ECMPride

A FLEXIBLE AND SCALABLE TOOL DEVELOPED FOR PREDICTING EXTRACELLULAR MATRIX PROTEINS

USER MANUAL FOR

ECMPride

Author: Mr. Binghui Liu

Address: State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing 102206, China

Content

Chapter 1. Software Overview

ECMPride is a flexible and scalable tool developed for predicting extracellular matrix (ECM) proteins. ECMPride can directly perform ECM prediction by taking UniProt IDs in CSV (*.csv) file format as input. The core of ECMPride was written in R 3.6.1 language on the RStudio 1.1.442 under Windows System. The function in ECMPride are based on R statistical environment.

Chapter 2. Installation

This chapter explains how to download and install ECMPride on the user’s computer.

2.1. Requirement

2.1.1 Hardware requirements

a) 2.0 GHz CPU minimum

b) 2 GB RAM minimum

2.1.2 Software requirements

a) Supported operating system (OS) versions (32-bit or 64-bit)

Windows 7

Windows 10

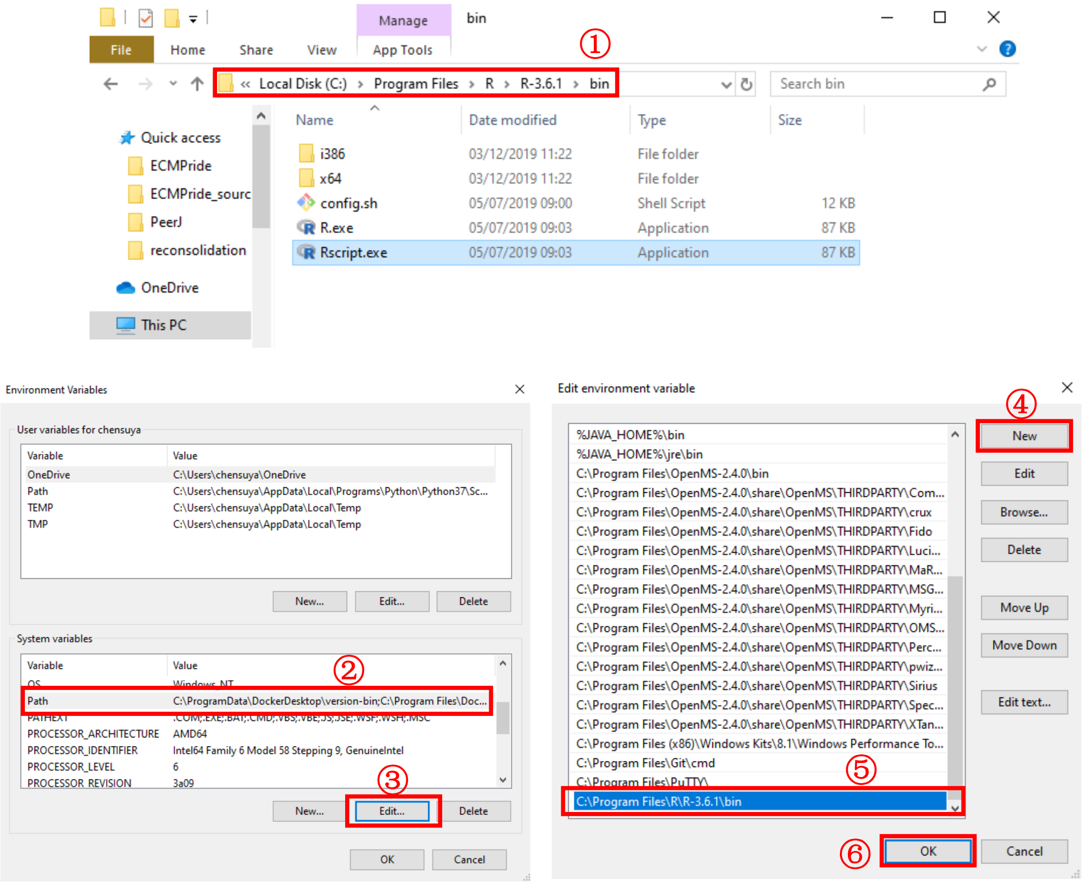
b) R 3.6.1 or higher (for Windows) from R project

