Cytidine-to-Uridine Recognizing Editor for Chloroplasts

A Chloroplasts C-to-U RNA editing site prediction tool

A User Manual

Pufeng Du, Liyan Jia and Yanda Li

MOE Key Laboratory of Bioinformatics and Bioinformatics Div, TNLIST
Department of Automation, Tsinghua University

Important Notice:
- Using CURE implies the agreement of the license terms. See the last section of this manual for details.
- It is strongly recommended that users read this manual and the algorithm document carefully before applying CURE in the research work. Using prediction software without understanding its mechanism may be very risky and lead to false discoveries.
CURE License

Whoever uses the CURE service is regarded as a "user" of CURE. Users agree that the downloaded package of CURE and result obtained from CURE service will be used for academic purpose ONLY, and agree that any work based on or directly related to the CURE software/service and/or its documentations will cite the original work by Pufeng Du and Yanda Li. (see below for details) We also should be informed for such publications. For potential commercial users, please contact the author of CURE before using it.

CURE uses NCBI BLAST as a part of its function. Please refer to the license of NCBI BLAST for the original license agreement of BLAST program.

No Warranty
Because the program is licensed free of charge, there is no warranty for the program, to the extent permitted by applicable law. Except when otherwise stated in writing the copyright holders and/or other parties provide the program "as is" without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the quality and performance of the program is with you. Should the program prove defective, you assume the cost of all necessary servicing, repair or correction.

In no event unless required by applicable law or agreed to in writing will any copyright holder, or any other party who may modify and/or redistribute the program, be liable to you for damages, including any general, special, incidental or consequential damages arising out of the use or inability to use the program (including but not limited to loss of data or data being rendered inaccurate or losses sustained by you or third parties or a failure of the program to operate with any other programs), even if such holder or other party has been advised of the possibility of such damages.

Citation requirement
Anyone who is publishing any work based on or directly related to this CURE software/service should cite at least one of the following publications
Contents

1. Basic knowledge of C-to-U RNA editing ...................................................... 4
2. Understanding CURE Architecture ............................................................. 4
3. Using online CURE service ........................................................................ 5
   3.1 The main interface ................................................................................... 5
   3.2 Entering sequences .................................................................................. 6
   3.3 Using advanced mode of CURE .............................................................. 6
   3.4 Adjusting the parameters ....................................................................... 7
4. Viewing the result .......................................................................................... 7
   4.1 Result browser online ............................................................................... 7
   4.2 Interpreting the downloaded result ............................................................ 8
5. Using local version of CURE ......................................................................... 9
   5.1 Downloading CURE local version and other resources ......................... 9
   5.2 Using command line ................................................................................ 9
6. Suggestions and comments from the designer .......................................... 11
7. Questions and comments ............................................................................. 12
1. Basic knowledge of C-to-U RNA editing

In the angiosperm mitochondrial genome, hundreds of cytidines are selectively converted to uridines in the primary transcripts. For the last two decades, such records of various organisms accumulate in the database GenBank. Until the year 2007, the total number of experimentally detected C-to-U RNA editing event in plant mitochondria is more than 3000.

Unfortunately, the site selective mechanism is still not clear for this type of RNA editing. Since the sequence of plant mitochondrial genome is accumulating very fast, the decision of the editing status of these genomes is very emerging required. CURE, the Cytidine-to-Uridine Recognizing Editor, which is a computational predictor for C-to-U RNA editing was developed. CURE can computationally predict over 80% of the C-to-U RNA editing with only the knowledge of sequence. At the same time, CURE produces almost no false positives in the prediction result.

2. Understanding CURE Architecture

The CURE predictor is implemented using B/S architecture. Due to the limitation of the server resources, the implementation of CURE service is based on a Java/PHP hybridization technology. The hybridization technology also gives the user an alternative choice of using CURE locally. There is a packaged command line based java program that can be executed on any PC, which is called CURE local version. Figure 1 shows the architecture of CURE.

![Diagram of CURE Architecture](image)

Fig. 1. The architecture of CURE. The arrows connecting the objects shows the data flow directions between the objects. The CURE core shadow is a partial copy of CURE core. CURE local shell is a command line shell for the users to call limited core function.
In the following part of this manual, the online system usage and the local version usage will be introduced. Especially, some skills of adjusting parameters will be demonstrated.

### 3. Using online CURE service

#### 3.1 The main interface

The CURE service has a well designed web-like interface which is based on JavaScript and DHTML technology. Figure 2 shows the main interface of CURE and the name of each part on this interface.

