User Guide of AP3

Version 1.0

Last revised February10, 2022

Zhiqiang Gao[1, 3||], Cheng Chang[2||], Jinghan Yang[1, 3],

Yunping Zhu[2, 4*], Yan Fu[1, 3*]

1, National Center for Mathematics and Interdisciplinary Sciences, Key Laboratory of Random Complex Structures and Data Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

2, State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China.

3, School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.

4. Anhui Medical University, Hefei 230032, China.

|| Contributed equally to this work

* To whom correspondence should be addressed:

Yan Fu, Email: yfu@amss.ac.cn

Yunping Zhu, Email: zhuyunping@gmail.com

Contents:

1 Introduction

AP3 (short for Advanced Proteotypic Peptide Predictor) is an improved proteotypic peptide prediction tool by taking the protein proteolytic digestion process into consideration. Firstly, a random forest classifier is trained for predicting the cleavage probability of tryptic sites. Then the peptide digestibility is calculated and integrated into the peptide detectability predictor as a feature when charactering peptides. Finally, AP3 trains the peptide detectability prediction model using the selected features. Experiments on three available public datasets demonstrate that AP3 has superior performance to existing methods for proteotypic peptides prediction. AP3 is freely available at http://fugroup.amss.ac.cn/software/AP3/AP3.html.

2 Installation

2.1 Requirements

2.1.1 Hardware Requirements

- 1 CPU processor at 2.4 GHz or higher
- 2G RAM or higher
- 100G of free hard disk space or higher

2.1.2 Software Requirements

- Verified operating system (OS) versions (64-bit)
  - Windows 7 SP1
  - Windows 10
- Download and install the R2014a MATLAB Runtime.
- .NET Framework 4.5 or above from Microsoft
- Download and install the Visual C++ Redistributable for Visual Studio 2013

2.2 Configuration of MATLAB Runtime

Two MATLAB R2014a scripts are called when implementing the AP3 algorithm, so the R2014a MATLAB Runtime should be configured before running AP3. Firstly, users should download the R2014a MATLAB Runtime 64-bit version (Fig. 1) from our website or the official website: https://ww2.mathworks.cn/products/compiler/matlab-

runtime.html. Then double click the "MCR_R2014a_win64_installer.exe" to install it. Finally, users should add the path of *MATLABRuntimeDirectory*\bin\win64 into the system environment variable before using AP3. The method for setting system environment variable can be found at http://www.computerhope.com/issues/ch000549.htm.

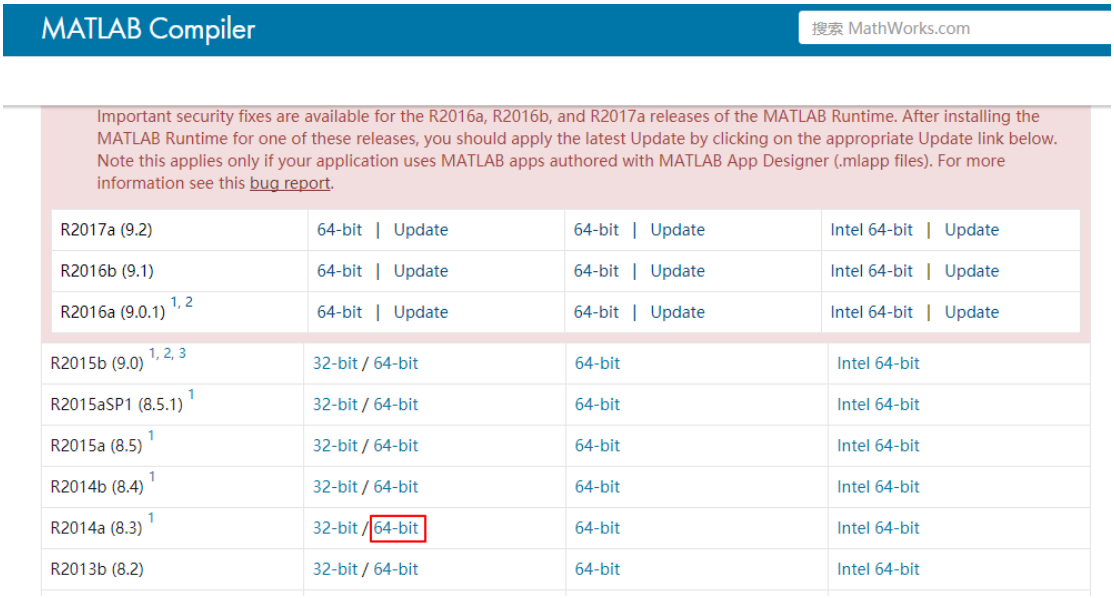

Figure 1. Illustration of downloading the R2014a MATLAB Runtime.

For example, if we install the MATLAB runtime into the directory "D:\Program Files", then we should make sure the following paths "D:\Program Files\MATLAB8.3 Compiler Runtime\v83\bin\win64" and "D:\Program Files\MATLAB8.3 Compiler Runtime\v83\runtime\win64" are added to the system environment variable "Path" (Fig. 2).

