

# ECMPrise

A FLEXIBLE AND  
SCALABLE TOOL  
DEVELOPED FOR  
PREDICTING  
EXTRACELLULAR  
MATRIX PROTEINS

## USER MANUAL FOR ECMPrise

Author: Mr. Binghui Liu

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

# Content

<b>Chapter 1. Software Overview .....</b>	<b>2</b>
<b>Chapter 2. Installation.....</b>	<b>3</b>
2.1. Requirement.....	3
2.1.1 Hardware requirements .....	3
2.1.2 Software requirements.....	3
2.2. Configuration of R Environment.....	3
2.2.1 Setting system environment variable .....	3
2.2.2 Installing R packages.....	4
2.3. Download and UnZip ECMPrize .....	5
<b>Chapter 3. Begin To Predict .....</b>	<b>7</b>
3.1. Format requirements for input data .....	7
3.2. Run ECMPrize from the command line.....	7
3.3. Command line output and result files.....	8
<b>Chapter 4. Support Services .....</b>	<b>12</b>
4.1. Contact.....	12
4.2. Copyright .....	12

# Chapter 1. Software Overview

ECMPrise is a flexible and scalable tool developed for predicting extracellular matrix (ECM) proteins. ECMPrise can directly perform ECM prediction by taking UniProt IDs in CSV (\*.csv) file format as input. The core of ECMPrise was written in [R 3.6.1 language](#) on the [RStudio 1.1.442](#) under Windows System. The function in ECMPrise are based on [R statistical environment](#). Both single-threaded and multi-threaded versions of ECMPrise are provided here.

## Chapter 2. Installation

This chapter explains how to download and install ECMPrise 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 ECMPrise, because ECMPrise is currently running from the command line by calling Rscript.exe. When there are several versions of R installed in a user's computer, ECMPrise 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.

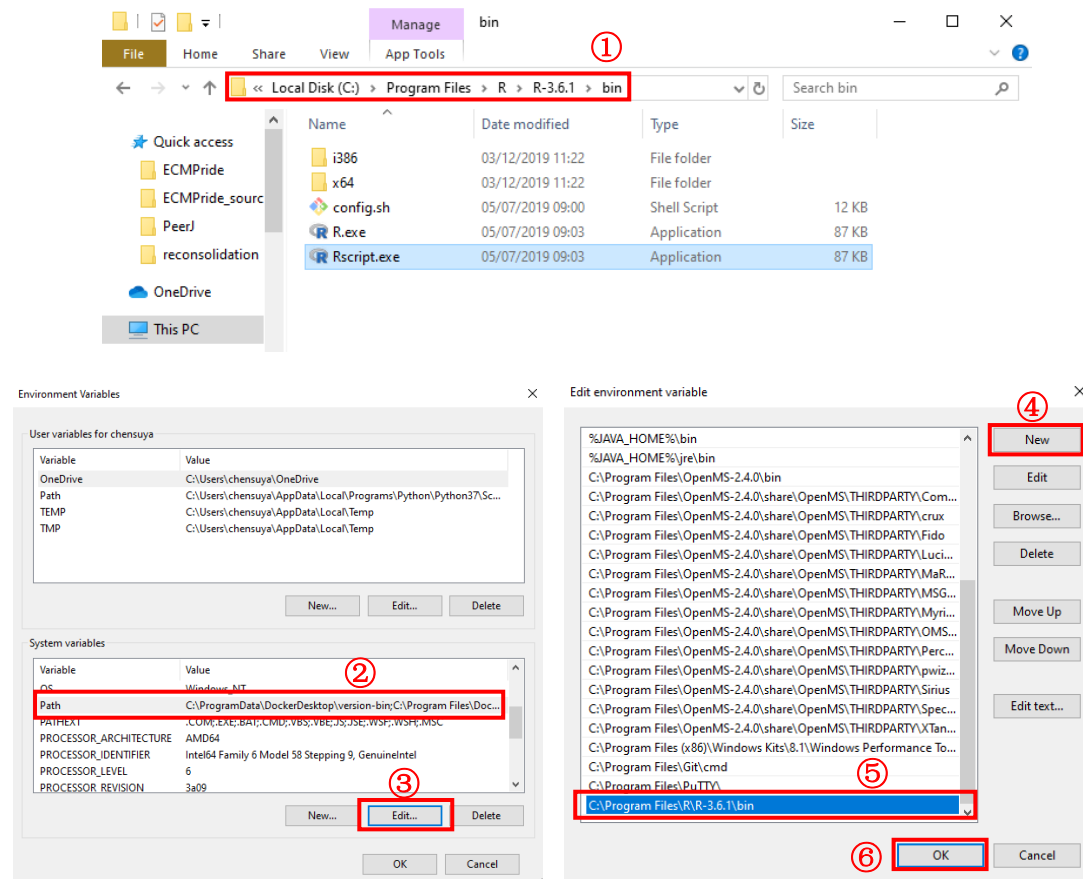


Fig. 1. The illustration of adding the RScript.exe path into system environment variable.

