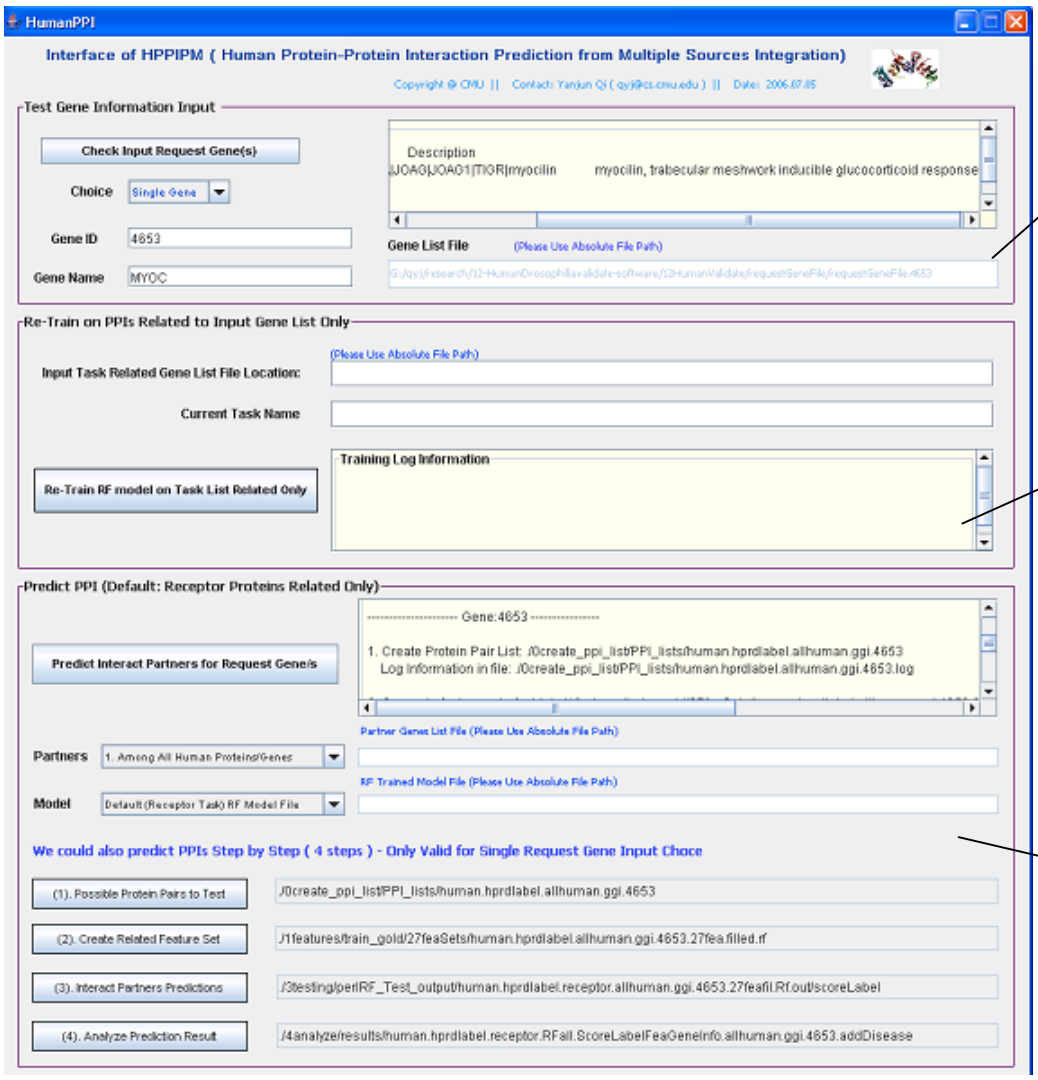
Yanjun Qi (qyj@cs.cmu.edu)

User Manual of HPPIPM

(Software for Human Protein-Protein Interaction Prediction from Multiple Sources Integration)

1. Introduction:

HPPIPM is the name of software aiming for predicting human protein-protein interactions by integrating multiple biological data sources based on the "Random Forest" classifier.
It is a Java based interface and could provide three different kinds of functions as described follows. The snapshot of its full interface looks like:



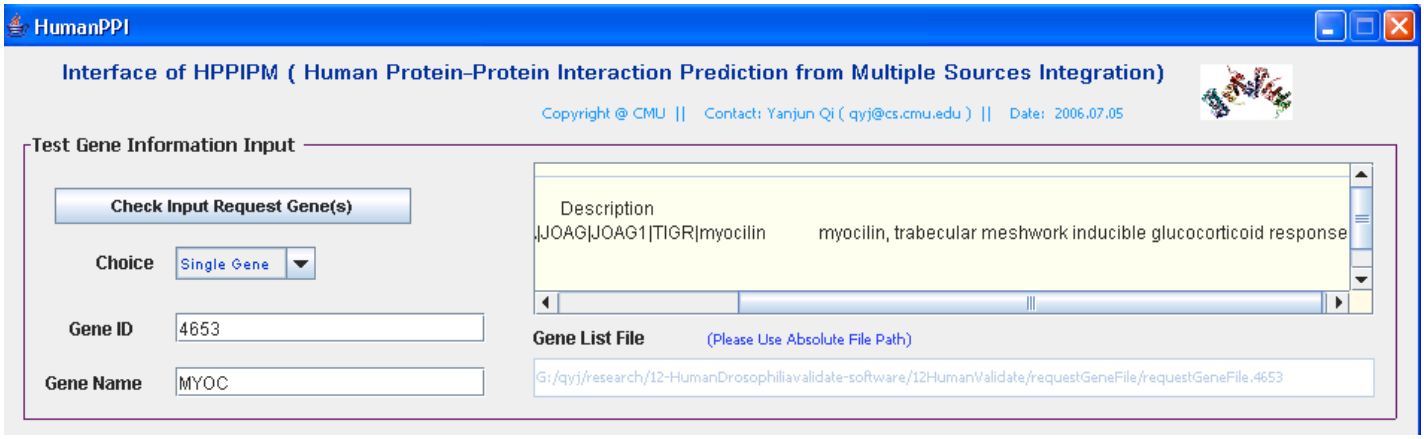
2. Functions

2.1. Check the input request gene / request genes list

There are two possible choices for the input request: (1). Input a single gene (2). Input a request genes list file.

(1). Input a single request gene :

- Input: GeneID and / or GeneSymbol
- Function: Use button "Check Input" to perform the checking
- Output: If this gene corresponding to a human protein in NCBI records or not.
- Output: We would also find and output other names and descriptions of this gene.



(2). Input a single request gene :

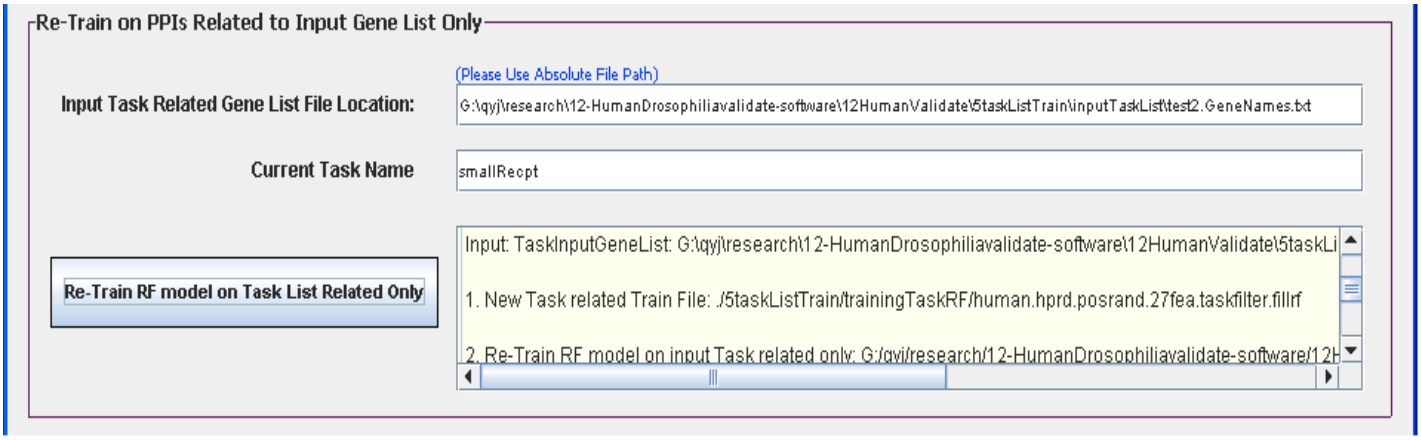
- Input: a file contains a list of request genes
- Function: Use button "Check Input" to perform the checking
- Output: If these genes corresponding to a human protein in NCBI records or not.
- Output: We would also find and output other names and descriptions of the gene in this list.