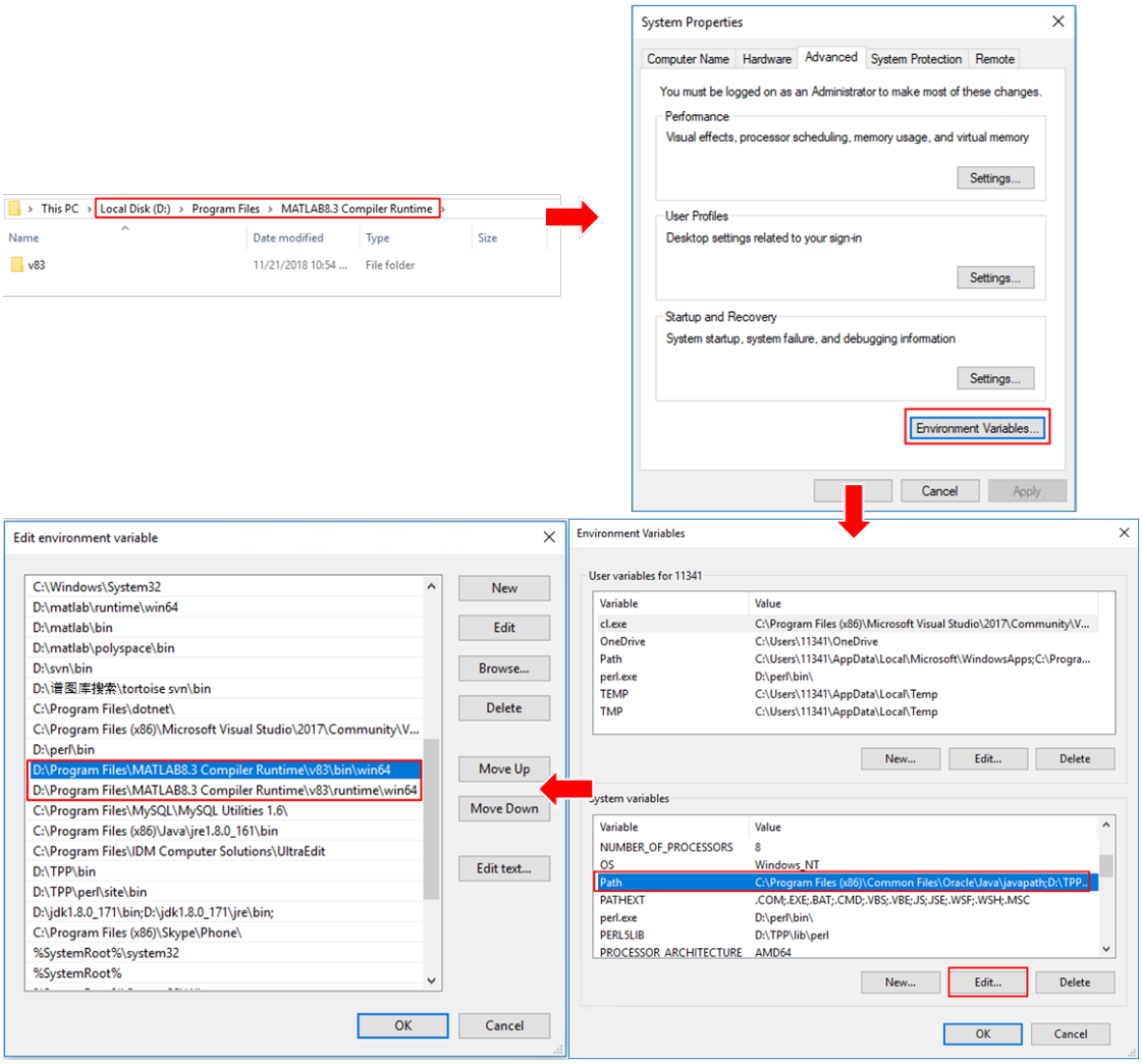

Figure 2. Illustration of adding paths into system environment variable "Path".

2.3 Downloading AP3

AP3 can be freely downloaded from http://fugroup.amss.ac.cn/software/AP3/AP3.html. Users can download the release version of AP3 "AP3.rar" (Fig. 3a), the source code of AP3 "SourceCodeOfAP3.rar" (Fig. 3b), the helping document "User Guide of AP3.pdf" (Fig. 3c) and the test data set "TestData.rar" (Fig. 3d) from this website.

