VisFeature: a stand-alone program for visualizing and analyzing statistical features of biological sequences

A User Manual

Contact	**Jun Wang wj0708@tju.edu.cn**
Date	June. 8th, 2019
Version	1.0

Contents

1 Introduction

In order to predict biological functions and cellular attributes, one of the most challenging problems is to find a valid approach to visually represent the features of biological sequences. Therefore, it is important to analyze the features of biological sequences. We developed VisFeature, a stand-alone program for visualizing and analyzing statistical features of biological sequences. It provides a function for retrieving a number of biological sequences automatically and integrates over 30 sequence representation modes for extracting the features of DNA, RNA and protein sequences rapidly. It can also convert sequences into curves and density maps based on their physicochemical properties or feature vectors. By using the feature visualization, the differences between different sequences or groups can be observed, which is convenient for sequence analysis.

There are four main functions in VisFeature: “Fetch data”, “Single sequence mode”, “Multiple sequences mode” and “Density map comparison”.

You can use “Fetch data” to download a large number of DNA, RNA and protein sequences by identifier or query expression. You can use “Single sequence mode” to visualize a DNA, RNA or protein sequence by physicochemical properties. You can use “Multiple sequences mode” to visualize multiple DNA, RNA or protein sequences by physicochemical properties. You can also perform multiple sequence alignment in this function. You can use “Density map comparison” to

calculate feature vectors from DNA, RNA and protein sequences. After the calculation, a label file should be uploaded to assign group labels to each sequence. You can use this function to visualizing and comparing feature vectors as density maps.

2 Usage

2.1 Installation

We have tested these codes on **Windows10-64bit platform** and **Ubuntu 16.04.5 LTS platform** . There is no guarantee that these codes can be compiled and executed on other platforms without modifications. The usage of the **Windows version** of VisFeature will be described later.

For ***Microsoft Windows platform***, just download the `VisFeature-win32-x64.zip` package from https://github.com/wangjun1996/VisFeature/releases. Unpack it to your favorite location and then open `VisFeature.exe`.

Note:

- Some anti-virus software may report a risky warning when you first run VisFeature. Please do not worry about it and you can ignore it.
- Recommended memory size: **8GB or larger**. If the memory is too small, it will cause a large file to open without response and the program will run unsmoothly.

The figure below shows the location of the executable files on Windows platform.

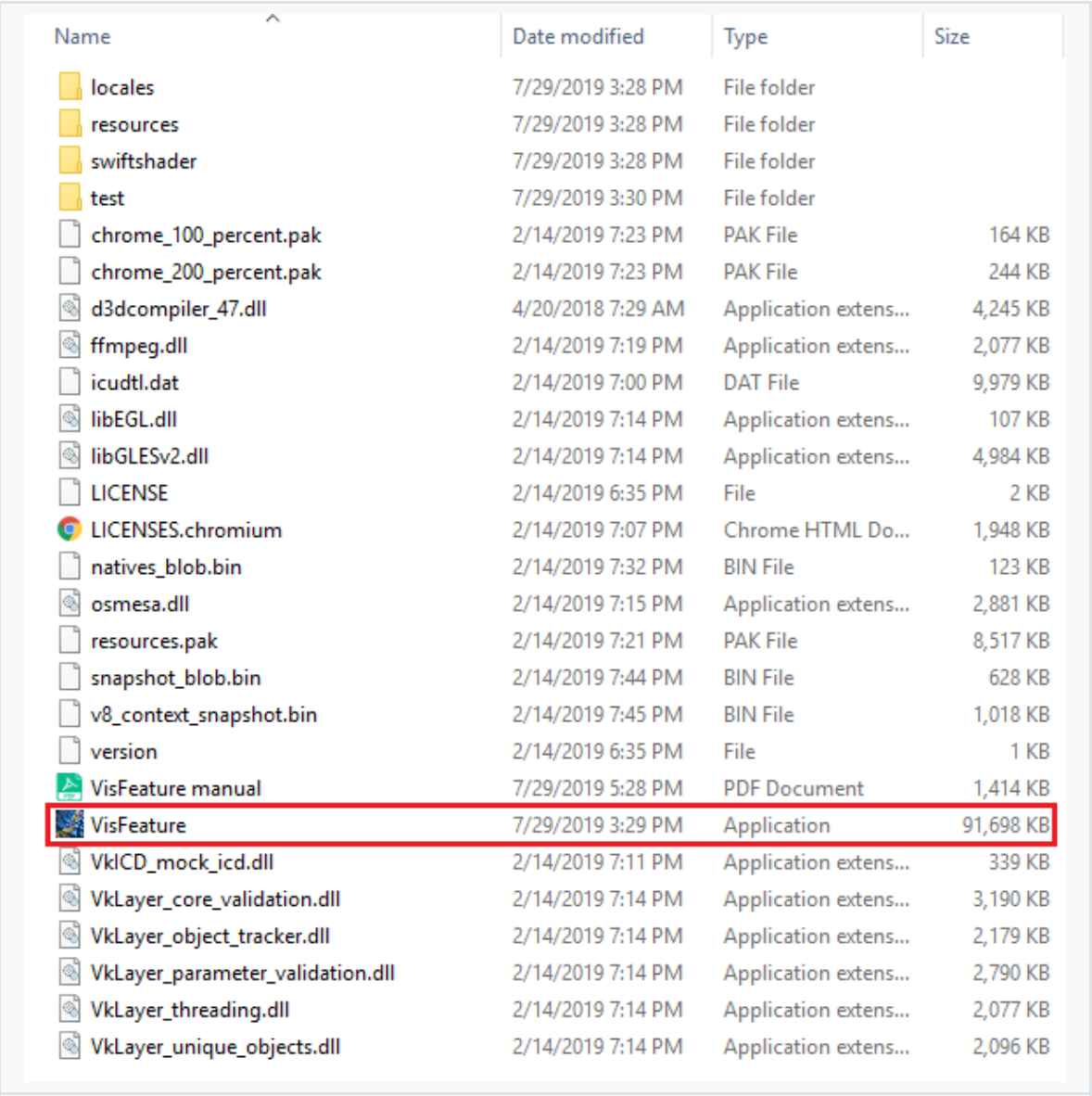

Figure 1 - Location of the executable files on Windows platform

Right-click on "VisFeature.exe" and select "Open" to start VisFeature. The figure below shows how to start VisFeature on Windows platform.

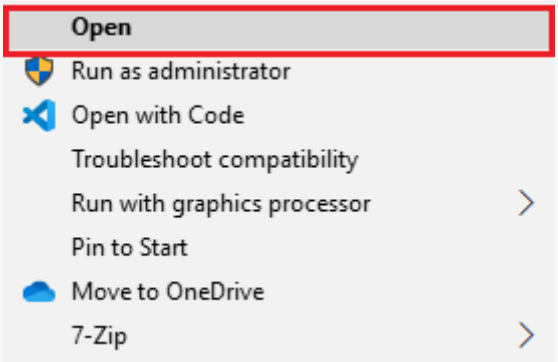

Figure 2 - Start VisFeature on Windows platform

2.2 Input

There are three ways to input into VisFeature: keyboard type, paste from clipboard, and open a file.

Note:

- **Start a quick experience**: Small example data are prepared for VisFeature. Tiny DNA, RNA, and protein sequences are used as sample data. See below for details. You can right-click on the edit area and then select the example of DNA, RNA, or protein sequences by clicking "Example Fasta" item to use these examples as input. See below for details.

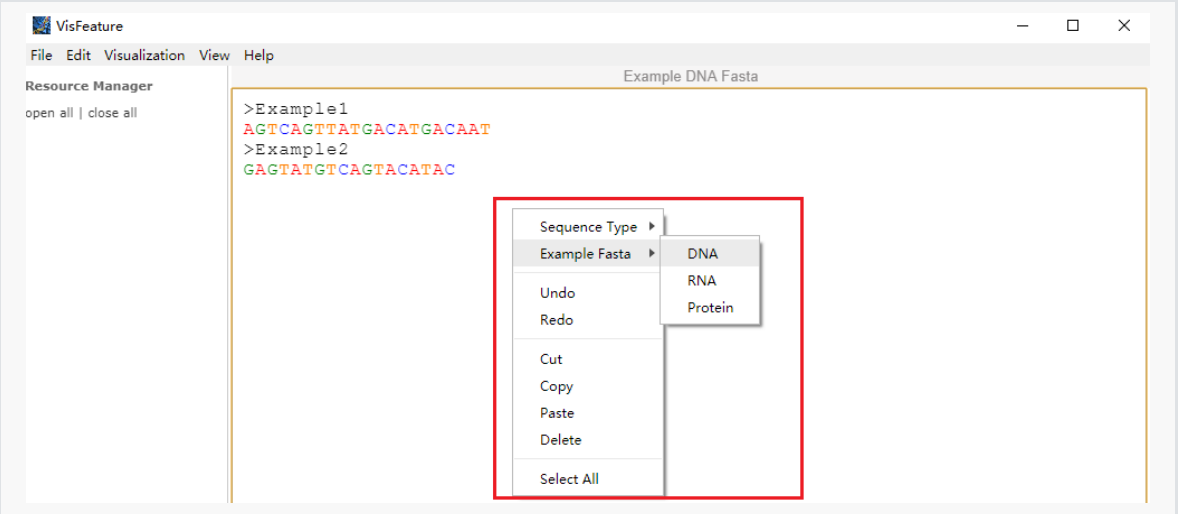

Figure 3 - Load sample data

- **Load test data**: 811 DNA sequences from iORI-PseKNC(Li et al., 2015) were used as test data. We divide these sequences into three files with FASTA format in the test folder, which are "ORI.fas", "non-ORI.fas" and "ORI-and-non-ORI.fas". The "label.csv" in the test folder is used for grouping these sequences before generating density maps. You need to click "Open Example Folder" from "File" in the menu bar to display the test folder in resource manager on the left side of the main page. Then, you can click a file or a folder in the resource manager to open it. See below for details.