## 2.2.2 Installing R packages

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

1. `install.packages("randomForest")`
2. `install.packages("plyr")`

```
3. install.packages("dplyr")
4. install.packages("xlsx")
5. install.packages("mRMRe")
6. install.packages("caret")
7. 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, ECMPrise 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 ECMPrise.

## 2.3. Download and UnZip ECMPrise

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

In the file folder after decompression, you can find the R file of "ECMPrise.R" and "ECMPrise\_parallel.R", which are single-threaded and multi-threaded versions of ECMPrise. When the number of proteins to be predicted is small (less than 2000), it is recommended that you choose single-threaded ECMPrise; when the number of proteins to be predicted is large, it is recommended that you choose multi-threaded ECMPrise. Next we'll illustrate ECMPrise with the example of single-threaded ECMPrise.

## USER MANUAL FOR ECMPrize

First of all, you should record the local path of “ECMPrize.R” (The path in my computer is: C:\ECMPrize\ECMPrize\_source\_code\ ECMPrize.R) as path\_1 (Fig. 2.)

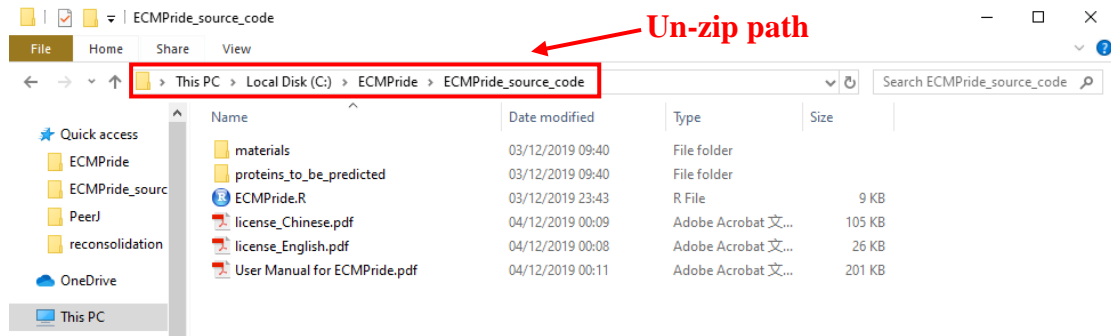


Fig. 2. The illustration of configuring the runtime environment.

## Chapter 3. Begin To Predict

### 3.1. Format requirements for input data

- a) The input data for ECMPrize is the UniProt IDs of proteins to be predicted. Therefore, you need to convert the IDs of proteins that need be analyzed by ECMPrize to UniProt IDs (The UniProt website provides multiple types of ID conversion services:

<https://www.uniprot.org/uploadlists/>);

- b) Write UniProt IDs of proteins into a CSV (\*.csv) file. We present a sample CSV (\*.csv) file under the proteinsToBePredicted folder in the ECMPrize 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) Record the local path of the CSV (\*.csv) file with UniProt IDs (The path in my computer is:  