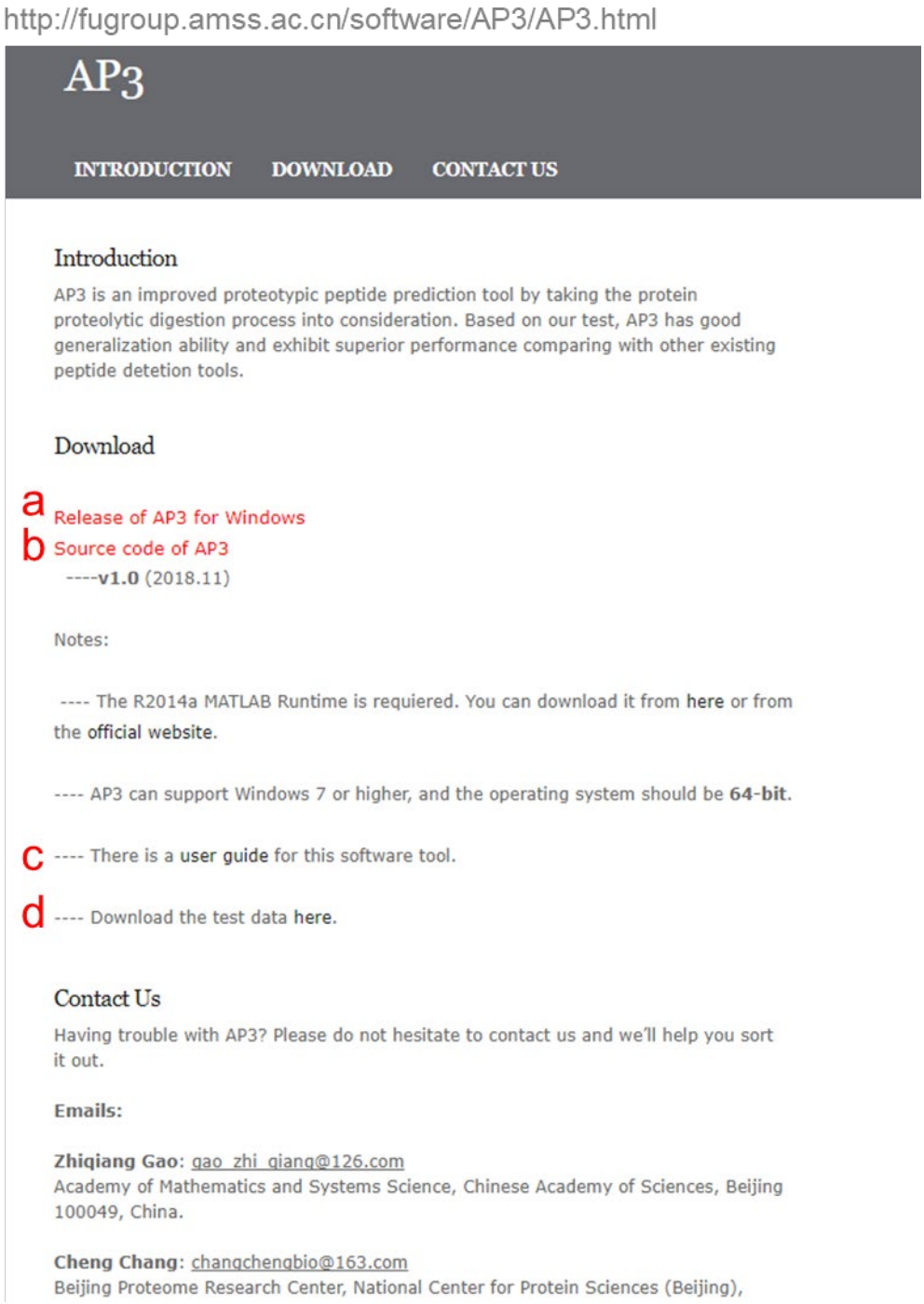

Figure 3. The screenshot of the website of AP3

Un-compress the zip package "AP3.rar" into a specified file folder. Double-click "GUI_AP3.exe" and the GUI of AP3 will appear as shown in Fig. 4.

