcc4569 on September 6, 2019.

Authors

- **Alena Orlenko**
  0000-0003-1757-293X · desmidium · desmidium
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Trang T. Le**
  0000-0003-3737-6565 · trang1618 · trang1618
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Weixuan Fu**
  0000-0002-6434-5468 · weixuanfu · weixuanfu
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Jason H. Moore**
  0000-0002-5015-1099 · EpistasisLab · moorejh
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

Abstract

Challenge link: https://www.synapse.org/#!Synapse:syn16924919/wiki/583955

Objective: Predict the parasite clearance rate of malaria parasite isolates based on in vitro transcriptional profiles.

Data source: [1]

Methods

Data harmonization

For this subchallenge, the training dataset's characteristics were very different compared to those of the test set. Importantly, training set contains *in vivo* transcription as described in [1] while test set contains *in vitro* transcription. Other differences in data collection, synchronization, microarray platforms, preprocessing steps, etc. have been detailed on the challenge website. We first impute the missing data points using the KNN imputation method [2] with k = 10 via the fancyimpute Python package. Next, to adjust for batch effects between the two datasets, we apply the *ComBat* algorithm [3] from the *sva* R package [4] on the transcription data. We assess the effect of the adjustment by examining the principal component analysis (PCA) plots on the raw and processed data.

Transcriptional feature selection

Using the STatistical Inference Relief (STIR) algorithm [5], we selected the genes with adjusted STIR P values < 0.05. Specifically, with the MultiSURF neighborhood, STIR nearest-neighbors to select transcriptional features whose association with an outcome may be due to epistasis.

Automated machine learning for model training

The Tree-based Pipeline Optimization Tool (TPOT) is a Python Automated Machine Learning (AutoML) tool that uses genetic programming to optimize machine learning pipelines for analyzing biomedical data [6]. However, like other AutoML tools, TPOT currently requires high computational expense when analyzing high dimensional data. Therefore, even though we already manually select features prior to executing TPOT, we also limit the final classifier of TPOT pipelines to LinearSVC to focus TPOT's effort on identifying the sequence of preprocessors (e.g., recursive feature elimination or function transformer). We used balanced accuracy as our scoring function.

Test sample selection

While there is one single sample per isolate in the training set, there are eight samples for each of the 32 isolates in the test set (most have 2 biological replicates, 2 time points: 6 hours and 24 hours post invasion, perturbed with 5nM DHA (DHA) or perturbed with DMSO (UT)). Therefore, to obtain one prediction per isolate, we need to take caution in selecting which test samples to predict on. First, because the majority of the training samples are estimated at 18 hours post invasion (hpi), we selected test samples at 24 hpi (closer to 18 hpi compared to 6 hpi). Second, we analyze the developmental stage of each biological replicate at seperate timepoints. We hypothesize that the group of isolates with asexual stage distribution close to that of the training set will yield the most accurate prediction for each isolate. While the stage of the parasites in the training set was determined by Mok et al. according to [1], transcriptional profiles in the test set were compared against the 3D7 sample from [7]. The Methods section in Mok et al.'s supplementary material

indicates that the two asexual stage estimation techniques are similar [8], but we carefully investigate the distributions of parasite stages in two datasets before making any specific selection.

Non-genetic features consideration

Because all samples from the test data were collected from the Thailand-Myanmar border, for now, we ignore the `Country` feature. In the future, however, perhaps we can place more weights on the training samples that are geographically closer to this region. We also ignore the `Kmeans.Grp` feature that is cluster groups corresponding to three types of transcription profile based on parasite developmental stage.

Availability

Detailed preprocessing, modeling and analysis code for this study is available at https://github.com/EpistasisLab/malaria-challenge.

Results

Batch effect adjustment

Before batch adjustment, the two datasets are clearly separated in the first two principal component dimensions (Fig. 1A). After being adjusted for batch effect, this dataset-specific effect is less evident (Fig. 1B). The amount of variance explained in each component also seems to be more balanced.

