sbv IMPROVER Species Translation Challenge report

Feng Luo

School of Computing

Clemson University

The Sub-Challenge 1:Intra-Species Protein Phosphorylation Prediction

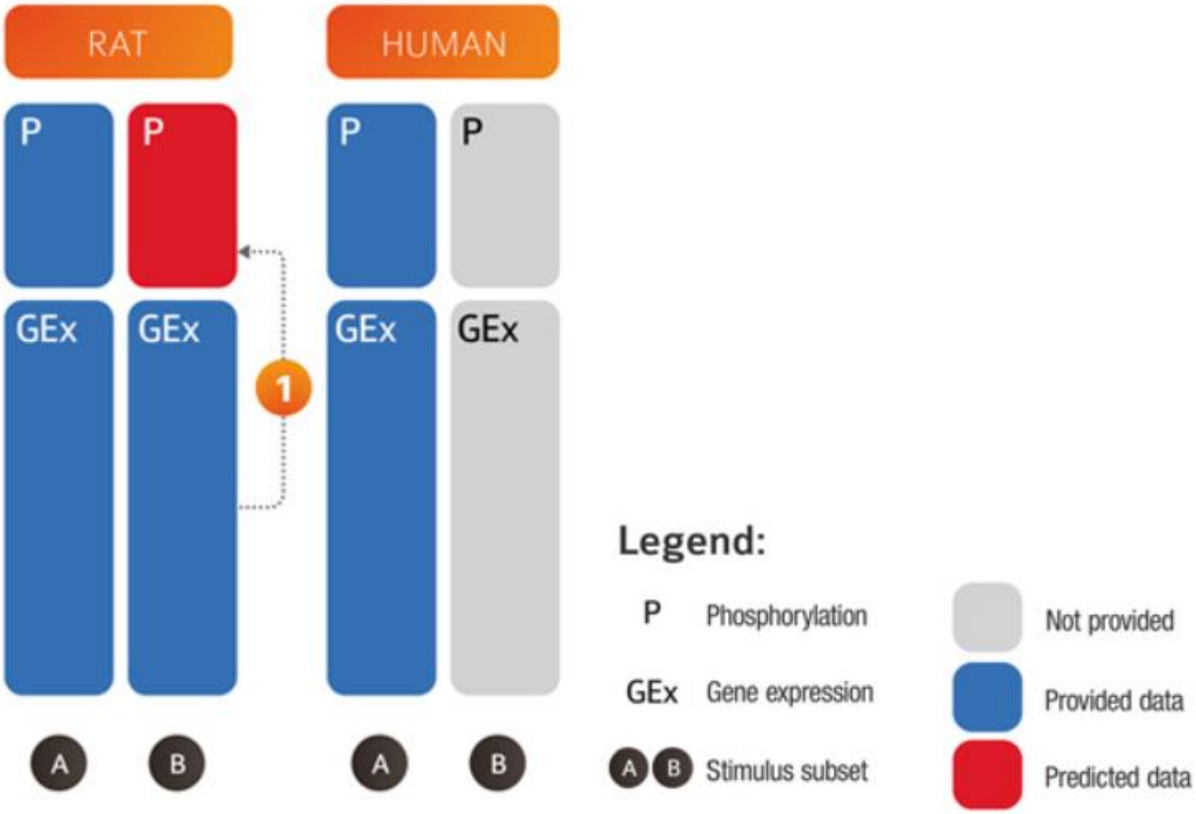
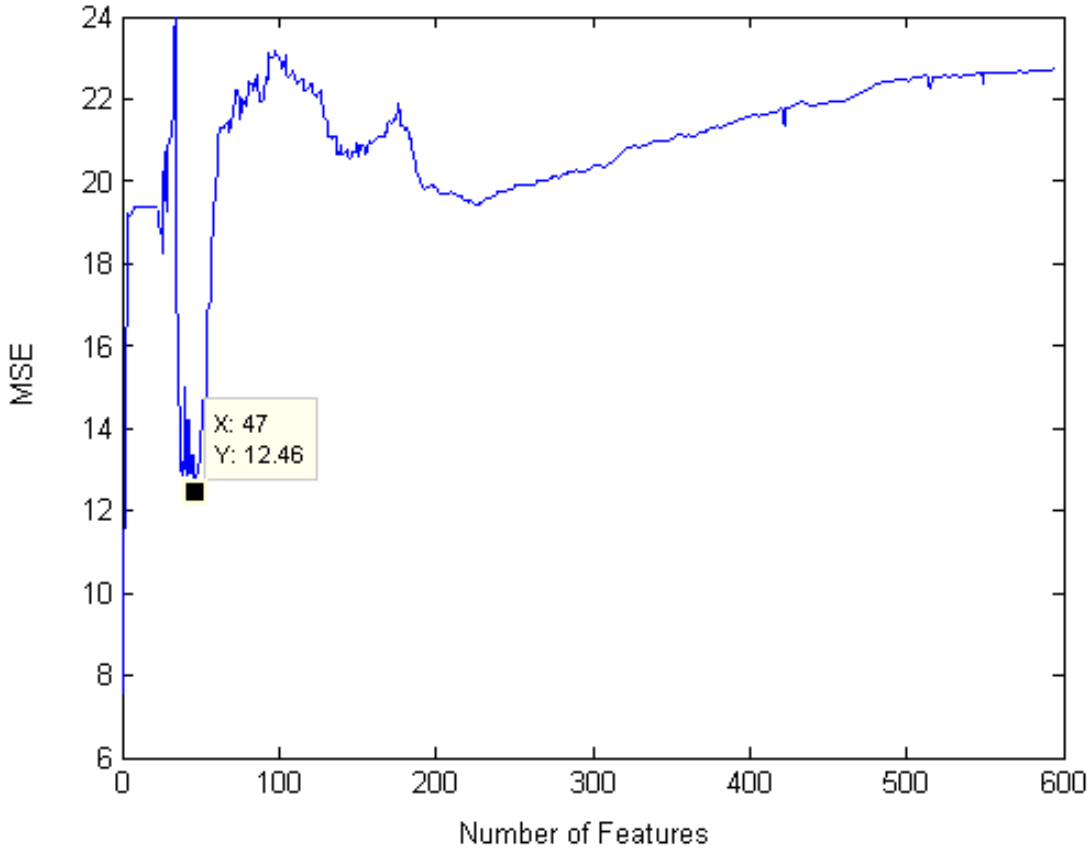