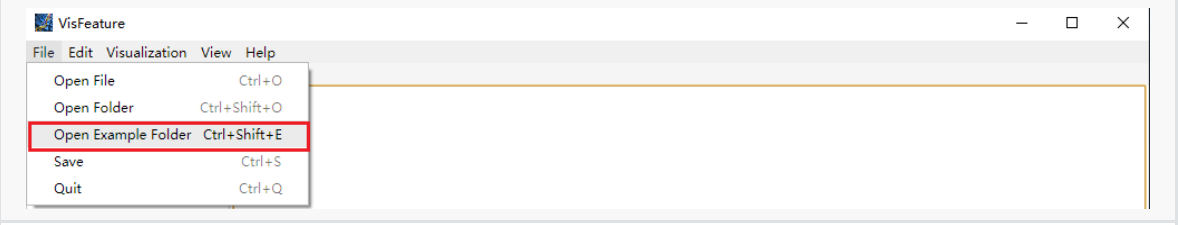

Figure 4 - Load test data: menu

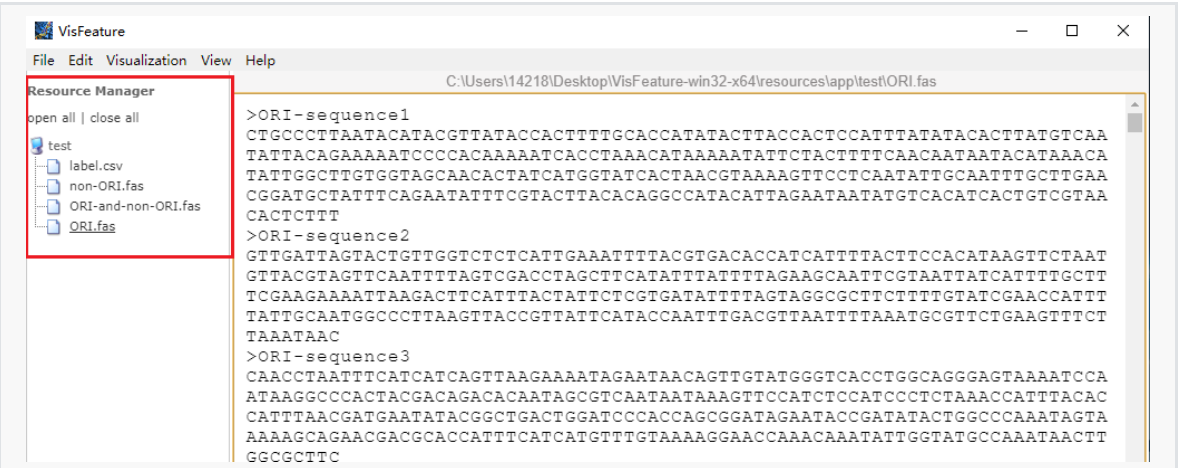

Figure 5 - Load test data: interface

- **Set sequence type to accelarate**: You can right-click on the edit area and select the type of sequence by clicking "Sequence Type" to change the color of letters. It is important to note that if you edit or paste or open a large file, please set the sequence type to **"None"** first, which can improve the speed of the program significantly. See below for details.

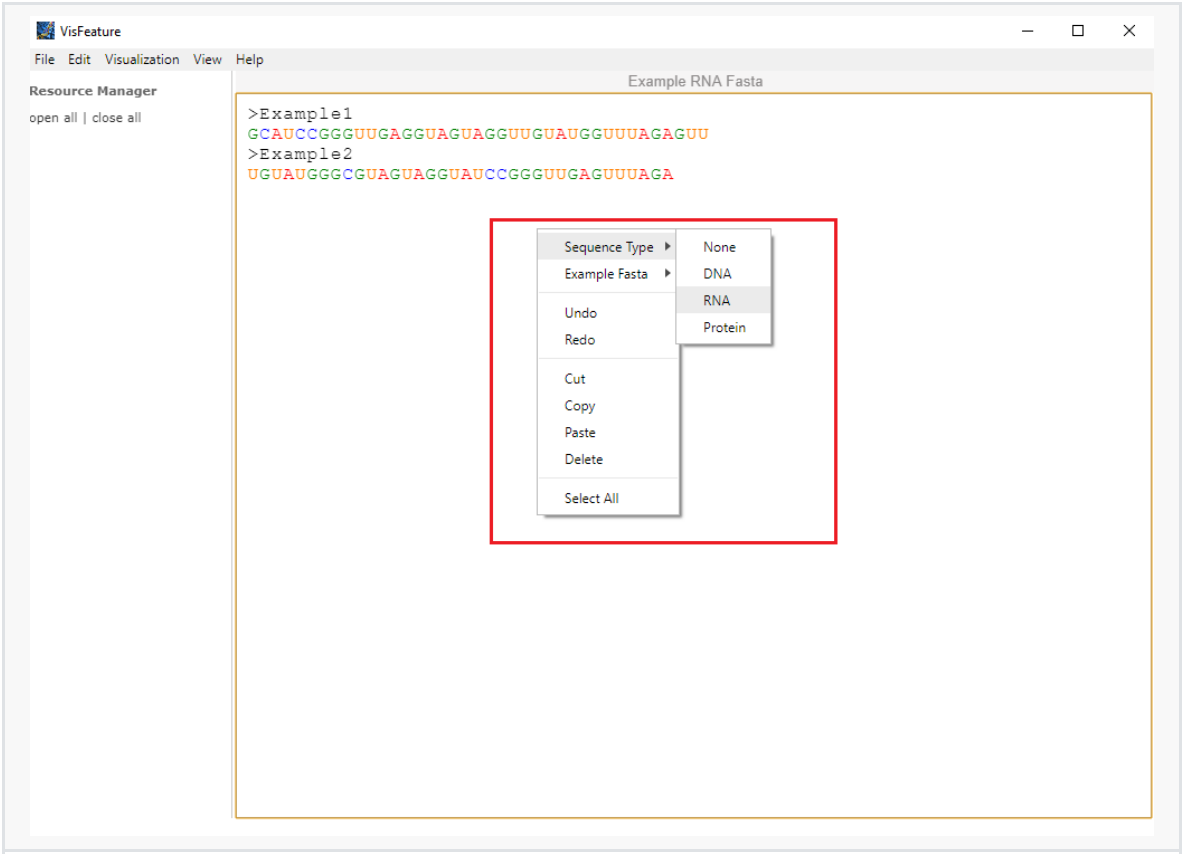

Figure 6 - Sequence type selection

2.2.1 Type by keyboard

You can type content in the editing area of the main page by the keyboard. See below for details.

Figure 7 - Type sequences

2.2.2 Copy and paste from "Fetch Result"

Click "Fetch Data" from "Tool" in the menu bar, then complete all parameters. Finally, input identifiers or query expression and click “Fetch” to fetch data of sequence. You can then save the fetched data as a FASTA file or a text file. Or, you can also copy-and-paste the fetched data from the fetch data window to the main window.

Note:

- There are two options in "Fetch priority". They are "Speed priority" and "Order priority". If you select “Speed priority”, this will make the speed of fetch faster. There is **no guarantee** that the order of sequences of the fetch result will be the same as the order of input. If you select "Order priority", the order of sequences of the fetch result will be the same as the order of input, but this result in a slower speed than "Speed priority".