2.2. Re-Train the existing Random Forest model based on the input task's gene list

- Input: one file contains a specific task related gene names (each gene name in a line). To distinguished the generated predictions from the default "receptor" task, we also ask use to give a name for this task in the "Task Name" input field.

- Function: We could re-train the Random Forest model by the input task gene list. The training data would be only those pairs related to this task of the whole training set. There are three steps related to this re-Training.

- Log information is provided in the "Re-Training Log Window" and the related files generated in each step are also provided to user if needed for detailed investigations.



2.3. Predict Human PPI for the request gene

- Input: Valid request gene or a list of request genes from the first step
- The purpose of this step is to predict protein interacting partners of the input gene/genes.

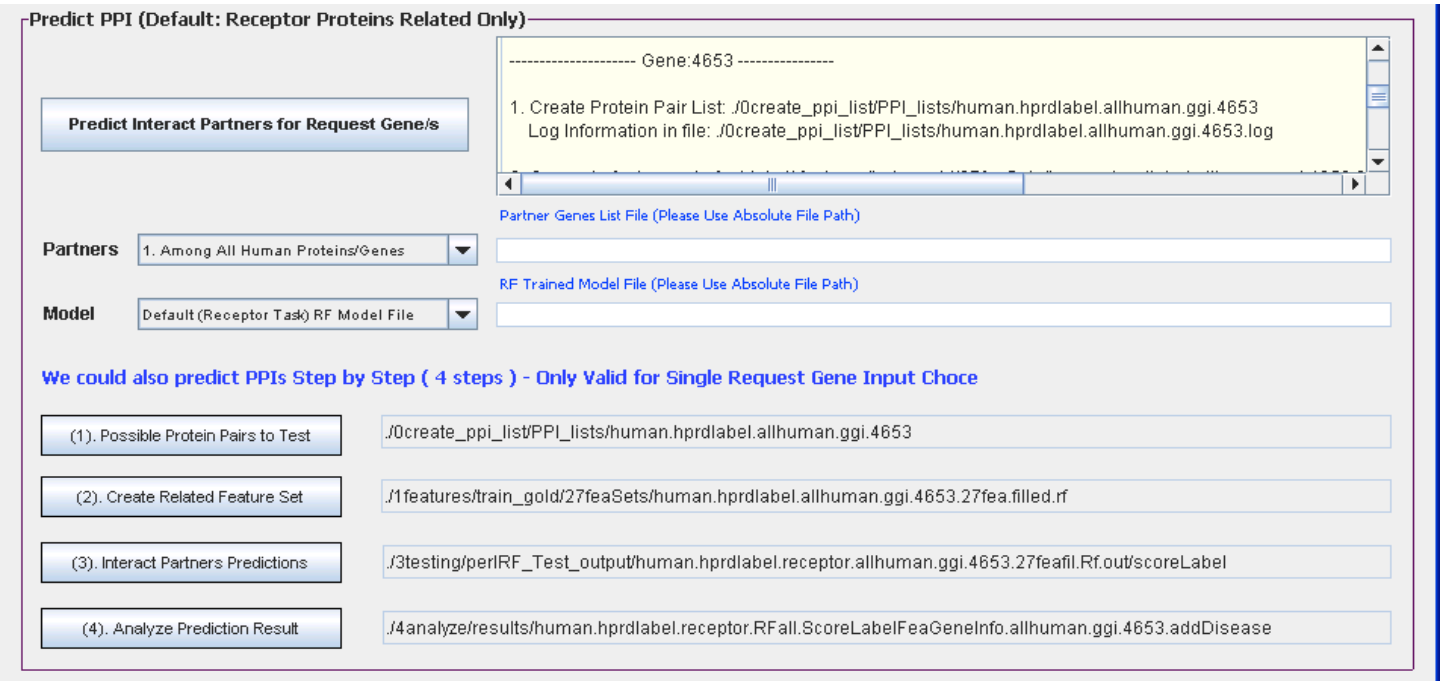
- Choice of Partners: There are three possible choices: (1). Among all possible human proteins/genes; (2) Among a input list of genes; (3) Among the specific task related genes.
- (Note: the available positive interaction information is purely from the HPRD PPI data set.)