![Fig. 2. The main interface of CURE](image)

There are two working mode of CURE, the basic mode and the advanced mode. Checking the checkbox which is the working mode switch can switch between these two working modes. The current working mode can be seen by the title bar of the interface. The details of using different working mode will be discussed in the following sections.

All the “pink books” can be clicked to open. They will show the instructions of the corresponding parameters. Of course, click the opened books will result in closing them.

The main submit button will collect the information you entered on the interface and transfer them to the server. The server will automatically begin the prediction
procedure according to your requirements.

### 3.2 Entering sequences

There are two ways of entering sequences. One is to paste single raw sequence in the input area; the other is to upload a FASTA file containing several sequences to predict. To switch between this two input modes, click the input mode switch (The text “[Change]”).

If you need to upload a FASTA file, remember to restrict your file size within 2MB. If you paste single raw sequence in the input area, remember to restrict the sequence length within 1Mb.

Since the prediction procedure includes a BLAST scanning, the longer the sequence you entered, the longer the time the server needs.

### 3.3 Using advanced mode of CURE

In the basic mode of CURE, no parameter is allowed to set. All parameters will be automatically set to default values. The default values of these parameters can be found in Figure 3. In the advanced mode of CURE, the input field of all parameters will be enabled. Figure 3 shows the interface after enabling the advanced mode.

![Advanced mode of CURE](image)

**Fig. 3.** Advanced mode of CURE

The EPES training set options can be set to CDS or genome-wide. The details of
CDS and genome-wide training set are discussed in the published paper.

### 3.4 Adjusting the parameters

The parameters of CURE affect its performance seriously. You need to adjust them for your own purpose. The most important parameter is the BLAST e-value cut-off parameter.

As everyone knows, the BLAST e-value is related to the size of searching database. In CURE, the sequence you entered is used as a database of BLAST. So the reasonable e-value cut-off depends on the sequence you entered. The longer sequence you entered, the larger cutoff you may consider. If you use a whole mitochondrial genome sequence as the prediction subject, the e-value can be set to 0.01. If you enter a short sequence, for example, 200nt, you may need to set the e-value cut-off to 0.001 to get more reliable results.

### 4 Viewing the result

#### 4.1 Result browser online

The prediction result of CURE online service will be returned in a page call “Result Browser”. This browser has 4 parts, the job summary, the sequence filtering report, the prediction details and the result downloading engine. Figure 4 shows the interface of this Result Browser.

The job summary gives a brief summary of the prediction you requested. The sequence filtering report is the sequence preprocessing report. This report is collapsed by default. You may click the link “Show detail” to expand it. This report will show all the filtering procedure on the sequences you submitted. All the symbols which are not standard ACGT will be kicked out from the sequence. The prediction result is a multi-page result list. Every 20 prediction result forms a new page in the prediction result part. You may navigate in these pages using the four links in yellow cells. The last line of this part shows the total number of predicted results and the page of result which you are viewing.

The last part is the download engine. You may download the result file using your job ID. The job ID is generated randomly and has no other meanings. Because of the garbage collection mechanism we used, all result files will be cleared automatically without warn at a time that is impossible to predict. You would better download your
result immediately after you have seen the result page. If you require a job ID that is not exists, you will receive an error message.

Fig. 4. The online result browser.

4.2 Interpreting the downloaded result
The downloaded result format is not the same as you have seen in the prediction details part. It is a raw text file with 2 columns. The first column is the sequence name. The second column is the editing site predicted. The most important is the interpretation of the “minus value” in the second column.

The “minus” symbols which are in the second column has exactly the same meaning as the “complementary ()” function in the GenBank remarks. You may refer to the document of GenBank for the detail of interpreting this “function”. Do not explain the “minus” symbol as the minus strand simply, that is not correct.

5 Using local version of CURE

5.1 Downloading CURE local version and other resources

The local resource of CURE include 5 parts

(1) The User Manual of CURE
   This file.

(2) The Algorithm Document of CURE
   The document describing the algorithm of CURE.

(3) The local version of CURE
   A JAR file containing the local version of CURE.

(4) The EPES Library of PM_CDS and GE_CS dataset
   A zip file containing 2 text files.

(5) The evaluation purpose sample dataset of CURE
   A FASTA file containing several sequences.