Figure 8 - Fetch Data: step 1

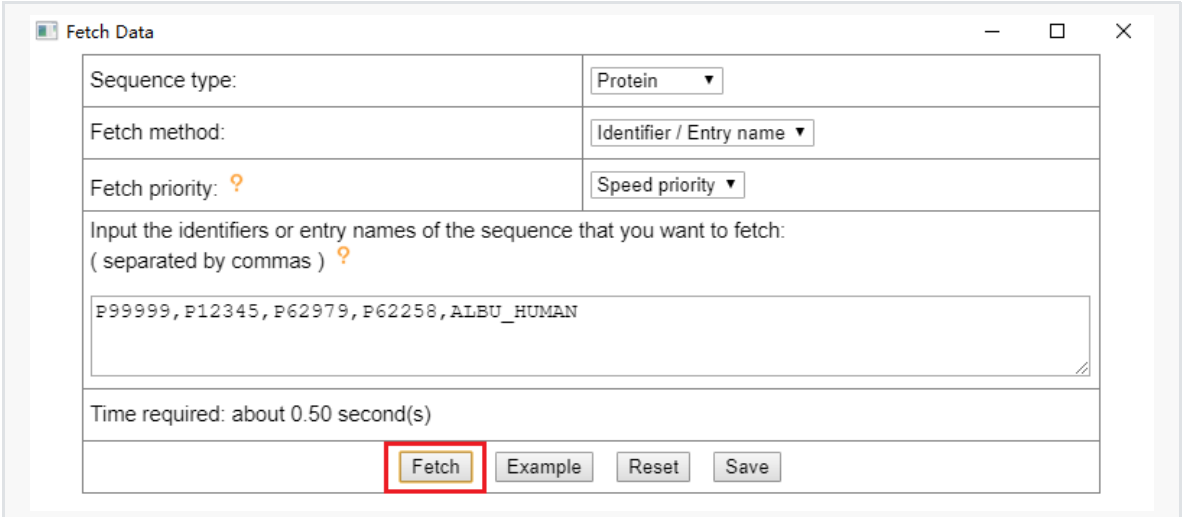

Figure 9 - Fetch Data: step 2

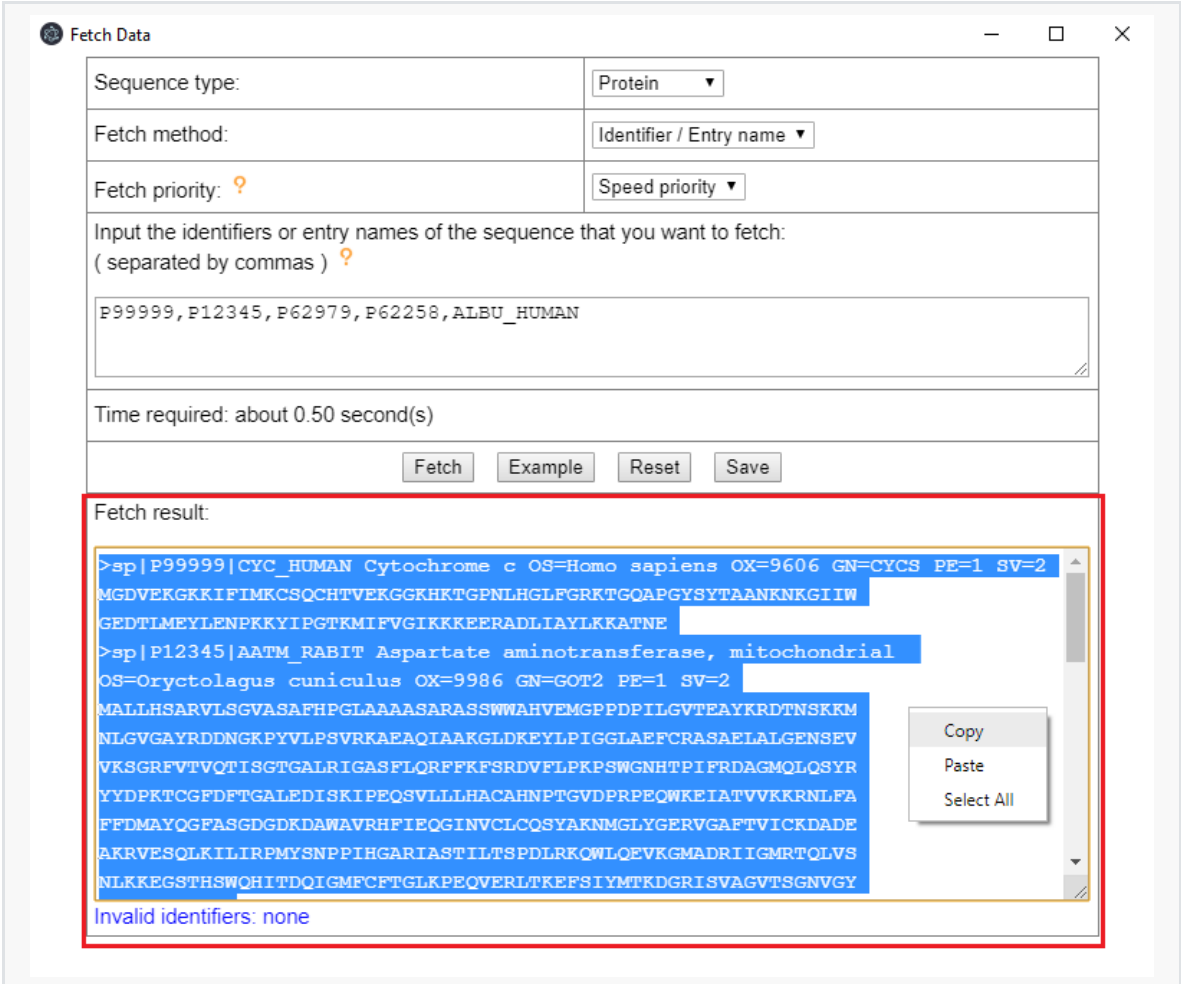

Figure 10 - Fetch Data: step 3

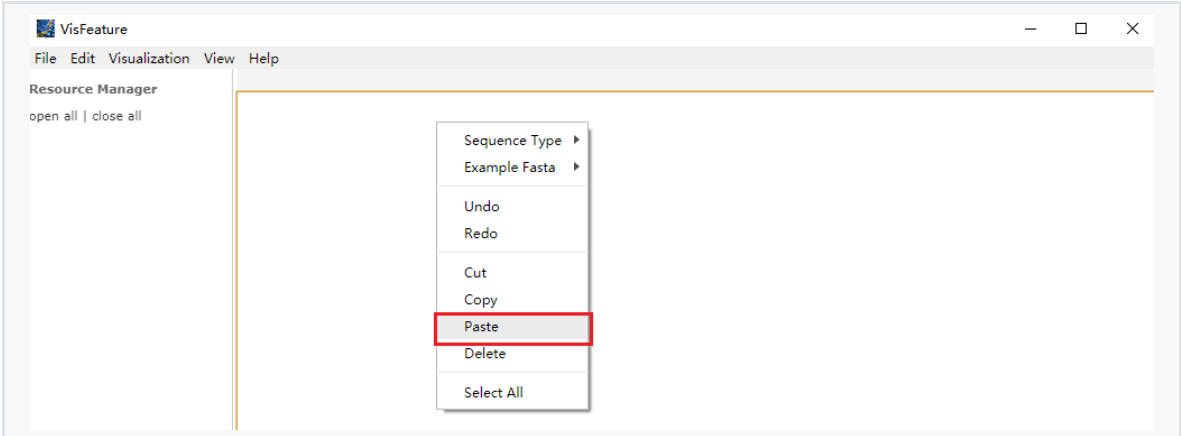

Figure 11 - Fetch Data: step 4

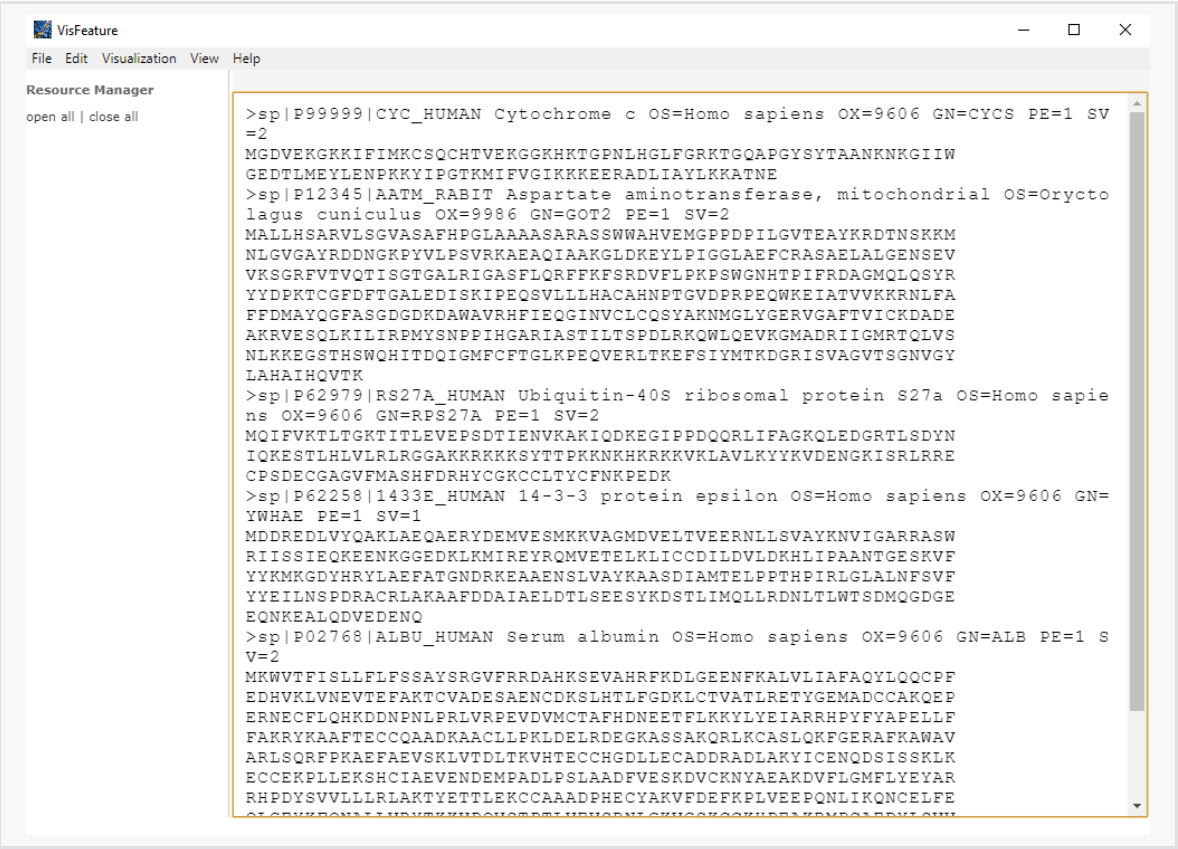

Figure 12 - Fetch Data: step 5

2.2.3 Open a file

Two methods to open a file in the VisFeature.

- **Method 1**: Open a file by clicking "Open File" from "File" in the menu bar. The maximum size of file that can be opened is **5MB**. If you want to use a file that larger than 5MB as input, you can **upload** it on the page of "Density map comparison" mode. Uploading file is much faster than opening file, so upload is more recommended.