2.2. Configuration of R Environment

2.2.1 Setting system environment variable

After installing R, users must add the path of RScript.exe into the system environment variable before using ECMPride, because ECMPride is currently running from the command line by calling Rscript.exe. When there are several versions of R installed in a user’s computer, ECMPride will call the Rscript.exe whose path is added into the system environment variable. The method for setting system environment variable can be found at https://www.computerhope.com/issues/ch000549.htm.

By default, Rscript.exe is in the path of “C:\Program Files\R\R-3.6.1\bin”. Then, this path should be added into the system environment variable. See Fig. 1. for details.



Name	Date modified	Type	Size
i386	03/12/2019 11:22	File folder	
x64	03/12/2019 11:22	File folder	
config.sh	05/07/2019 09:00	Shell Script	12 KB
R.exe	05/07/2019 09:03	Application	87 KB
Rscript.exe	05/07/2019 09:03	Application	87 KB

User variables for chensuya

Variable	Value
OneDrive	C:\Users\chensuya\OneDrive
Path	C:\Users\chensuya\AppData\Local\Programs\Python\Python37\Sc...
TEMP	C:\Users\chensuya\AppData\Local\Temp
TMP	C:\Users\chensuya\AppData\Local\Temp

System variables

Variable	Value
OS	Windows_NT
Path	C:\ProgramData\DockerDesktop\version-bin;C:\Program Files\Doc...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64
PROCESSOR_IDENTIFIER	Intel64 Family 6 Model 58 Stepping 9, GenuineIntel
PROCESSOR_LEVEL	6
PROCESSOR_REVISION	3a09

```
%JAVA_HOME%\bin
%JAVA_HOME%\jre\bin
C:\Program Files\OpenMS-2.4.0\bin
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\Com...
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\crux
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\Fido
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\Luci...
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\MaR...
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\MSG...
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\Myri...
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\OMS...
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\Perc...
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\pwiz...
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\Sirius
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\Spec...
C:\Program Files\OpenMS-2.4.0\share\OpenMS\THIRDPARTY\XTan...
C:\Program Files (x86)\Windows Kits\8.1\Windows Performance To...
C:\Program Files\Git\cmd
C:\Program Files\PuTTY\
C:\Program Files\R\R-3.6.1\bin
```

Fig. 1. The illustration of adding the RScript.exe path into system environment variable.

2.2.2 Installing R packages

The required R packages and their installation commands are listed below:

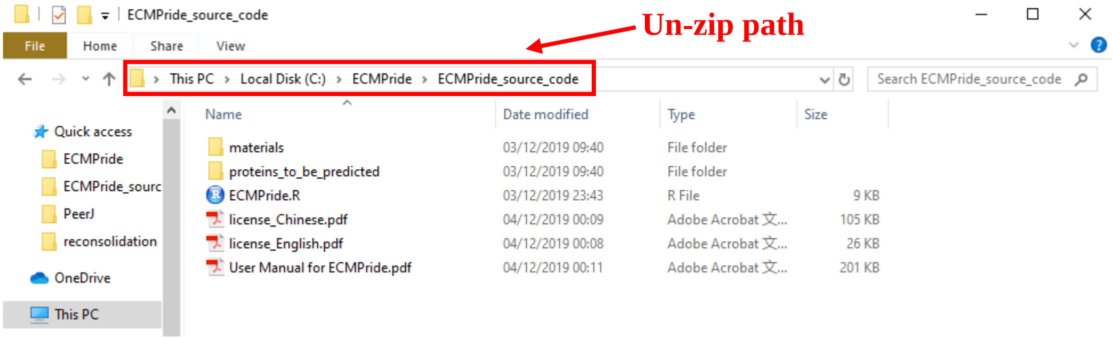
```
install.packages("randomForest")
install.packages("plyr")
install.packages("dplyr")
install.packages("xlsx")
install.packages("mRMRe")
install.packages("caret")
install.packages("parallel")
```

You should install these R packages by R 3.6.1 (not R 3.5.3 or the older version) ahead of time. In fact, ECMPride will install these R packages itself the first time it runs, but this approach may face some unknown errors. Therefore, we recommend users to install these R packages before running ECMPride.

