DEARGEN

To bring biological insights to everyone in the world

DREAM CHALLENGE PROTEOGENOMICS sub task 2 overview

Contents

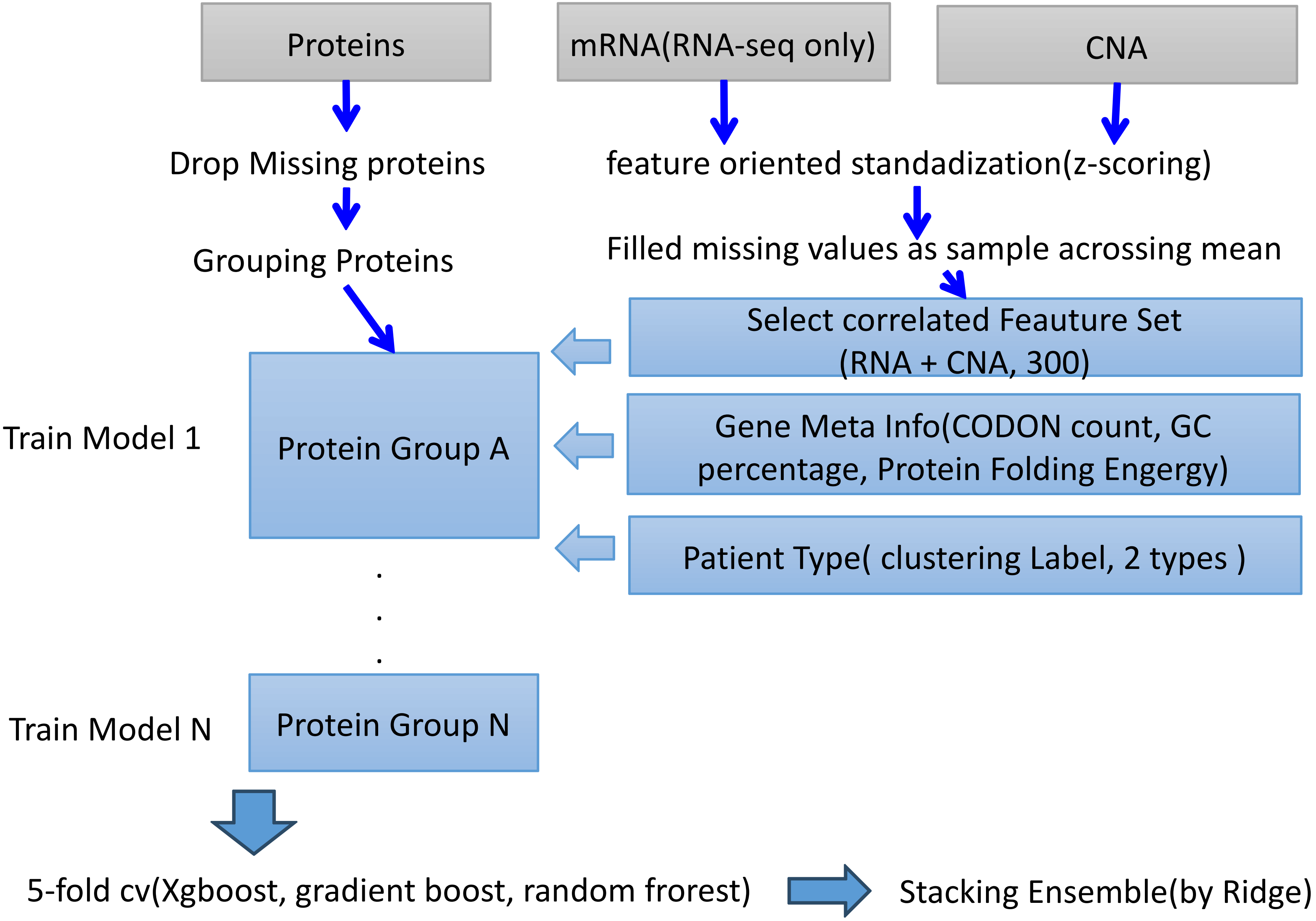
1. Early Approaches

2. Decisions (What decisions we made)

3. Key processes or components improving accuracy

4. Further suggestions

Overall architecture



Proteins

mRNA(RNA-seq only)

CNA

Drop Missing proteins

Grouping Proteins

feature oriented standadization(z-scoring)

Filled missing values as sample acrossing mean

Select correlated Feauture Set (RNA + CNA, 300)

Gene Meta Info(CODON count, GC percentage, Protein Folding Engergy)

Patient Type(clustering Label, 2 types)

Train Model 1

Protein Group A

.
.
.

Train Model N

Protein Group N

5-fold cv(Xgboost, gradient boost, random forest)

Stacking Ensemble(by Ridge)

Early approaches

1. Early Approaches

- Basically we defined this sub-challenge as a traditional regression problem that predicts the abundance of protein.

- We wanted to know whether the computational engineering approach works well or not.

- What we thought as important things are how to select the feature that has the high predictability, and how to reveal the external features to increase that power.

- We couldn't decide the number of proteins to train in early stage.

- We wanted to apply recent deep learning tech such as “Relational Network” to this sub-challenge.

Decisions

2. Decisions (What decisions we made)

- Training the model proteins as a group not as individual.

Single protein model produced high score in local, but got a low score in public. I think this result because validation sample size is too small to trust in single protein model.

- Inserting Gene meta informations

Gene meta informations what we select are CODON counts, GC percentage, protein folding energy

- Inserting patient type label

We Inserted patient type label to each sample, We used the PCA and K-means clustering algorithms.

- Training models

Each Training model has three primitive regressors. (xgboost, random forest, gradient boost). We used stacking ensemble method to submit final prediction. We did 5-fold cross validation

Key components

3. Key points (Key processes or components improving accuracy)

- Normalization / handling missing values

We did feature oriented standadization(z-scoring) on RNA / DNA data. This gave us a huge improvement.

We filled the missing values as mean value of across the samples in RNA / DNA data. We didn’t use missing protein.

- Feature selecting

Basically, we included the coding gene of each protein. Additionally, We selected about 300 other features have high pearson correlation score with current group of proteins. (features might be mRNA or CNA)

- Protein grouping

We used three different way to grouping. These are Pathway based grouping, correlation score based grouping, protein name based grouping. Because of the lack of number of protein in pathway grouping we used it only at sub challenge 3.

Further suggestions

4. Further suggestions.

- Changing the normalization method.

We can sample oriented normalization, or min-max normalization

- Imputing missing protein value.

We can sub1’s method and can increase the training sample size

- Optimizing feature selection

We can set the different feature group size to train. In final submission we fixed the feature size. More over, we can insert relative features manually using domain knowledge.

- Using Deep Learning Method.

Once adapting Relation Network, We had a better score both local cv score and public test score. But we couldn't submit that model since docker limitations

Thank you

‘To bring biological insights to everyone in the world’