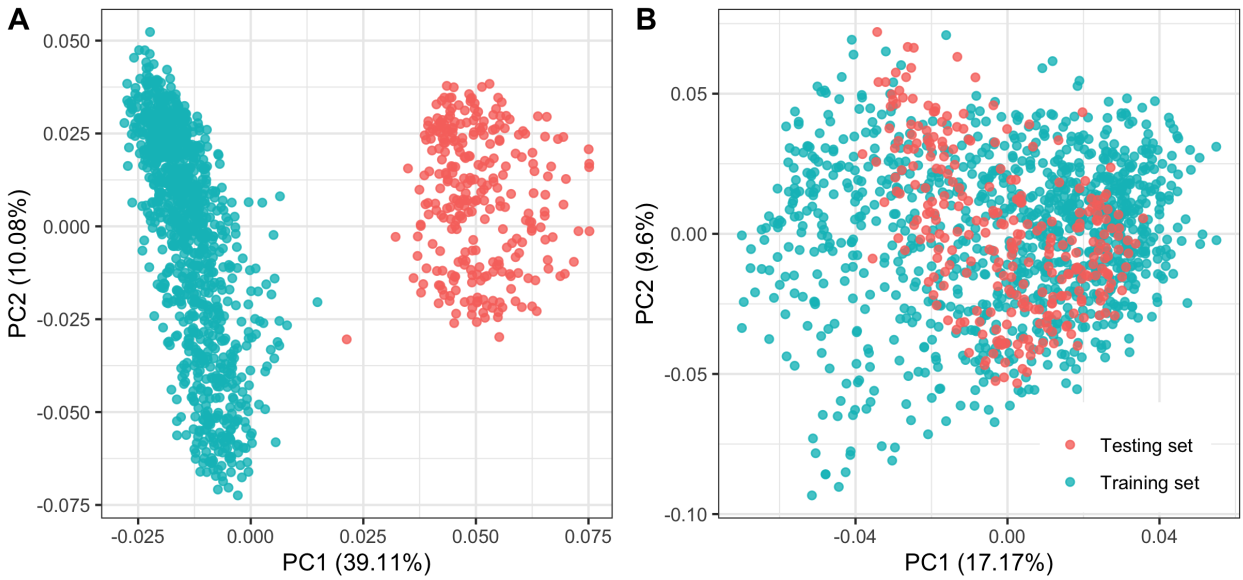

Figure 1: Principal component analysis plots before (A) and after (B) adjusting for batch effects

STIR feature selection and TPOT cross-validated balanced accuracy

The training dataset consists of 1043 in vivo parasite isolates, each with 4952 transcriptomic features, out of which STIR selects 1068.

We first run 50 replicates of the full configuration TPOT for 1000 generations (or 24 hours - whichever happened first) with the pipeline population size of 1000. We have identified that the majority of pipelines contain Logistic Regression or Linear SVC as a classifier. We further run TPOT for the second time using the same simulation parameters but reducing configuration to only two classifiers in the list: Logistic Regression and Linear SVC. To evaluate a pipeline performance on the training dataset we

used balanced accuracy implemented in TPOT as a custom metric. Balanced accuracy computes each class accuracy on per class base and in the binary case is equal to the arithmetic mean of sensitivity and specificity. Using balanced accuracy is highly advised in case of an imbalanced dataset to avoid inflated performance estimates. We chose the top 20 best performing pipelines to be used for the testing dataset predictions and then averaged these predictions to get the final score.

Test sample selection

In general, the parasite developmental stages in the training set are smaller than those in the test set (Fig. 2). Because of this difference in distribution, we could not use stage of parasites as guideline to select the test samples. Instead, at 24 hpi, we decided to compute the average of the two lowest probabilities per isolate (across two types of treatments and two biological replicates). In other words, setting our goal to be detecting isolates with SLOW parasite clearance rate (SLOW = 1), we want to decrease the false negative rate (SLOW isolates predicted as FAST) by sacrificing some false positives (FAST isolates predicted as SLOW).