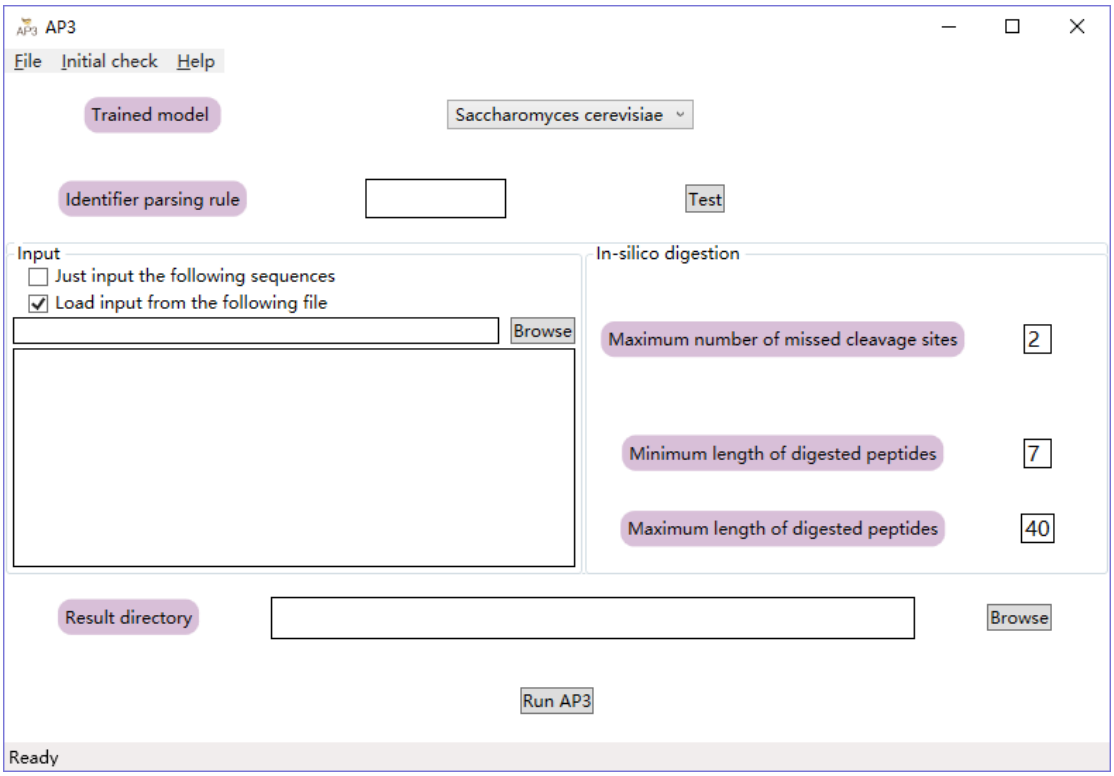

Figure 4. The screenshot of the GUI of AP3.

3 Running AP3

3.1 Using the GUI

1.1.1 Initial check for the program environment

In the menu “Initial check”, click “Matlab runtime check” (Fig. 5a) to test if the R2014a MATLAB Runtime is installed successfully. AP3 will show a message to the user that the MATLAB runtime has been installed (Fig. 5b) or not (Fig. 5c).