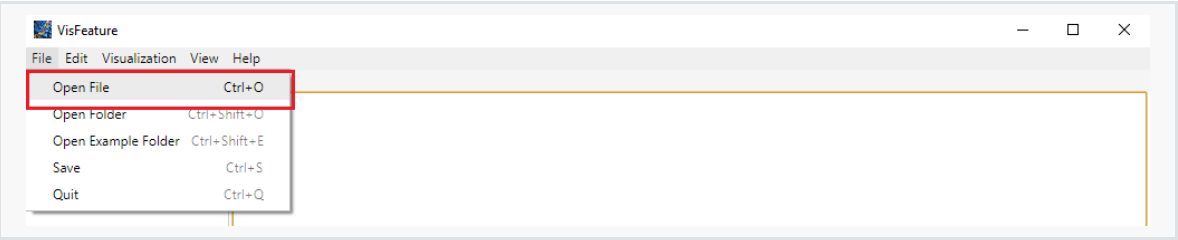

Figure 13 - Directly open a FASTA file: step 1

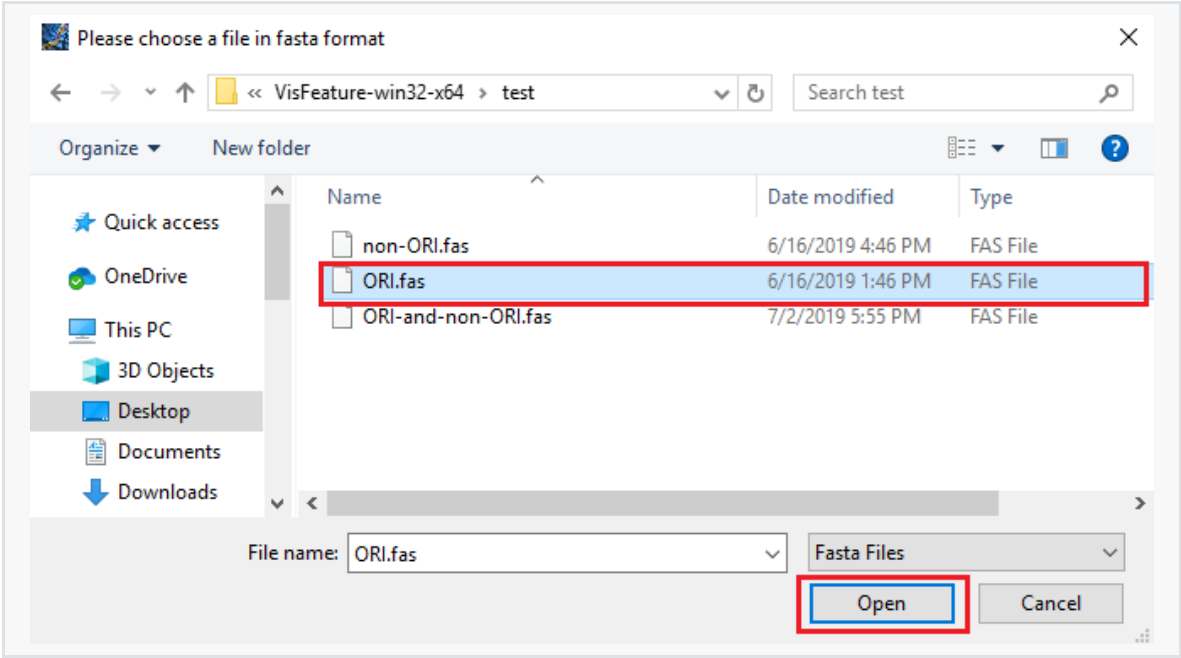

Figure 14 - Directly open a FASTA file: step 2

Figure 15 - Directly open a FASTA file: step 3

- **Method 2**: Open the file from the resource manager. First, you need to open a folder, then the resource manager will display the contents of this folder. You can open a file by clicking on it. Files with fas, fasta, csv and txt suffixes will be displayed in the resource manager.

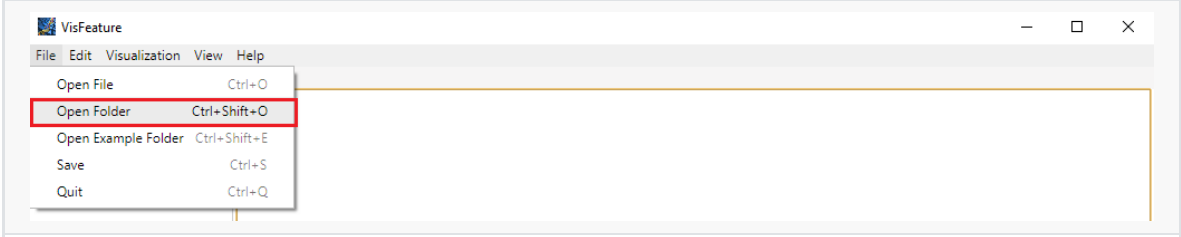

Figure 16 - Open a folder: step 1

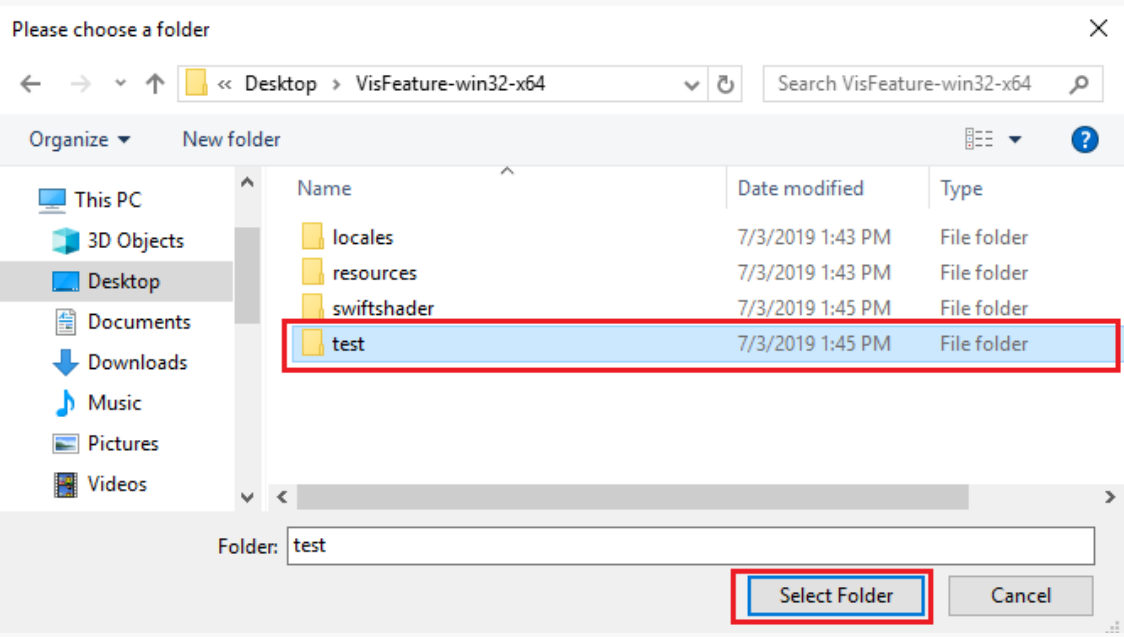

Figure 17 - Open a folder: step 2

Figure 18 - Open a folder: step 3

2.3 Visualization

There are three visual modes in VisFeature, which are “Single sequence mode”, “Multiple sequences mode” and “Density map comparison”.

In order to meet different requests, you can use "Zoom In" and "Zoom Out" for scaling from "View" in the menu bar. Of course, you can use them on other pages. Details will be given below.

2.3.1 Single sequence mode

First, you need to click "Single sequence mode" from the "Visulization" in the menu bar, then complete all parameters and click “Submit”. The program will jump to the visual page. Visual page will display the curves according to the sequence and physicochemical properties that you select, where the dotted lines represent the average level of that physicochemical properties.

Note:

- You can set the **maximum number** of selectable physicochemical properties on the page for setting parameters. The excessive number of physicochemical properties selected will result in too many curves on the page of visualization, which will affect the visual effect.

When the number of physicochemical properties selected is large, the rendering speed of curves will become slower and the delay of the operation will increase.