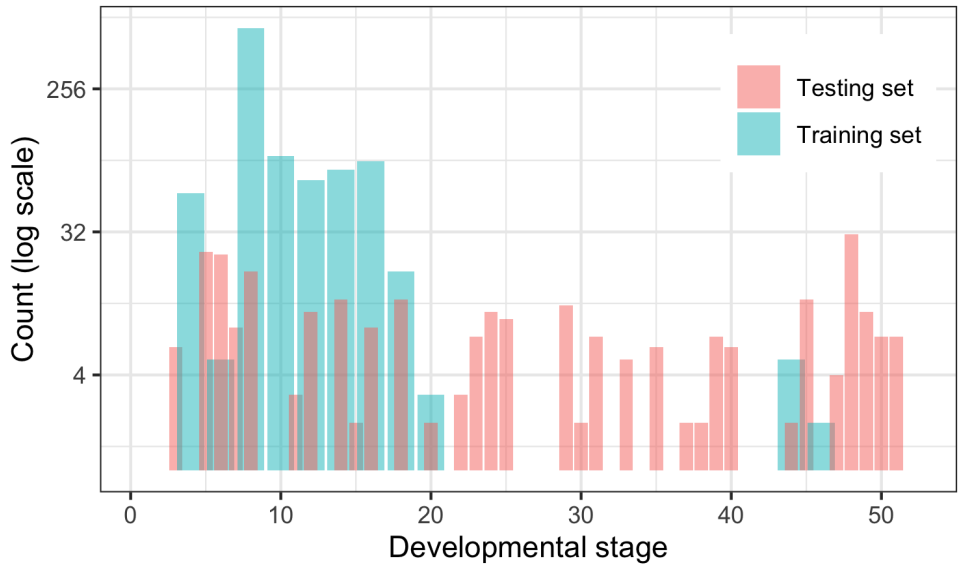

Figure 2: Developmental stages in training set and testing set

Conclusion

Even with autoML, at the moment, we still need to be clever at preprocessing and integrating the data for this type of problems where the test set is completely independent of the training set. To reduce computation time, further preprocessing (feature selection for dimensionality reduction) also needed before feeding the data into the autoML tool. Additional non-genetic information is important. Even though we did not find an efficient way to harmonize the asexual stage feature from the two datasets, because there are a lot of training samples, we should probably have removed the 30 samples with developmental stages of larger than 17 out of the training dataset.

It is surprising to us that there is not a separate holdout set for this challenge. In an attempt to prevent overfitting, the organizers only released the ranking instead of the absolute score of each team's performance. However, a team can still assess the trajectory of their performance compared to other teams and tweak their method in that direction.

References

1. **Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance**
S. Mok, E. A. Ashley, P. E. Ferreira, L. Zhu, Z. Lin, T. Yeo, K. Chotivanich, M. Imwong, S. Pukrittayakamee, M. Dhorda, ... Z. Bozdech
Science (2014-12-11) https://doi.org/f3ph2f
DOI: 10.1126/science.1260403 · PMID: 25502316 · PMCID: PMC5642863

2. **Data mining with R: learning with case studies**
Luís Torgo
CRC Press (2011)
ISBN: 9781439810187

3. **Adjusting batch effects in microarray expression data using empirical Bayes methods**
W. Evan Johnson, Cheng Li, Ariel Rabinovic
Biostatistics (2006-04-21) https://doi.org/dsf386
DOI: 10.1093/biostatistics/kxj037 · PMID: 16632515

4. **The sva package for removing batch effects and other unwanted variation in high-throughput experiments**
Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, John D. Storey
Bioinformatics (2012-01-17) https://doi.org/fx9kw2
DOI: 10.1093/bioinformatics/bts034 · PMID: 22257669 · PMCID: PMC3307112

5. **STatistical Inference Relief (STIR) feature selection**
Trang T Le, Ryan J Urbanowicz, Jason H Moore, Brett A McKinney
Bioinformatics (2018-09-18) https://doi.org/gfbdwx
DOI: 10.1093/bioinformatics/bty788 · PMID: 30239600 · PMCID: PMC6477983

6. **Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science**
Randal S. Olson, Nathan Bartley, Ryan J. Urbanowicz, Jason H. Moore
Proceedings of the 2016 on Genetic and Evolutionary Computation Conference - GECCO '16 (2016) https://doi.org/gfgqv2
DOI: 10.1145/2908812.2908918

7. **The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum**
Zbynek Bozdech, Manuel Llinás, Brian Lee Pulliam, Edith D Wong, Jingchun Zhu, Joseph L DeRisi
PLoS Biology (2003-08-18) https://doi.org/cd68hh
DOI: 10.1371/journal.pbio.0000005 · PMID: 12929205 · PMCID: PMC176545

8. **Quantitative Time-course Profiling of Parasite and Host Cell Proteins in the Human Malaria ParasitePlasmodium falciparum**
Bernardo Javier Foth, Neng Zhang, Balbir Kaur Chaal, Siu Kwan Sze, Peter Rainer Preiser, Zbynek Bozdech
Molecular & Cellular Proteomics (2011-05-10) https://doi.org/cwb67f
DOI: 10.1074/mcp.m110.006411 · PMID: 21558492 · PMCID: PMC3149090