Figure SC1.A: The objective of sub-challenge 1 is the prediction of the activation status of phosphoproteins based on gene expression data in Subset B for 26 stimuli. Data in Subset A, collected with 26 different stimuli, is provided for training.

Construction of Model

- 1. Using ridge regression to obtain the weights for all genes.
- 2. Sorting the genes based on the magnitudes of regression weights in descending order.
- 3. Selecting one gene with largest weight in the remaining genes and adding to the feature space;
- 4. Constructing support vector regression (SVR) model using selected features (genes). Evaluating the performance of SVR model using leave-one-out method on training data.
- 5. Repeating step 3 and 4 until MSE becoming stable.
- 6. Selecting the genes that can give us the lowest MSE on training data as our final features.

MSE of training sets



Prediction Based on Regression

- Used the selected features to construct SVR model from all training data to predict the phosphorylation change of testing data.
- The final phosphorylation statuses were predicted based on the changes.
- RBF kernel is used.

Performance

Team	AUPR	Pearson	BAC	Rank
49	0.42	0.71	0.68	1
50	0.38	0.72	0.68	1
75	0.38	0.71	0.72	1

MSE=6.0291

The frequency of genes

Number of models	*Number of genes*
7	5
6	32
5	25
4	54
3	115
2	186
1	710

Top frequent genes

Genes	Number of models included	Description
LYST	7	lysosomal trafficking regulator
AGPS	7	alkylglycerone phosphate synthase
KEL	7	Kell blood group, metallo-endopeptidase
BEAN	7	brain expressed, associated with Nedd4
TRIM46	7	tripartite motif-containing 46

Change Kernel Function

- **Using Polynomial kernel with degree 3:**
  - MSE=9.5584
  - TN/N：0.9055
  - TP/P：0.6
  - Balanced Accuracy (BAC): 0.7528.

Prediction Based on Classification

- **Constructing a SVM model using selected feature.**
- **The final phosphorylation statuses were predicted using the SVM model.**
- **Perfomance**
  - TN/N: 0.9554
  - TP/P: 0.6286
  - Balanced Accuracy (BAC): 0.7920

Data is imbalance

- **Positive: 35**
- **Negative: 381**

Prediction Based on Weighted Classification

- **Constructing a SVM model using selected feature with different weights for positive and negative samples**
  - 1 for positive
  - 0.03 for negative
- **The final phosphorylation statuses were predicted using the SVM model.**
- **Better performance:**
  - TN/N: 0.9475
  - TP/P: 0.6571
  - Balanced Accuracy (BAC): 0.8023

Next Research

- Deep Learning!

Team Member

- Dr. Zhiming Wang
- Dr. Feng Luo

Acknowledge

- sbv IMPROVER organizer
- PMI
- IBM

Questions?

Thomas Green Clemson (1807-1888)