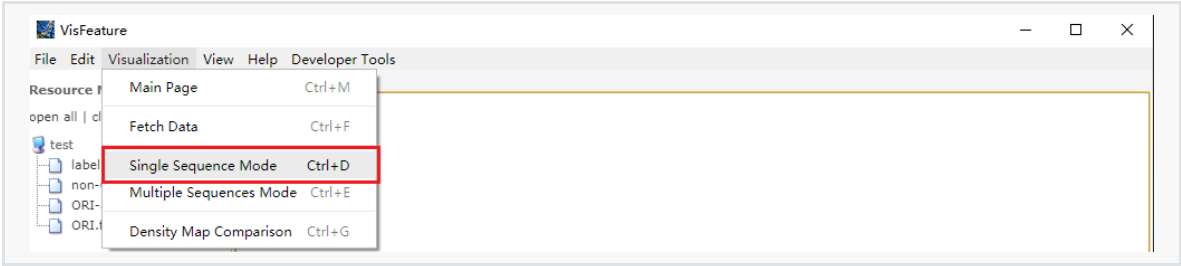

Figure 19 - Single sequence mode: step 1

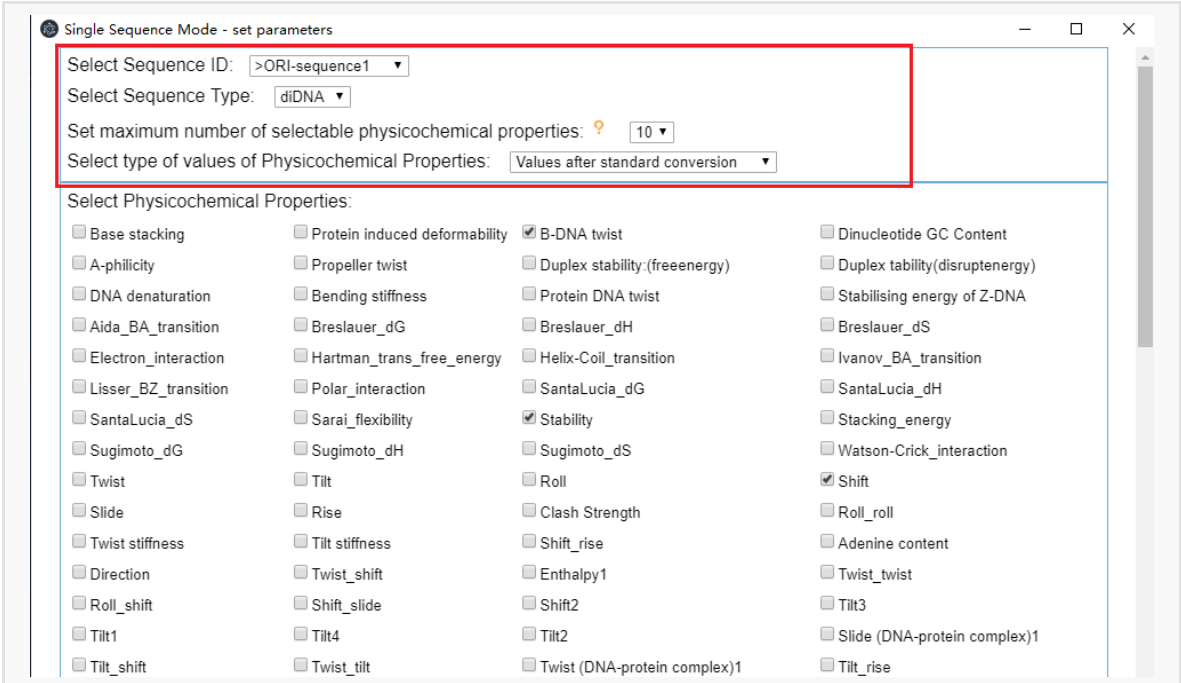

Figure 20 - Single sequence mode: step 2

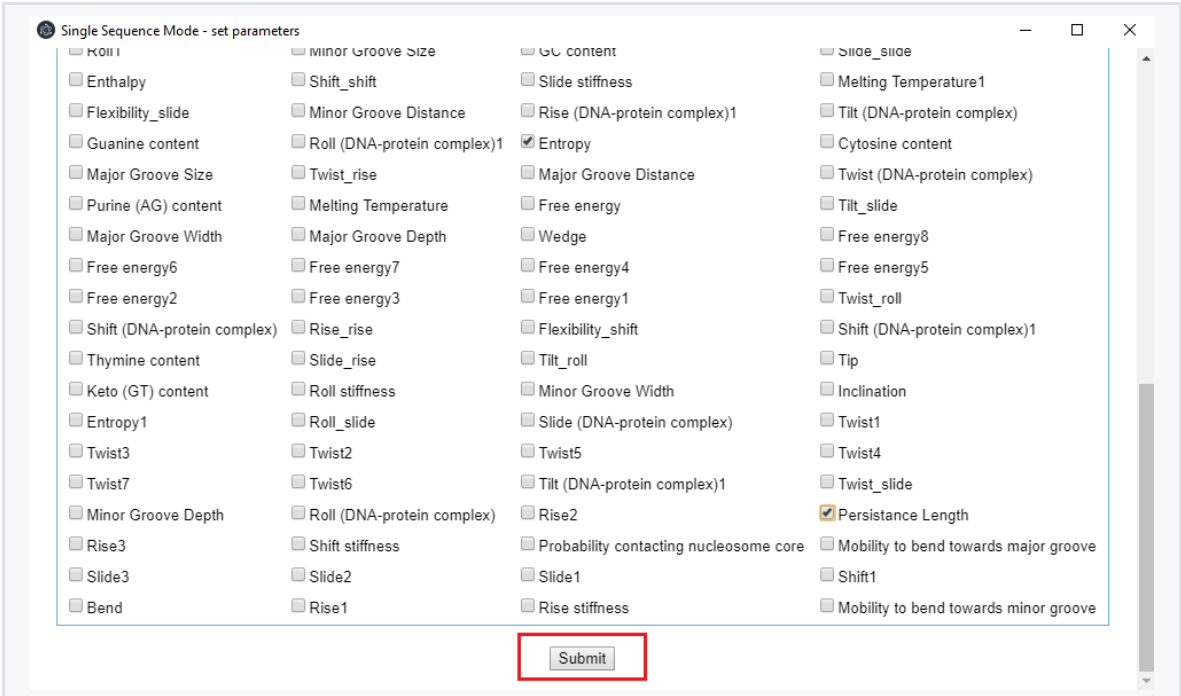

Figure 21 - Single sequence mode: step 3

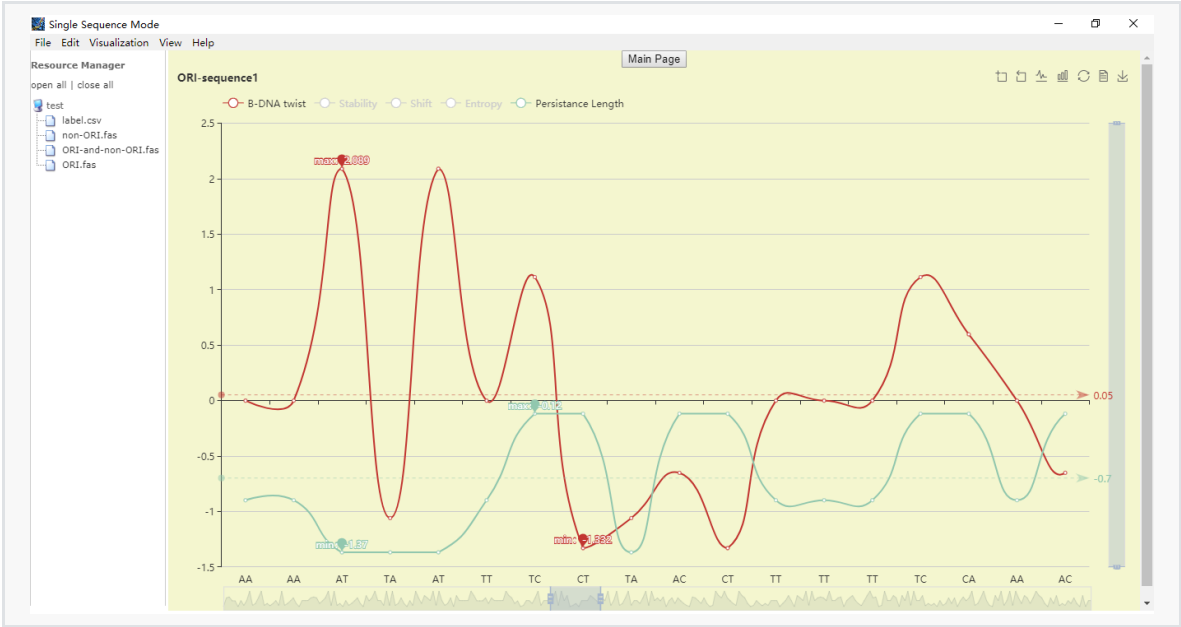

Figure 22 - Single sequence mode: step 4

There are many tools on this page. They can help you with further analysis.

- You can click on the name of the physicochemical property to show or hide its corresponding curve.

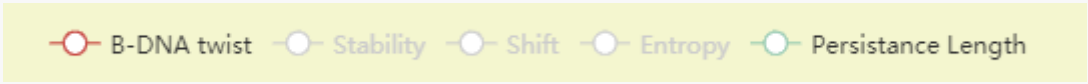

Figure 23 - Interactive tool: curve selection

- You can drag the sidebar of the x-axis and the y-axis or roll the mouse wheel for data scaling. You can also drag the mouse to view the local information on the chart. In these ways, you can focus on the details of data or outline the data as a whole. It's worth mentioning that you can observe sidebar on the x-axis to find the trend of data change.

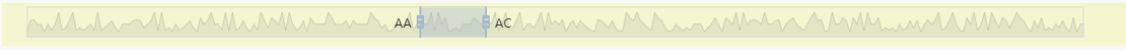

Figure 24 - Interactive tool: sidebars

- You can hover over any point in the chart to observe all data for that point.

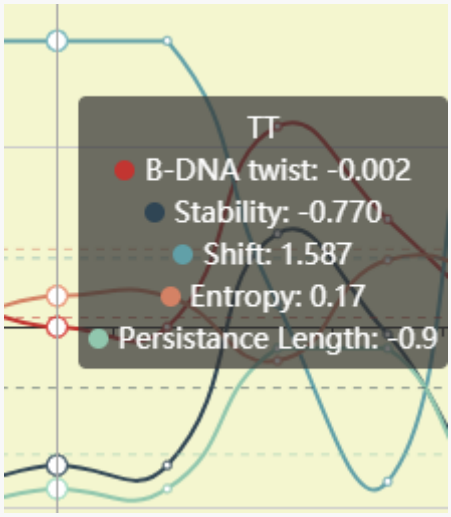

Figure 25 - Interactive tool: Data details

- There are seven buttons in the upper right corner of the page, they are “Zoom In Area”, ”Restore Zoom”, ”Line Chart”, ”Bar Chart”, “Restore”, “Data View” and “Save”. Click “Zoom In Area” to expand the area that you selected. Click “Restore Zoom” to undo expand. Click “Line Chart” to display the line chart. Click “Bar Chart” to display the bar chart. Click “Restore” to restore the chart to its original state. Click “Data View” to view or edit data for all curves. After editing the data, you can click "Update" to update the chart. Click "Save" to save the chart to your computer hard drives.

Figure 26 - Interactive tool: Corner buttons

2.3.2 Multiple sequences mode

First, you need to click "Multiple sequences mode" from "Visualization" in the menu bar, then complete all parameters and click “Submit”. Finally, the program will jump to the visual page. The visual page will display the curves according to the sequences and physicochemical properties that you select, where the dotted lines represent the average level of that physicochemical properties. There are three sub-modes in this mode, which are “default”, “truncation” and “clustalw2”. The functions of the toolbar are the same as that in Single sequence mode.

- If you select " truncation ", the program will truncate the sequences according to the shortest sequence of all the sequences you select, which means that the length of all sequences in this sub-mode is the same.
- If you select " clustalw2", the program will use clustalw2 for multiple sequence alignment and then visualize the alignment result.