2.3. Download and UnZip ECMPride

ECMPride can be freely downloaded from https://github.com/Binghui-Liu/ECMPride.git. Un-compress the zip package (or 7z) into a specified local folder.

In the file folder after decompression, you can find an R file "ECMPride.R", which is the main file for program running and you should record the local path of this file (The path in my computer is: C:\ECMPride\ECMPride_source_code\ ECMPride.R) as path_1 (Fig. 2.)



Un-zip path

This PC › Local Disk (C:) › ECMPride › ECMPride_source_code

Name	Date modified	Type	Size
materials	03/12/2019 09:40	File folder	
proteins_to_be_predicted	03/12/2019 09:40	File folder	
ECMPride.R	03/12/2019 23:43	R File	9 KB
license_Chinese.pdf	04/12/2019 00:09	Adobe Acrobat 文...	105 KB
license_English.pdf	04/12/2019 00:08	Adobe Acrobat 文...	26 KB
User Manual for ECMPride.pdf	04/12/2019 00:11	Adobe Acrobat 文...	201 KB

Fig. 2. The illustration of configuring the runtime environment

Chapter 3. Begin To Predict

3.1. Format requirements for input data

a) The input data for ECMPride is the UniProt IDs of proteins to be predicted. Therefore, you need to unify the characterization of proteins to be predicted as UniProt IDs (The UniProt website provides multiple types of ID conversion services: https://www.uniprot.org/uploadlists/);

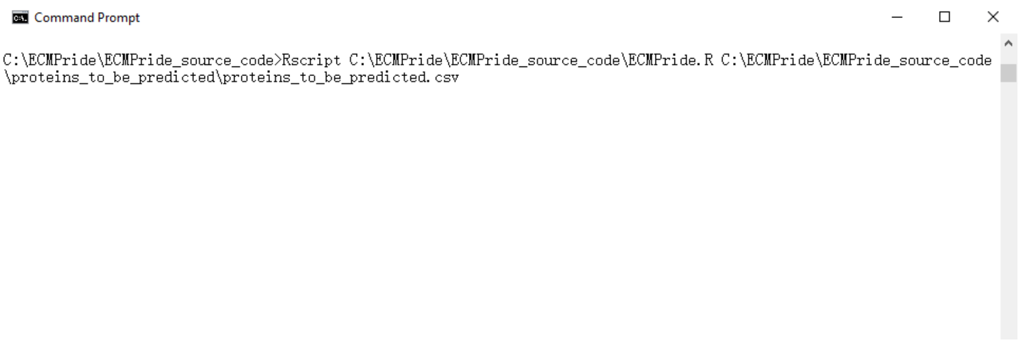
b) You need to write UniProt IDs of proteins to be predicted to a CSV (*.csv) file. We present a sample CSV (*.csv) file under the proteinsToBePredicted folder in the ECMPride archive (proteinsToBePredicted.csv): The file consists of N rows and 1 column, the first row is titled "UniProtID", and the second through the N-th row is UniProt IDs of proteins to be predicted, one for each row.

c) After you enter UniProt IDs of proteins to be predicted into the CSV (*.csv) file according to the above requirements, record the local path of this file (The path in my computer is: C:\ECMPride\ECMPride_source_code\proteins_to_be_predicted\ proteins_to_be_predicted.csv) as path_2.