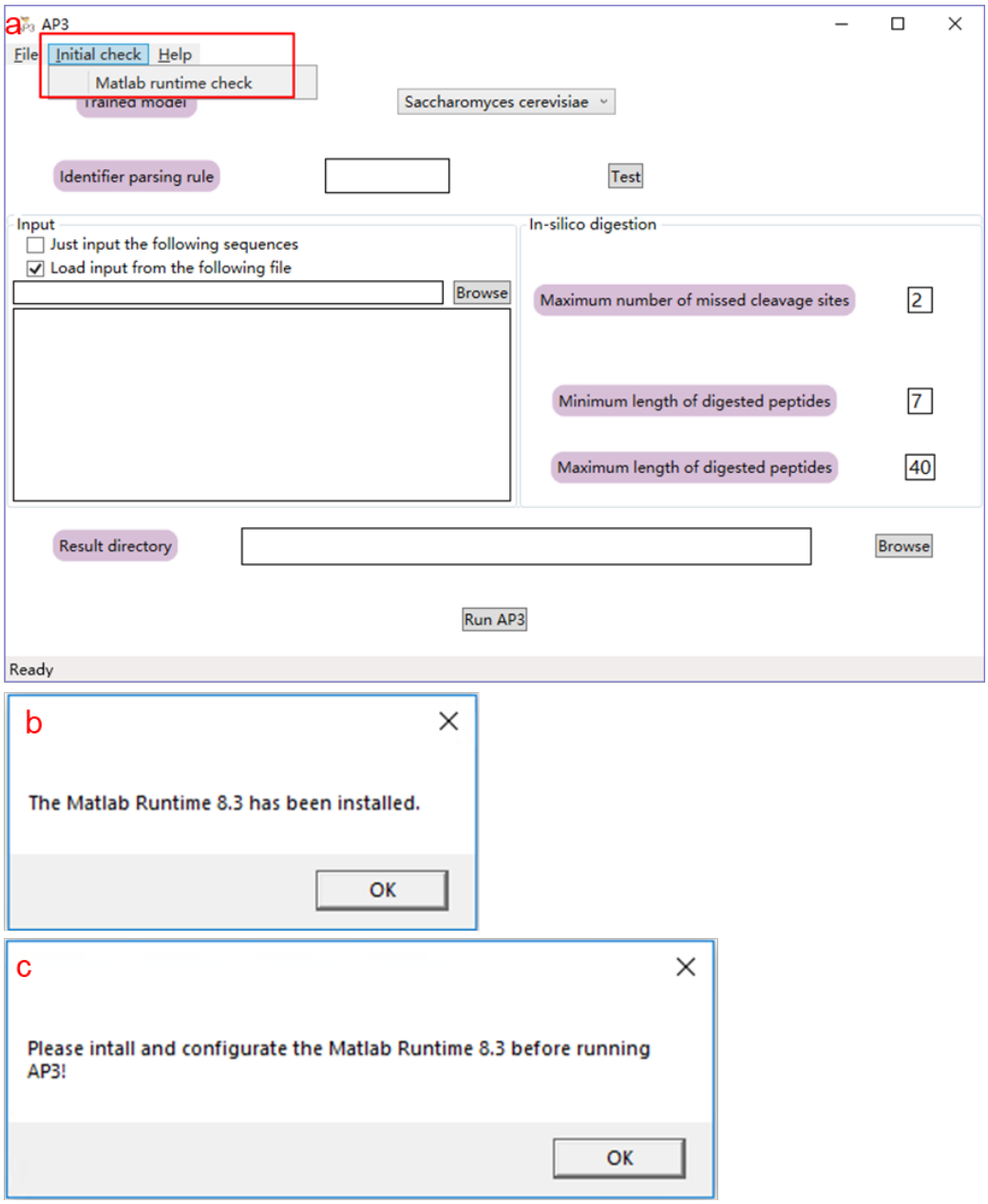

Figure 5. The screenshot of the initial check function of AP3.

3.1.2 Setting parameters in GUI

After completing the preparation work, the user can double-click "GUI_AP3.exe" and set parameters to run AP3. The parameters of AP3 can be divided into three parts: parameters about input, parameters about model and parameters about output.

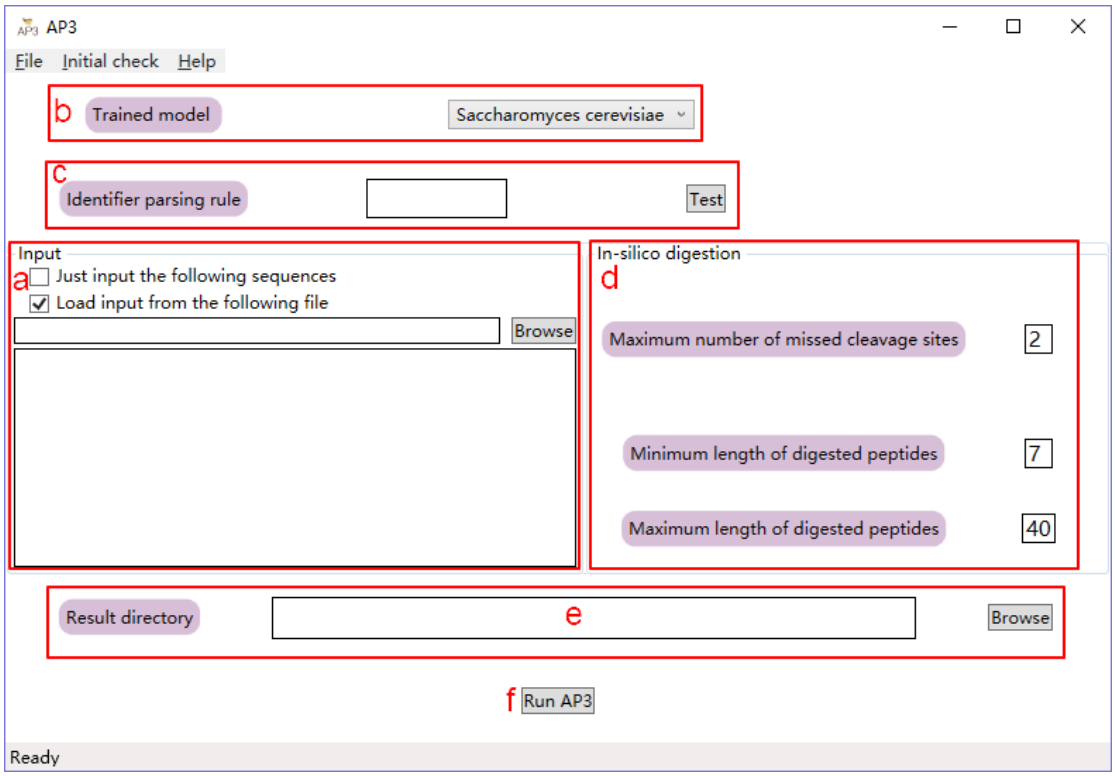

Figure 6. Parameter setting interface for the AP3

Setting parameters about input

1) Load input (Fig. 6a). AP3 needs only the sequences of proteins of interest as input. There are two ways to input protein sequences. The first way is choosing the checkbox "Just input the following sequences" and writing the protein sequences into the textbox. The second way is choosing the checkbox "Load input from the following file" and clicking the "Browse" button to load the protein sequences file in FASTA format. The following textbox will show the first 10 proteins of the input file.

Setting parameters about model

2) Choose the trained model which users would like to use (Fig. 6b). Currently, AP3 provides five kind of models: *Saccharomyces cerevisiae, Mus, Homo sapiens, E.coli* and Custom.

3) Set the identifier parsing rule which is used to extract the protein ID from the protein header line (Fig. 6c). Users can click the button "Test" to open the dialog "Test identifier parsing rule". In the "Test identifier parsing rule" dialog (Fig. 7), users can specify the regular expression which is used to extract the protein identifiers

from the protein sequence as follows:

(a) Click “Browse” button to choose the protein sequence file in FASTA format (Fig. 7a).