- Choice of Trained Model: (1). The default "receptor" task trained RF model ; (2). User provide a RF trained model (for other tasks).

- The prediction for a request gene includes totally four steps. We could pursue this function by running the steps in batch style using button "Predict PPI for Request Gene"; we could also run the process step by step. There are buttons to initialize the action of each step. The step by step running is only available for the "single gene" input request case.

- Four steps: (1). Create the candidate pairs with the request gene; (2). Generate feature set for these candidate pairs; (3). Test RF models on the feature file; (4). Analyze and add gene information / disease information on the predicted score file. The resulting file from this step could be used by user directly then.

- Log information is provided in each step running in the "Running Log Window" and the related files generated in each step are also provided to user if needed for detailed investigations.



3. Installation Guide and System Requirement

- This software includes (1). the java interface and (2). the perl interface for programs used for the training / testing steps.
  - (1) The java interface are under sub-directory: "humanTaskPpi07.jar"
  - (2) The train-test programs (perl interface) are under sub-directory: "HumanValidate-perl06"
- Together it needs roughly 750M space.
- The computer system should contain JDK1.5 or higher.

- Furthermore, this software requires the system to contain the "Perl" and "G77" packages already. For windows system, the user could download "Active Perl" free perl software and "MinGW" free fortran 77 software online.
  - Remember to add the 'g77' and 'perl' execute directory in your system's "PATH" variable.
  - How to add path into the system PATH variable
    - unix (just set PATH variable depending on your SHELL)

  - window (control panel -> system -> advance -> environment vairlabe -> path)

- User needs to set a string parameter to contain the directory of this software's "HumanValidate-perl06" programs ➔ in the parameter file "humanPPIsoftDirPara.txt" (this file is used by the Java wrapper "humanTaskPpi07.jar)
- Put file 'humanPPIsoftDirPara.txt' in the same directory with the JAR file: "humanTaskPpi07.jar"
- Then just click the "humanTaskPpi07.jar" ➔ The java interface should be running.

4. Documents for Developers

- This software includes (1). the java interface and (2). the programs used for the training / testing steps.
- (1) The java interface has fully JavaDoc documents support for further developing
- (2) The predicting functions are mainly the perl codes. Each code has brief comment in the header lines. Developers could easily use and change them.
- (3) The re-Training functions used perl and fortran code. The fortran code is changed from the Berkeley "Random forests - classification" package. It has a complete manual and developing guide online.

5. Perl Interface of the Software (to Run in Batch Mode)

- The subdirectory './HumanValidate-perl06' has the perl interface for the receptor task human PPI
- We could run it using the following wrapper in batch mode from command line:

```
Perl Batch_PrePartAlHum_HprdRecptTrain.pl interested_genelist
```

6. Please Cite

Yanjun Qi, Harpreet K. Dhiman, Neil Bhola, Ivan Budyak, Siddhartha Kar, David Man, Arpana Dutta, Kalyan Tirupula, Brian I. Carr, Jennifer Grandis, Ziv Bar-Joseph and Judith Klein-Seetharaman (2009) "Systematic prediction of human membrane receptor interactions" PROTEOMICS (In Press)

- (Supplementary Web): http://www.cs.cmu.edu/~qyj/HMRI/