Note:

- You can set the **maximum number** of selectable physicochemical properties on the page for setting parameters. The excessive number of physicochemical properties selected will result in too many curves on the page of visualization, which will affect the visual effect. When the number of sequences or the number of physicochemical properties selected is large, the rendering speed of curves will become slower and the delay of the operation will increase.

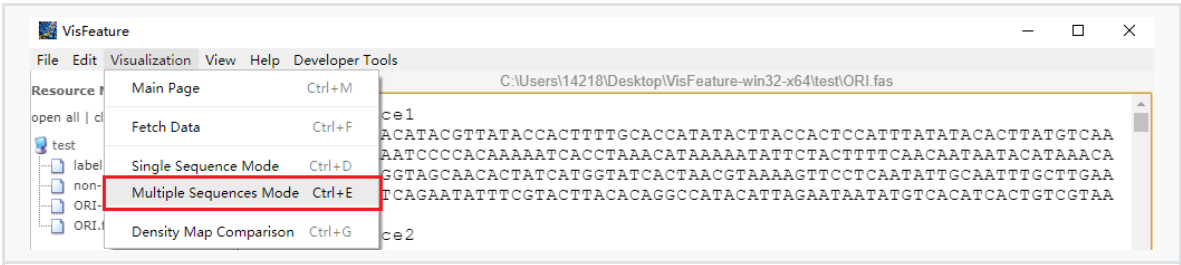

Figure 27 - Multiple sequences mode: step 1

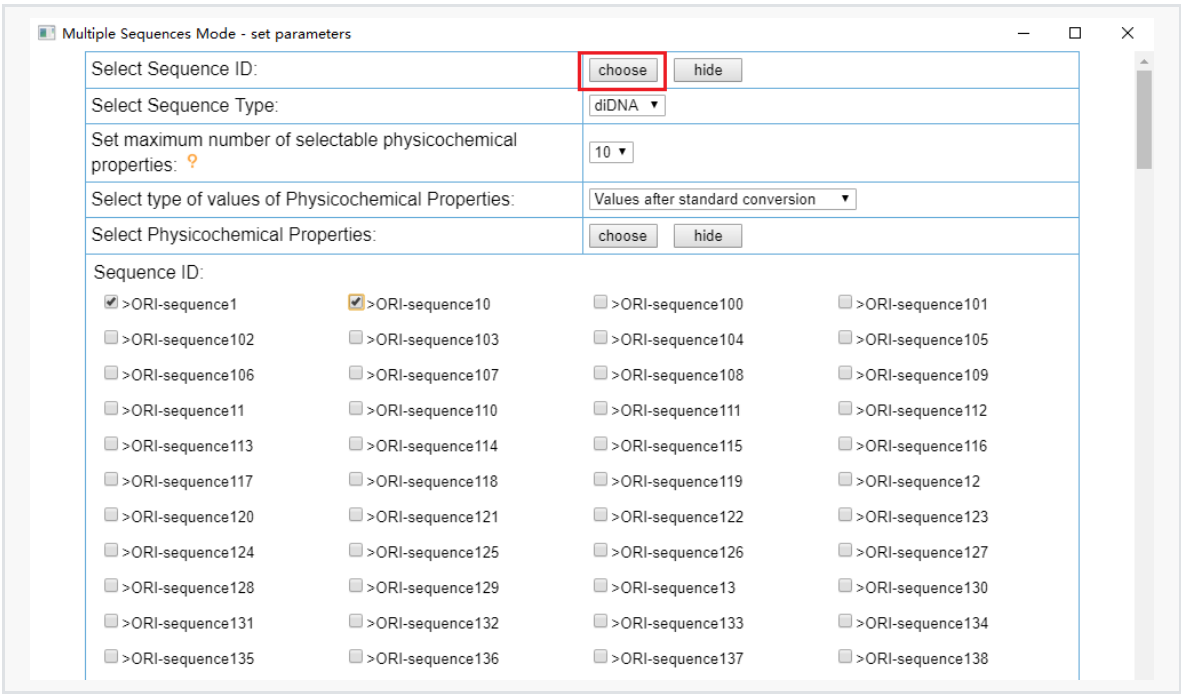

Figure 28 - Multiple sequences mode: step 2

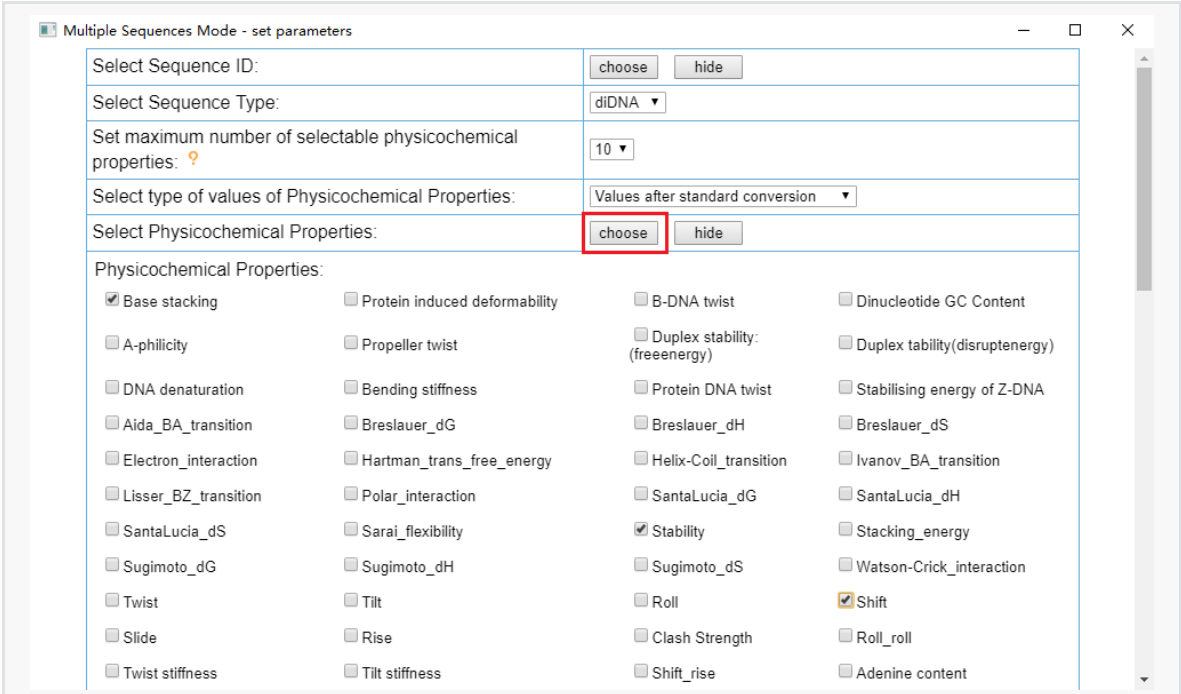

Figure 29 - Multiple sequences mode: step 3

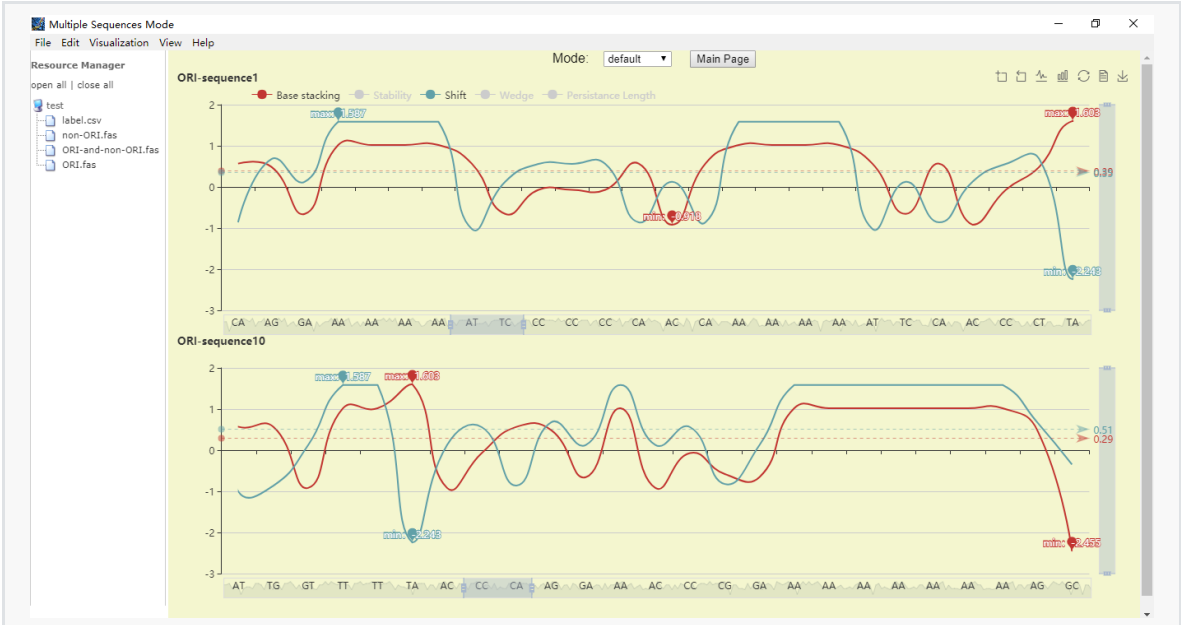

Figure 30 - Multiple sequences mode: step 4

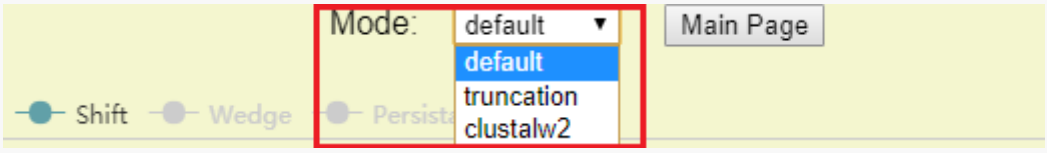

Figure 31 - Multiple sequences mode: step 5

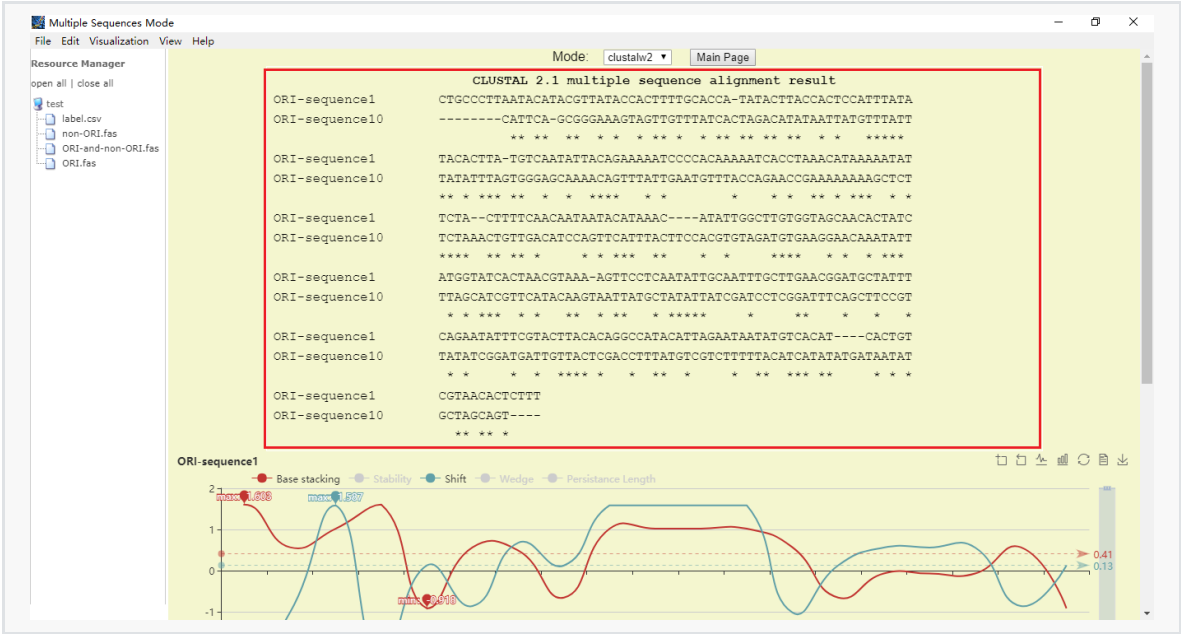

Figure 32 - Multiple sequences mode: step 6

2.3.3 Density map comparison

First, you need to click "Density map comparison" from "Visualization" in the menu bar, then complete all parameters and click “Submit”. Finally, the program will generate the vectors that can effectively reflect its key features according to sequences in your FASTA file. In this page, you can set the **maximum dimension** of the visual feature vector and upload a label file for assigning group labels to each sequence, then click on "Visualization" to visualize the feature vectors. Waiting for about 10 seconds, the program will jump to the visual page. This page will display the density maps of the feature vectors. You can save the chart to your computer hard

disk by clicking "save". There are two sub-modes in this mode, which are “single composition” and “multiple compositions”.

- If you select "single composition", the visual page will display density maps of the features on each dimension. For example, if the feature vector of every sequence has *n* dimensions, this mode includes *n* density maps.
- If you select "multiple compositions", the visual page will display density maps for each dimension in each group separately. The density maps of different dimensions in the same group are stacked together using transparent figures. This sub-mode will generate 2 charts, which are the charts that display by column and display by row.

The label file must follow this format: **Identifier,Group.** When the output format that you choose is csv(comma separated) and tsv(tab-separated), the identifier in the label file is the **first** element of each result. When the output format that you choose is svm(libSVM), the identifier in the file is the string after the “#” in each feature vector. Group is a string used to represent the category of a sequence. The label file in csv(comma separated) format is **recommended**.