(b) Fill in the regular expression (Fig. 7b). There are some examples in the below for reference.

(c) Click “Test” button to check the regular expression (Fig. 7c).

(d) Click “OK” button if the regular expression is confirmed (Fig. 7d).

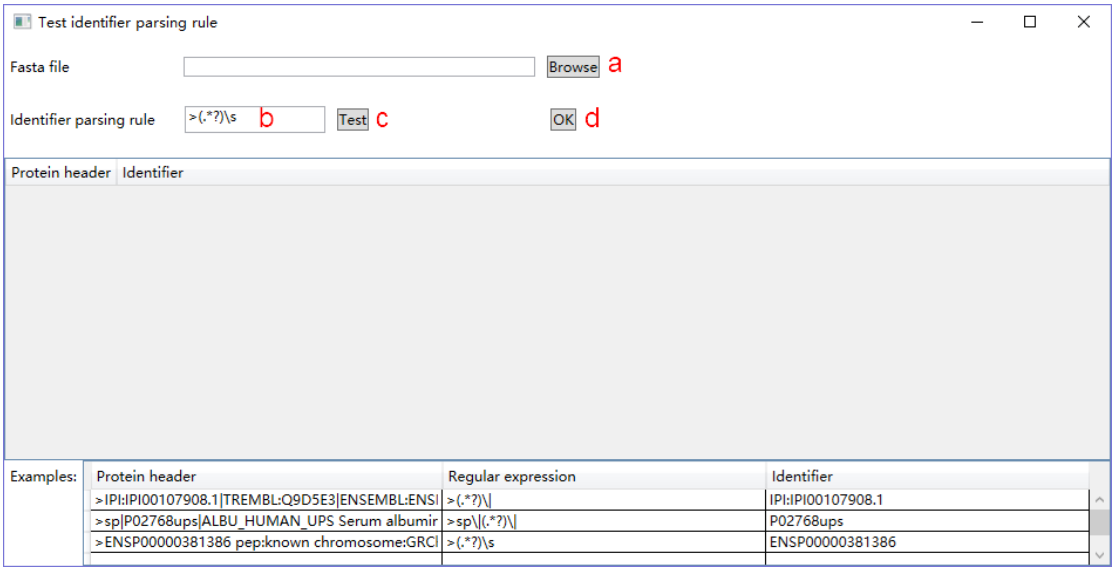

Figure 7. Parameter setting interface for the identifier parsing rule.

4) Set parameters used in the in-silico digestion (Fig. 6d).

Setting parameters about output

5) Choose the directory where the result files will be saved by clicking the “Browse” button (Fig. 6e).

3.1.3 Clicking “Run AP3”

Make sure that all the parameters are appropriate, and then click the “Run AP3” button to start AP3 (Fig. 6f). A parameter file named as the local time (e.g. parameters20181130-13-55-45.param) containing all the parameters setting in the GUI will be generated in the result folder. When the background program is running, the status bar at the bottom of the window would show “AP3 is running!” (Fig. 8). If the background program is done successfully, a dialog showing “AP3 has finished successfully!” would appear (Fig. 9) and the status bar would show “Finished!”.

