Predicting Molecular Initiating Events from High Throughput Transcriptomic Screening using Machine Learning

J. L. Bundy1, R. Judson1, A.J. Williams1, C. Grulke1, I. Shah1, L. Shah1, L. Everett1
1) USEPA, Research Triangle Park, NC

The views expressed in this presentation are those of the author(s) and do not necessarily represent the views or policies of the Agency.

Introduction

Goal: U.S. EPA is developing new approach methodologies (NAMs) to identify potential toxicity pathways. Some NAMs are using mechanistic data, such as high throughputs (HTTs), to connect adverse effects with molecular initiating events (MIEs). To meet this challenge, we are developing a machine learning based methodology that integrates HTTs and chemical-MIEs to predict MIEs.

Key points:

- Integrated LINC3 L1000 CMAP gene expression compendium [1]
- Used RefChemDB database of chemical-protein target interactions [2]
- Trained binary classifiers on integrated data sets with the following parameters:
  - 51 MIEs
  - 3 Feature Sets
  - 6 Classification Algorithms
  - 2 Cell Types
  - MCF7 profiles
  - PC3 profiles

Method:

1. Chemical treatments associated with LINCS L1000 profiles were matched to EPA Chemical-MIE labels and a large gene expression compendium (Figure 1).

2. All genes

3. Feature sets

4. Classification algorithms

5. Cell types

6. Training sets

7. Validation sets

8. Models

9. Predictive accuracy

Figure 1. Data processing and classifier training workflow

The prediction of chemical biodata at the level of MIEs required the integration of chemical-MIE labels and a large gene expression compendium (Figure 1).

Classifier Training Overview

Selection of Training Data

Classifier Optimization Continued

Classifier Optimization

Comparison of Training Feature Types for SVM Linear

Figure 2. Example of data structure for Support Vector Machine (SVM) Linear classifiers were trained on integrated data sets with the following parameters:

- 51 MIEs
- 3 Feature Sets
- 6 Classification Algorithms
- 2 Cell Types
- MCF7 profiles
- PC3 profiles

Figure 3. Comparison of internal accuracy for Support Vector Machine (SVM) Linear classifiers trained on different feature types. P-values are from a two-tailed, paired Wilcoxon test.

- To optimize MIE models, we evaluated model performance across all the 51 types of gene expression feature sets (Figure 3) and 6 classification algorithms (Figure 4).

- Classifiers were trained using three different sets of features:
  - Landmark genes
  - Pathway scores
  - Gene set enrichment scores

- Gene set enrichment scores were calculated from “All genes” features using ssGSEA [5], and the resulting scores used as training features

- Cross-fold validation accuracies were compared for the 51 MIE classifiers trained on different feature types using a paired Wilcoxon test

- Landmark Gene based classifiers consistently out-performed “All Gene” and “Pathway Score” based classifiers

Figure 4. Comparison of internal accuracy across different training algorithms. P-values are from a two-tailed, paired Wilcoxon test.

- Explored differences in classifier performance as a function of classification algorithm

- Internal accuracies were compared across algorithms using a paired Wilcoxon test

- svmPoly classifiers achieved significantly higher internal accuracies than svmLinear and svmRadial classifiers

Test for Overfitting

Figure 5. Comparison of internal and hold-out accuracies across classification algorithms. P-values are from a one-tailed, paired Wilcoxon test.

- Comparison of internal and hold-out accuracies from svmPoly-based models indicated a significant overall lift for SVM Poly classifiers

- Restricted further analysis to the runner up svmLinear based classifier

Discussion / Conclusions

In this study we integrated RefChemDB chemical-MIE annotations with LINCS chemical identifiers and gene expression profiles for the purpose of predicting MIE induction from gene expression profiles. We trained binary classifiers to predict 51 distinct MIEs and explored factors that affected model accuracy such as feature type and classification algorithm. Finally, we trained classifiers on both MCF7 and PC3 derived data and compared accuracies, identifying scores that are well modeled in both cell types. A subset of classifiers showed a disparity in performance as a function of cell type and shed light on MIEs that may be better screened in one cell type over another (AR_Positive in MCF7 cells over PC3 cells). Some classifiers showed cell-type specific differences in accuracy.

References