The local version of CURE is a JAR package which can be executed directly. To execute the local version of CURE, you need to obtain the cure.jar file and the EPESLIB.zip. Do not extract the files in cure.jar and just put the downloaded files in an independent directory. Extract two text files from the EPESLIB.zip file. Next step is to setup and configure the Java Runtime Environment. You may refer to http://www.java.com for details of this step. After the JRE is installed and configured. You may start a console in your operating system (Windows or Linux, either is OK) and begin to use CURE local version.

5.2 Using command line

The syntax of executing CURE local version is:
### Java –jar cure.jar <operation> <parameters>

The `<operation>` is a parameter which will be passed to the local version of core system to specify the operating mode of the core system. Currently, only “-predict” is allowed in the CURE local version.

The `<parameters>` is a serial of parameters which will be passed to the core system to make the prediction. The format of this `<parameters>` is:

```
epes_lib fasta_file e_value_cut_off word_size up_bound low_bound k
```

The meaning of each parameter is listed in Table 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>epes_lib</td>
<td>A filename for a EPESLIB. This can be set to either the CDS EPESLIB or the genome-wide EPESLIB.</td>
</tr>
<tr>
<td>fasta_file</td>
<td>The FASTA file containing the sequence which you want CURE to predict their RNA editing sites.</td>
</tr>
<tr>
<td>e_value_cut_off</td>
<td>The parameter which will be passed to BLAST as the e-value cutoff for selecting HSPs. It should be a very small positive number, like 0.001.</td>
</tr>
<tr>
<td>word_size</td>
<td>The parameter which will be passed to BLAST as the word size parameter. It should be an integer between 4 and 11.</td>
</tr>
<tr>
<td>up_bound</td>
<td>The upper boundary of the conservative ratio region where the micro-analyzers are used to decide the editing status of a cytidine.</td>
</tr>
<tr>
<td>low_bound</td>
<td>The lower boundary of the conservative ratio region where the micro-analyzer are used to decide the editing status of a cytidine.</td>
</tr>
<tr>
<td>k</td>
<td>The micro-analyzer parameter k in the K-NN algorithm.</td>
</tr>
</tbody>
</table>

A sample of executing the local CURE is here

```
Java –jar cure.jar –predict epes_cds_all.txt test.fas 0.001 4 0.7 0.5 1
```

This sample uses the EPESLIB of CDS to predict the sequences in test.fas and the
parameters a set to the default value of CURE.

6 Suggestions and comments from the designer

The local version of CURE is only an alternative choice of CURE service. The implementation of CURE local version is based on the core system of CURE service. However, the pre-process and post-process of the CURE online service are not included in the CURE local version. The result of CURE local version may have little difference to the result of CURE online service. Especially, the local package require some extra condition to run. To make sure that the local version of CURE can be executed normally, you need to check the following two points before you begin to type any command.

(1) The local version of CURE needs to call NCBI BLAST. You need to make the program `formatdb` and the program `blastall` executable from the current directory.

(2) The local version of CURE needs to connect to NCBI through internet to parse the BLAST result. You need to make sure that you have an internet connection to NCBI and the firewall is properly configured to allow CURE to access internet.

From the view of designer, there are still some important points we need to make it clear.

(1) Though we have not use any web tech that is related to a particular web browser, we still recommend the users to use Internet Explorer version 6 or higher with the latest JavaScript support. Because we do not test the service on other browsers, we are not sure whether it will look the same as that on IE.

(2) Theoretically, the length of input sequence has no restrictions. But, the longer sequence you input, the more false positive you will got. The CURE designing purpose is to make sure that when scanning a 1Mb genome sequence containing about 500 real editing sites, CURE will report more than 400 of real editing sites and less than 100 false positives.

(3) The computation speed, as out experience while evaluating CURE, is mainly depend on the internet connection speed to NCBI. The calculation time of scanning the whole *Arabidopsis thaliana* mitochondrial genome is less than one second if we ignore the time of communications with NCBI (On a PC with Pentium D 820 CPU and 2GB memory.).

(4) Though the CURE can predict over 85% of all editing sites, we still recommend
you to try different algorithms for your own research work. We have to admit that no algorithm is perfect. Some other algorithm may be more useful than CURE for your particular situation.

(5) Computational prediction is not a 100% reliable result. So use the result at own risk.

7 Questions and comments

The questions and suggestions are welcome. You may email your questions and comments to Pufeng Du using the address dpf05@mails.tsinghua.edu.cn. We would like to try our best to improve the CURE service.