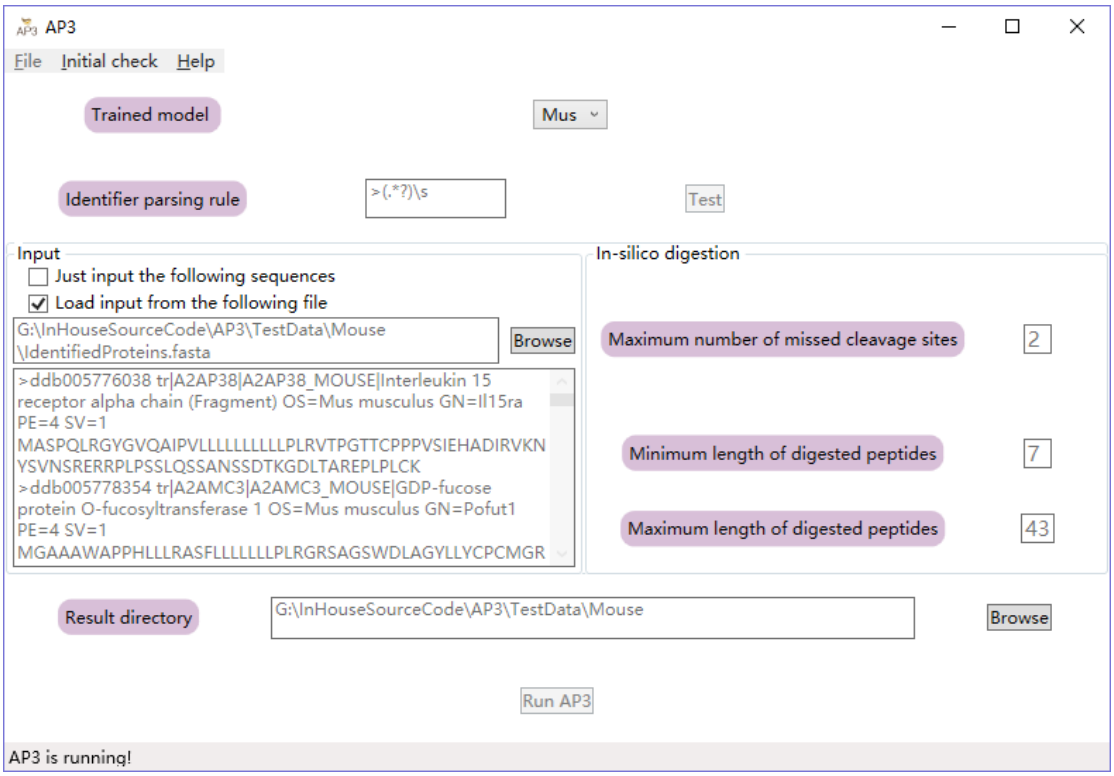

Figure 8. The interface when the background program is running.

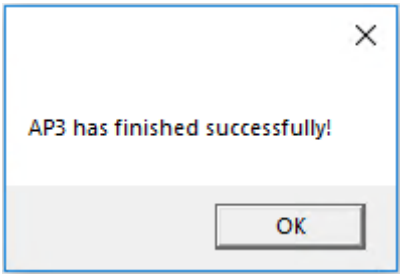

Figure 9. The pop-up dialog when the program finished successfully.

3.2 Using the command line

AP3 can also be run in command line mode. Users should configure the parameter file at first and then call it in the command interpreter.

3.2.1 Configuring parameter file

The running of AP3 requires one parameter file. This file contains the following parameters:

Table 1. Annotations of the parameters in the parameter file

Parameter name	Meaning
InputFilePath	The file path of protein sequence file in FASTA format (.fasta)

Model	The trained model which will be used to predict peptide detectability. Currently, AP3 provides five kind of models: *Saccharomyces cerevisiae, Mus, Homo sapiens, E.coli* and Custom
IdentifierParsingRule	The regular expression used to extract protein id from the protein sequence file
ResultPath	The directory path of result files
AllowMinPeptideLength	The allowed shortest length of peptides in in-silico digestion
AllowMaxPeptideLength	The allowed longest length of peptides in in-silico digestion
AllowMissingCutNumber	The allowed maximum number of missed cleavage sites in in-silico digestion

3.2.2 Running the command

Open the command interpreter "cmd.exe" and run "AP3.exe" by the following format:

>AP3.exe *the path of parameter file*

An example is as follows (Fig. 10):

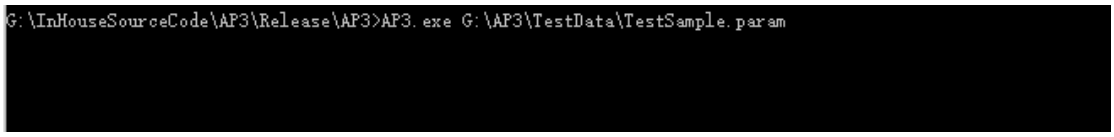

Figure 10. Illustration of running AP3 in command line mode.

4 Output file formats

Once the calculation is done, AP3 generates several files in the result directory (Table 2).

Table 2. Annotations of AP3 result files.

File name	Annotations
parameters*20181130-13-55-45*.param	The parameter file named as the local time

	contains all the parameters used (only generated when running AP3 by GUI).
Log.txt	The log file which is used to record the program process.
InSilicoPeptides.txt	This file contains all in-silico digested peptides of the imported protein sequences.
AllPepCleavagePros.txt	This file contains the predicted digestibility of all digested peptides and associated cleavage sites.
DetectabilitiesOfPeptides.txt	This file contains the predicted peptide detectability of all digested peptides.
PeptideProperties.txt	This file contains the predicted peptide properties of all digested peptides.

4.1 Annotation of AllPepCleavagePros.txt

This file contains the predicted digestibility of all digested peptides and associated cleavage sites. The detailed description of every column in this file is given in Table 3.

Table 3. Descriptions of headers in AllPepCleavagePros.txt.

Name	Description
Peptide sequence	The sequence of the peptide
Protein id	The protein id where the peptide is digested from
Digestibility	The predicted digestibility
L_NineMer	The 9-mer sequence fragment of the left cleavage site of the peptide
L_NineMer_Probability	The predicted cleavage probability of the left cleavage site of the peptide
R_NineMer	The 9-mer sequence fragment of the right cleavage site of the peptide
R_NineMer_Probability	The predicted cleavage probability of the right cleavage

	site of the peptide
M_NineMers	The 9-mer sequence fragment of the missing cleavage sites of the peptide
M_NineMer_Probabilities	The predicted cleavage probability of the missing cleavage sites of the peptide

4.2 Annotation of DetectabilitiesOfPeptides.txt

This file contains the predicted peptide detectability of all digested peptides of the imported protein sequences. The detailed description of every column in this file is given in Table 4.