C:\ECMPrize\ECMPrize\_source\_code\proteins\_to\_be\_predicted\  
proteins\_to\_be\_predicted.csv) as path\_2.

### 3.2. Run ECMPrize 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 ECMPrize.R, which is the main file for program running, and path\_2 is the local path of input file. Press enter and ECMPrize begins to predict proteins to be predicted.

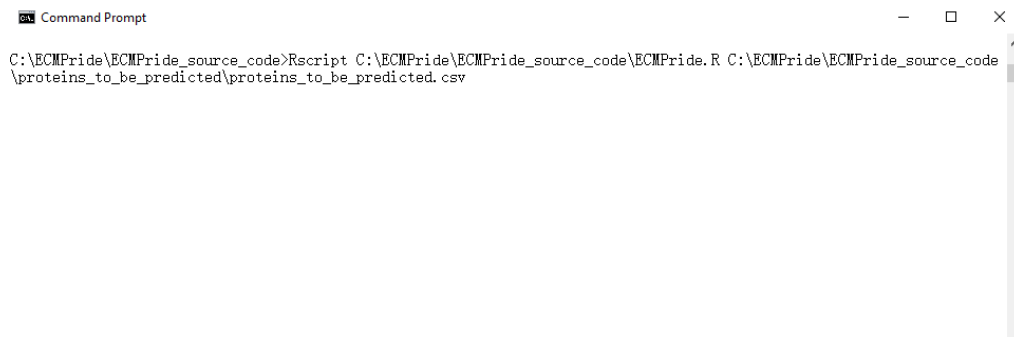


Fig. 3. The illustration of the commands entered on the command line

### 3.3. Command line output and result files

- a) Running progress of the ECMPrize. 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.).

## USER MANUAL FOR ECMPrise

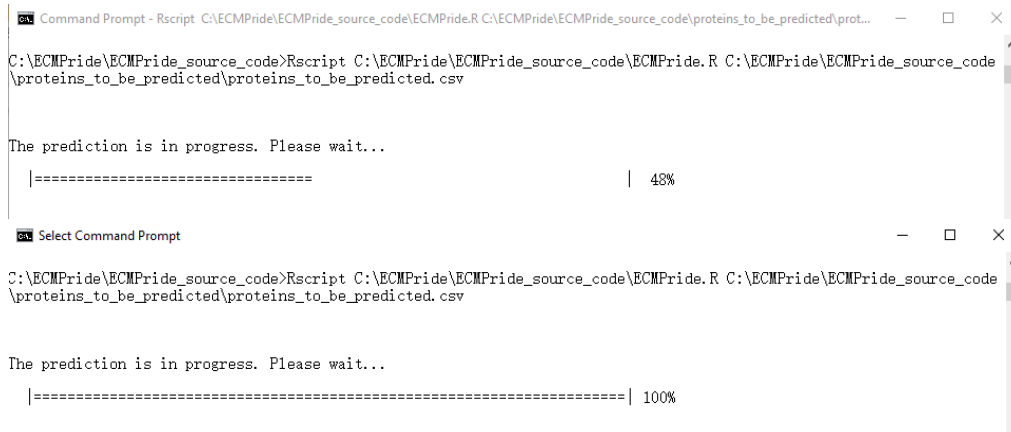


Fig. 4. The progress bar that shows the progress of a program

- b) Prediction results. When all proteins to be analyzed 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.

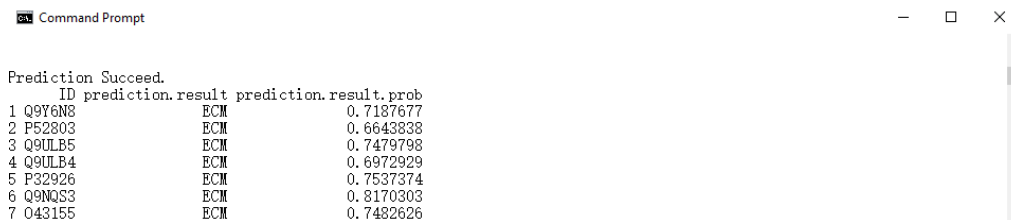


Fig. 5. Display successful predicted protein IDs and their predicted results

The prediction result displayed on the command line contains three columns of data: UniProt IDs of proteins to be predicted, the prediction result (ECM or non-ECM), and 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 the

probability calculated by ECMPrize of this protein being a ECM or a non-ECM is 85% and 15%, respectively; 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.

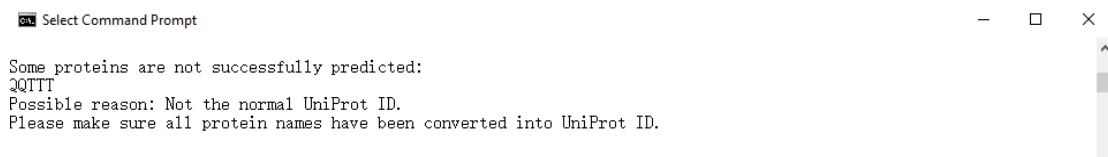


Fig. 6. Display unsuccessfully predicted protein IDs and suggest possible causes

- 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.

## USER MANUAL FOR ECMPrize



Fig .7. Prompt where to save the results file

## **Chapter 4. Support Services**

### **4.1. Contact**

For any questions involving ECMPrize, please contact Mr. Binghui Liu (Email: [l\\_binghui@163.com](mailto: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.