3.2. Run ECMPride from the command line

a) Open the command line (An easy way to do this: press Win + R at the same time, then type "cmd" and press enter to open the command line)

b) As shown in Fig. 3., enter <Rscript path_1 path_2> on the command line (< > contains the input: Rscript+ space + path_1+ space + path_2). Path_1 is the local path of ECMPride.R, which is the main file for program running, and path_2 is the local path of input file. Press enter and ECMPride begins to predict proteins to be predicted.

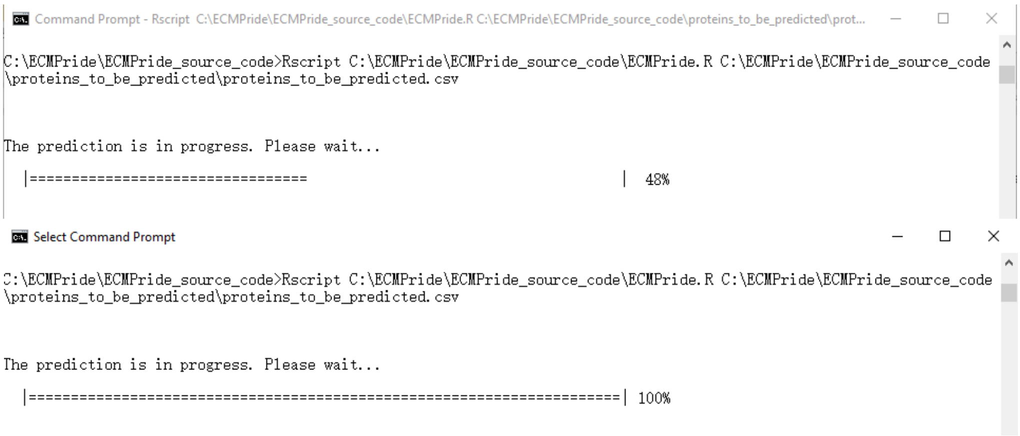


```
C:\ECMPride\ECMPride_source_code>Rscript C:\ECMPride\ECMPride_source_code\ECMPride.R C:\ECMPride\ECMPride_source_code
\proteins_to_be_predicted\proteins_to_be_predicted.csv
```

Fig. 3. The illustration of the commands entered on the command line

3.3. Command line output and result files

a) Showing the progress of the program. A progress bar appears in the center of the screen to indicate the progress of the program. When the progress bar is loaded to 100%, the program is predicted to complete (Fig. 4.).

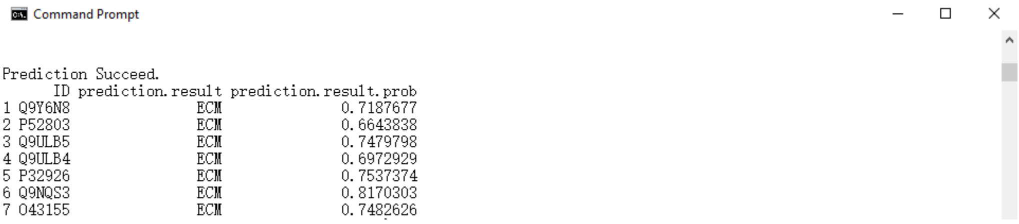

Command Prompt - Rscript C:\ECMPride\ECMPride_source_code\ECMPride.R C:\ECMPride\ECMPride_source_code\proteins_to_be_predicted\prot...

```
C:\ECMPride\ECMPride_source_code>Rscript C:\ECMPride\ECMPride_source_code\ECMPride.R C:\ECMPride\ECMPride_source_code
\proteins_to_be_predicted\proteins_to_be_predicted.csv


The prediction is in progress. Please wait...
  |================================                                   |  48%
```

```
C:\ECMPride\ECMPride_source_code>Rscript C:\ECMPride\ECMPride_source_code\ECMPride.R C:\ECMPride\ECMPride_source_code
\proteins_to_be_predicted\proteins_to_be_predicted.csv


The prediction is in progress. Please wait...
  |===================================================================| 100%
```

Fig. 4. The progress bar that shows the progress of a program

b) Description of prediction results. If all proteins to be predicted are successfully predicted, the command line will prompt "Prediction Succeed.", and simple prediction results of all proteins to be predicted will be displayed on the command line, as shown in Fig. 5.