Table 4. Descriptions of headers in DetectabilitiesOfPeptides.txt.

Name	Description
Peptide sequence	The sequence of the peptide
Protein id	The protein id where the peptide is digested from
Peptide detectability	The predicted peptide detectability of the peptide

5 Retraining solution of AP3

Although we have provided four trained model with different organisms, these models cannot hold good for all cases. So we provide a retraining solution of AP3 for users' custom data set. The source code and executable file for retaining AP3 can be downloaded from our website. Users can run the executable files directly to retrain AP3 as the following steps:

S1) Configure the parameter file for the ConstructAP3SeparateTrainSet.exe. The annotations of these parameters in the parameter file is shown in Table 5. Then run this program using the following command:

> ConstructAP3SeparateTrainSet *parameter_file_path*

This program ConstructAP3SeparateTrainSet will generate the training set of cleavage probability model and save it into the file "CCleavagePredictorTrainXY.txt".

Table 5. Annotations of the parameters in the parameter file of ConstructAP3SeparateTrainSet

Parameter name	Meaning
IdentResultDirectoryPath	The directory of identification result files of MaxQaunt
ProteinSequenceFilePath	The file path of protein sequence file in FASTA format (.fasta)
IdentifierParsingRule	The regular expression used to extract protein id from the protein sequence file
IfExistDecoyProtein	This parameter indicates if the MaxQuant results contain decoy proteins or not
IfExistContaminantProtein	This parameter indicates if the MaxQuant results contain contaminant proteins or not
PrefixOfContaminantProtein	If the parameter "IfExistContaminantProtein" is true, this parameter is the specific prefix of the contaminant proteins
MaxMiss	The SC_M of positive sites in the training set of the cleavage probability model should be less than or equal to MaxMiss
MinLeft	The SC_L of positive sites in the training set of the cleavage probability model should be more than or equal to MinLeft.
MinRight	The SC_R of positive sites in the training set of the cleavage probability model should be more than or equal to MinRight.
MaxLeft	The SC_L of negative sites in the training set of the cleavage probability model should be less than or equal to MaxLeft.
MaxRight	The SC_R of negative sites in the training set of the

	cleavage probability model should be less than or equal to MaxRight.
MinMiss	The SC_M of negative sites in the training set of the cleavage probability model should be more than or equal to MinMiss.
AP3ResultPath	The directory path of the result files

S2) Change two variables in the MATLAB script "CleavageModelByRT.m": the variable "datapath" which means the file path of "CCleavagePredictorTrainXY.txt" and the variable "resultpath" where the trained model will be saved. This script will output the trained cleavage probability model into the file "CleavageModel.mat".

S3) Copy the file "CleavageModel.mat" to the AP3ResultPath. Run the executable file ConstructAP3SeparateTrainSet.exe again. This program can automatically detect if there is a trained cleavage probability model in the AP3ResultPath. If the "CleavageModel.mat" exists, ConstructAP3SeparateTrainSet will generate the training set of the peptide detectability model and save it into the file "DetectabilityTrainData.txt".

S4) To run the feature selection method "mRMR", the file "DetectabilityTrainData.txt" should be changed to the CSV format "DetectabilityTrainData.CSV", where each row is a sample and each column is a feature. Make sure the data is separated by comma, but not blank space of other characters! The first row must be the feature names, and the first column must be the classes for samples. Refer to the website http://home.penglab.com/proj/mRMR for details. One example of running mRMR is as follows (**Figure 11**):

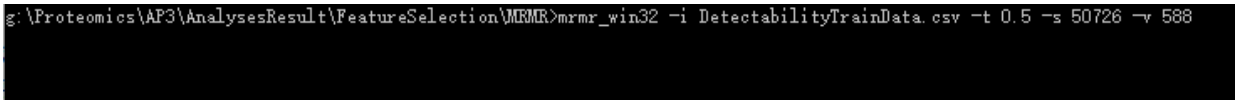

Figure 11. Illustration of running mRMR in command line mode.

Save the feature selection result into the file "MrmrResult.txt" and the first 29 features into the file "SelectedFeaturesOfMRMR.txt".

S5) Change three variables in the MATLAB script "DetectionModelByRT.m":

variable “datapath” which means the file path of “DetectabilityTrainData.txt”, variable “featurepath” which means the file path of “SelectedFeaturesOfMRMR.txt”, variable “resultpath” where the trained model will be saved. This script will save the trained model into the file “DetectionModel.mat” and parameters used in normalization into the file “MuSigma.ini”.

S6) Copy the files “CleavageModel.mat”, “DetectionModel.mat”, “SelectedFeaturesOfMRMR.txt” and “MuSigma.ini” into the directory “SupportFiles\Models\Custom” in the AP3 project. At this point, the retraining of AP3 has been finished.

S7) When using the retrained AP3, the parameter “Model” should be set to “Custom” and other parameters should be set following the description in **Table 1**.