- When output format that you choose is **csv**(comma separated), do not enter commas in the identifier of sequence in your FASTA file if you want to visualize the feature vectors. For example, If the computed result of your FASTA file is as follows:

```
sp|P99999, 5.714, 1.905, 2.857, 7.619, 2.857, 12.381, 2.857, 7.619, 17.143,
5.714, 3.810, 4.762, 3.810, 1.905, 1.905, 1.905, 6.667, 2.857, 0.952, 4.762,
CYC_HUMAN Cytochrome c OS=Homo sapiens OX=9606 GN=CYCS PE=1 SV=2
```

Your label file **MUST** have a line with this format:

```
sp|P99999,GroupA
```

- When output format that you choose is **tsv**(tab-separated), do not enter tab character in the identifier of sequence in your FASTA file if you want to visualize the feature vectors. For example, If the computed result of your FASTA file is as follows:

```
|gene_id|100040529|transcript_id|XR_875063     Gm2824 Mus musculus lncRNA
31.599    19.360       19.900    29.140    (null)
```

Your label file **MUST** follow this format:

```
|gene_id|100040529|transcript_id|XR_875063     Gm2824 Mus musculus lncRNA,GroupB
```

- When output format that you choose is **svm**(libSVM), do not enter space in the identifier of sequence in your FASTA file if you want to visualize the feature vectors. For example, If the computed result of your FASTA file is as follows:

```
0  1:6.564 2:1.544 3:3.475 4:5.405 5:5.019 6:6.564 7:2.703 8:5.019 9:5.792
10:10.811 11:1.544 12:3.861 13:8.494 14:5.019 15:4.247 16:6.178 17:6.564
18:6.564 19:1.158 20:3.475 # sp|Q8N2K1
```

Your label file **MUST** follow this format:

```
sp|Q8N2K1,GroupC
```

Note:

- It is optional to upload a file in FASTA format on the page of set parameters. If you do not upload a file, the program will take the contents of the input area as input. If you upload a file, the program will take this file as input. If both methods are used, the program will take the file that you upload as input. Please **upload** a FASTA format file on this page when your FASTA file is large. Because open a large file is slow, upload a large file on this page is very fast.
- Please make sure the sequence database for searching is properly configured before executing the "**PsePSSM**" sub-mode, otherwise, VisFeature cannot be executed. You should prepare three database files with the prefix "uniprot" in the *VisFeature-win32-x64\resources\app\UltraPse* directory, namely "uniprot.phr", "uniprot.pin" and "uniprot.psq".
- In some modes, you can set the maximum number of selectable physicochemical properties. The excessive number of physicochemical properties selected will increase calculation time.
- The maximum dimension of the visual feature vector in "single composition" and "multiple compositions" sub-mode is **150**, exceeded parts will be ignored. For example, if the dimension of each feature vector larger than 150, VisFeature will visualize the top 150 dimensional features.
- In the label file, groups with fewer than two data points will be **dropped**. This means that if the number of sequences in a group is less than two, the chart of this group will be **empty**.