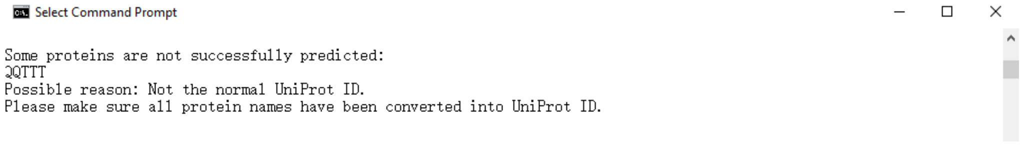


```
Prediction Succeed.
      ID prediction.result prediction.result.prob
1 Q9Y6N8               ECM              0.7187677
2 P52803               ECM              0.6643838
3 Q9ULB5               ECM              0.7479798
4 Q9ULB4               ECM              0.6972929
5 P32926               ECM              0.7537374
6 Q9NQS3               ECM              0.8170303
7 O43155               ECM              0.7482626
```

Fig. 5. Display successful predicted protein IDs and their predicted results

The prediction result displayed on the command line contains three columns of data: the first column contains UniProt IDs of proteins to be predicted, the second column contains the prediction result (ECM or non-ECM), and the third column contains the probability of the second column. The sum of the probabilities of each protein being predicted to be ECM or non-ECM is 1. If the predicted result of a protein is ECM and the probability value is 0.85, it means that ECMPride thinks that the protein has an 85% probability of being ECM and a 15% probability of being non-ECM, so the protein is more likely to be ECM in general. It is worth noting that if there are too many proteins to predict, the command line will only show partial results, and the full results will be saved in the results file.

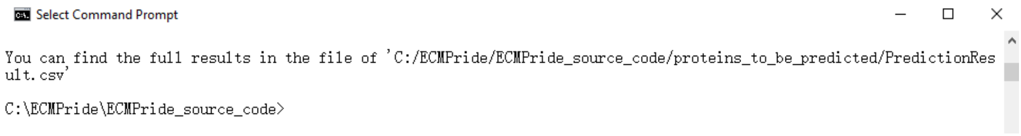
If Some of proteins to be predicted are not successfully predicted, the command line prompts "Some proteins are not successfully predicted:", and then displays the IDs of all the proteins that are not successfully predicted on the command line with the possible reasons, as shown in Fig. 6.



```
Some proteins are not successfully predicted:
JQTTT
Possible reason: Not the normal UniProt ID.
Please make sure all protein names have been converted into UniProt ID.
```

Fig. 6. Display unsuccessfully predicted protein IDs and suggest possible causes

c) Prompt for the location of the result file. All successfully predicted proteins and their complete predicted results (including annotations of proteins to be predicted) will be saved in a CSV (*.csv) result file in the same path as the input file, and the program will eventually prompt for the address of the file, as shown in Fig. 7.



```
You can find the full results in the file of 'C:/ECMPride/ECMPride_source_code/proteins_to_be_predicted/PredictionRes
ult.csv'

C:\ECMPride\ECMPride_source_code>
```

Fig .7. Prompt where to save the results file

Chapter 4. Support Services

4.1. Contact

For any questions involving ECMPride, please contact Mr. Binghui Liu (Email: l_binghui@163.com).

4.2. Copyright

This software product is developed by Mr. Binghui Liu from the National Center of Protein Sciences (Beijing)-Bioinformatics group. All titles and intellectual property rights, which is generated by the software product including, but not limited to, relative images, data, texts, additional program and other software products (dll, exe, etc.), incidental help materials, and any copies of the Software Products are protected by Copyright Law of People’s Republic of China and international copyright treaties and other intellectual property laws and treaties. Users only get the right to use this software product for non-commercial uses.