Protein-protein interaction based on pairwise similarity

By Nazar Zaki, Sanja Lazarova-Molnar, Wassim El-Hajj, Piers Campbell

Citation: Zaki, NM., Lazarova-Molnar, S., El-Hajj, W., Campbell, P.:Protein-protein interaction based on pairwise similarity. BMC Bioinformatics 2009, 10:150.

Data and programs for experiments no. 1, 3 and 4.
To run in **Cygwin**

Files and programs needed:
The following files are needed and should be included in the same directory.

- **fasta34.exe** downloadable from http://www.ebi.ac.uk/Tools/fasta/index.html
- **gist-train-svm.exe** and **gist-classify.exe** from Gist software (Windows-Cygwin) downloadable from http://bioinformatics.ubc.ca/gist/download.html

Experimental 1

In this experimental work, we tested the performance of our method on randomly selected 15 protein sequences from the yeast protein interaction. The datasets are prepared as follow:

Training Dataset		**Testing Dataset**	
YAR003W-YBR175W	*interact*	YCR077C-YDL160C	*interact*
YBR126C-YML100W	*interact*	YPR072W-YIL038C	*interact*
YNR006W-YOR025W	*non-interact*	YNL137C-YOR025W	*non-interact*
YMR203W-YNL029C	*non-interact*	YMR261C-YOR321W	*non-interact*

To prepare the training files, please type the following in the command line:

```
$ perl train.pl [interaction_file] [sequence_file] [n] [pos]
```

interaction_file = File contains the training protein pairs.
sequence_file = File contains the training protein sequences.
n = the window size, (n = 2, …, 20,000).
pos = number of positive examples in the training set.

To prepare the testing files, please type the following in the command line:

```
$ perl test.pl [interaction_file] [sequence_file] [pos]
```

interaction_file = File contains the testing protein pairs.
sequence_file = File contains the testing protein sequences.
pos = number of positive examples in the training set.

Example:

```
perl train.pl data_ex1/tr_int.txt data_ex1/tr_seq_8.txt 1500 2
perl test.pl data_ex1/ts_int.txt data_ex1/ts_seq_8.txt 2
```

Results:

```
Training results: FP = 0  FN = 0  TP = 2  TN = 2
Training ROC: 1.00000
Test results: FP = 0  FN = 0  TP = 2  TN = 2
Test ROC: 1.00000
*****
RFP is: 0
Sensitivity is: 1
Specificity is: 1
Precision is: 1
F value is: 1
Overall Accuracy is: 1
```

Experimental 3

In this experiment we furthermore split the 100 interacted protein pairs into 2 sets A (50 pairs) and B (50 pairs). We also split the 100 non-interacted protein pairs into 2 sets C (50 pairs) and D (50 pairs). We then combined A with C to create a training dataset and B with D to create a testing dataset.

Example:

```
perl train.pl data_ex3/tr_int.txt data_ex3/tr_seq_100.txt 14000 50
perl test.pl data_ex3/ts_int.txt data_ex3/ts_seq_100.txt 50
```

Results:

```
Training results: FP = 0  FN = 1  TP = 49  TN = 50
Training ROC: 1.00000
Test results: FP = 11  FN = 0  TP = 50  TN = 39
Test ROC: 0.86360
*****
RFP is: 0.22
Sensitivity is: 1
Specificity is: 0.78
Precision is: 0.819672131147541
F value is: 0.79934412789506
Overall Accuracy is: 0.89
```

Experimental 4

In the fourth experimental work, we assess the recognition ability of our method on the dataset created by Xue-Wen et al[1].

Example:

```
perl train.pl data_ex4/tr_int.txt data_ex4/tr_seq_8917.txt 2000 4917
perl test.pl data_ex4/ts_int.txt data_ex4/ts_seq_8917.txt 4917
```

Results:

```
Training results: FP = 732  FN = 690  TP = 4227  TN = 3268
Training ROC: 0.92763
Test results: FP = 1165  FN = 872  TP = 4045  TN = 2835
Test ROC: 0.83392
*****
RFP is: 0.29125
Sensitivity is: 0.822656091112467
Specificity is: 0.70875
Precision is: 0.776391554702495
F value is: 0.741030392224967
Overall Accuracy is: 0.771559941684423
```

[1] Xue-Wen C. and Mei L. (2005) Prediction of protein–protein interactions using random decision forest framework. Bioinformatics, 21, 4394–4400.