Figure 33 - Density map comparison: step 1

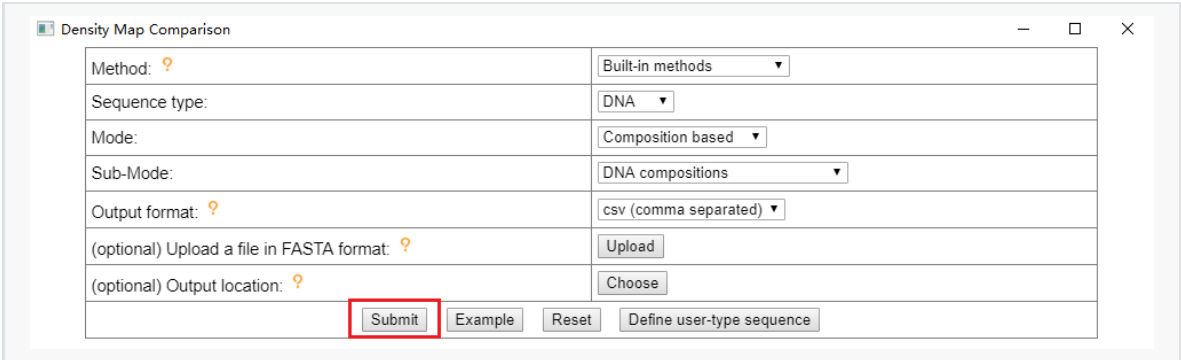

Figure 34 - Density map comparison: step 2

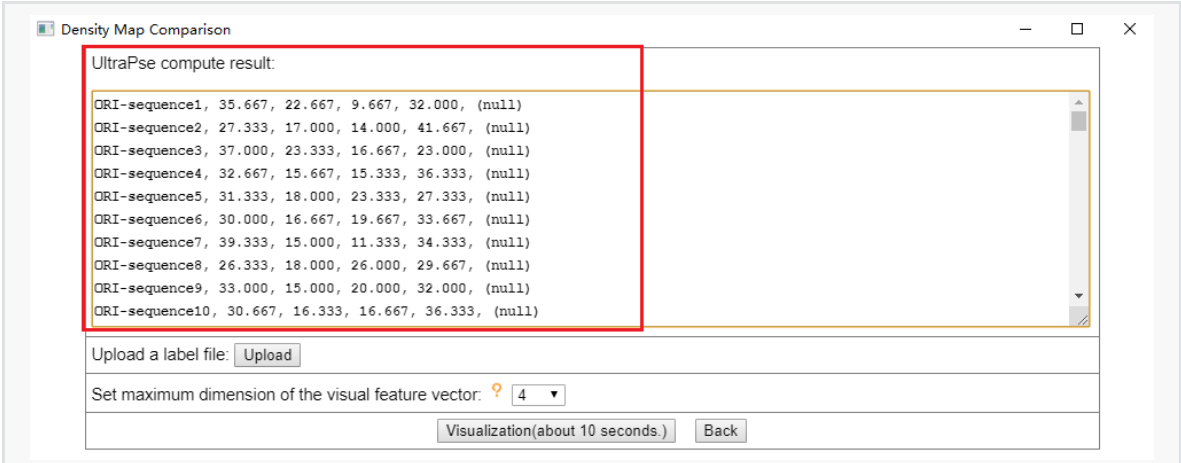

Figure 35 - Density map comparison: step 3

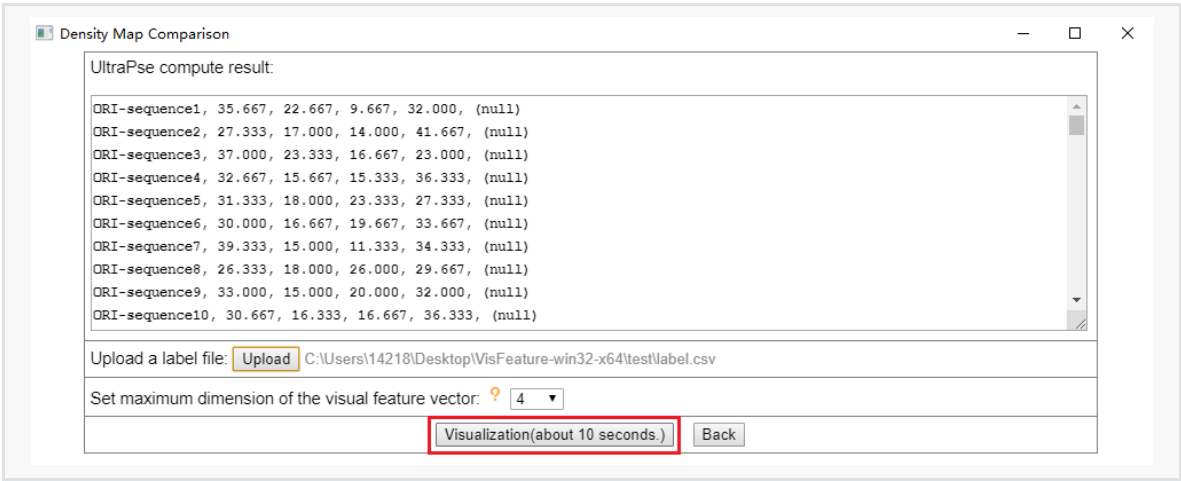

Figure 36 - Density map comparison: step 4

If you don't define groups of sequences, then the groups of these sequences are “Group:undefined”. The following is an example of “single composition” in which groups of all sequences are not defined. The visual page will display density maps of a dimension of all feature vectors in a chart.

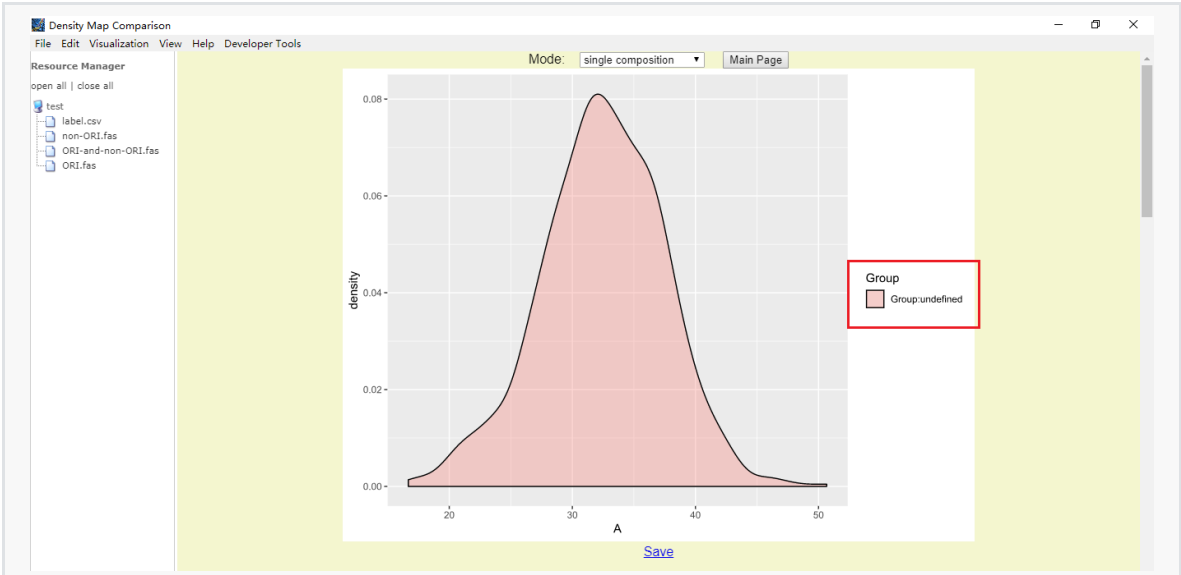

Figure 37 - Density map: undefined group

The following is an example of “multiple compositions” in which groups of all sequences are undefined. The visual page will display density maps of all dimensions of all feature vectors in a chart.

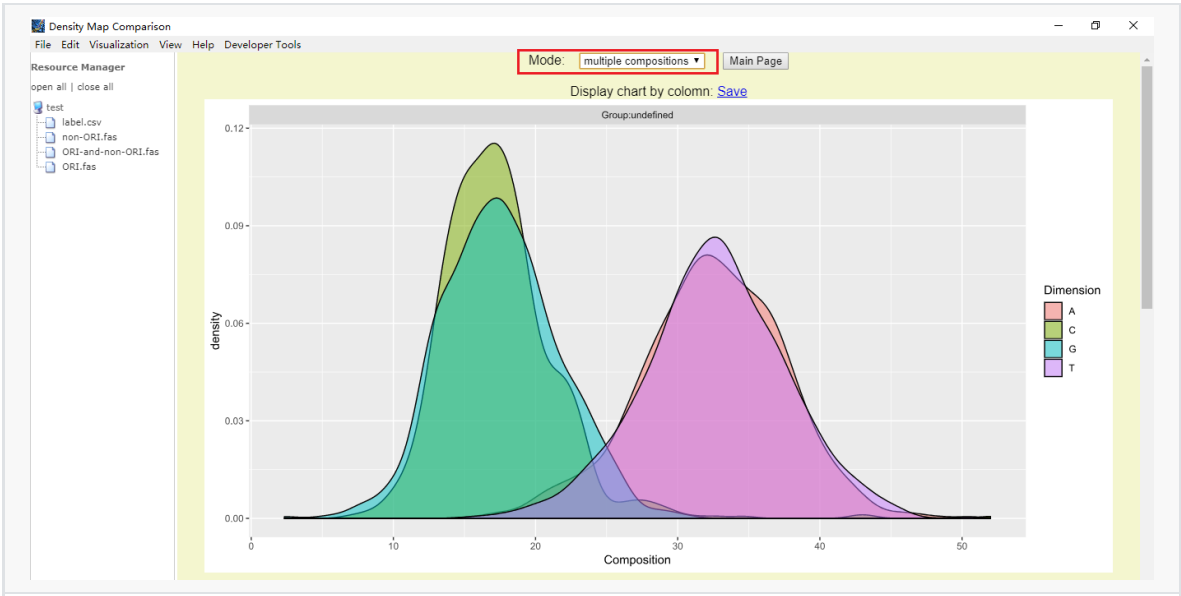

Figure 38 - Density map: multiple compositions

The following is an example of “single composition” and all sequences are divided into two groups. The visual page will display density maps of a dimension of all feature vectors from two groups in a chart.

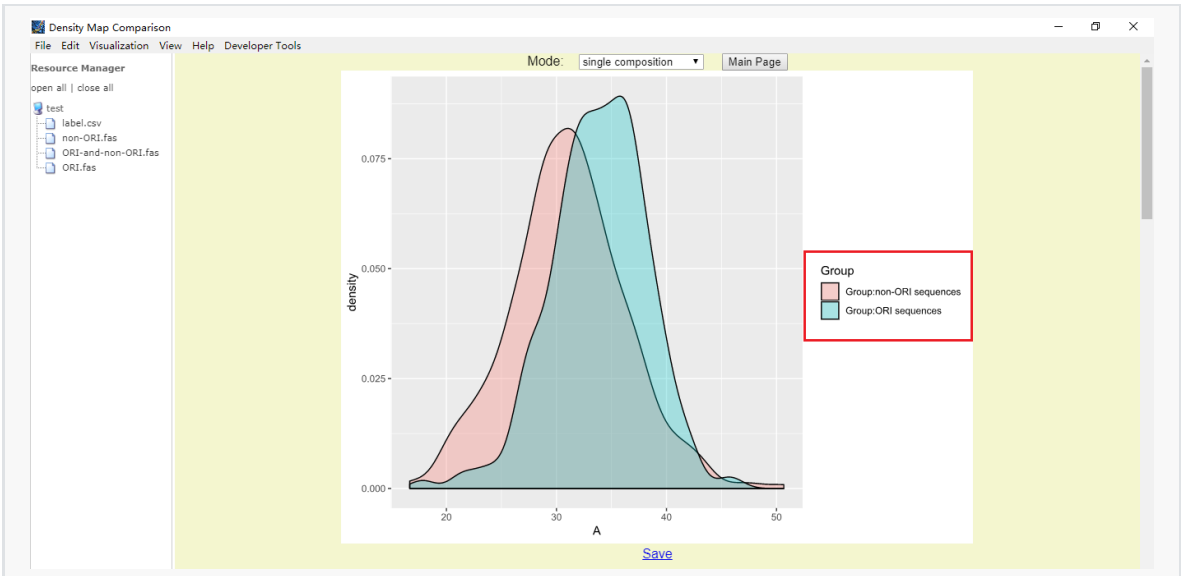

Figure 39 - Density map: single composition

The following is an example of “multiple compositions” and all sequences are divided into two groups. The visual page will display density maps of all dimensions of all feature vectors from two groups in a chart.

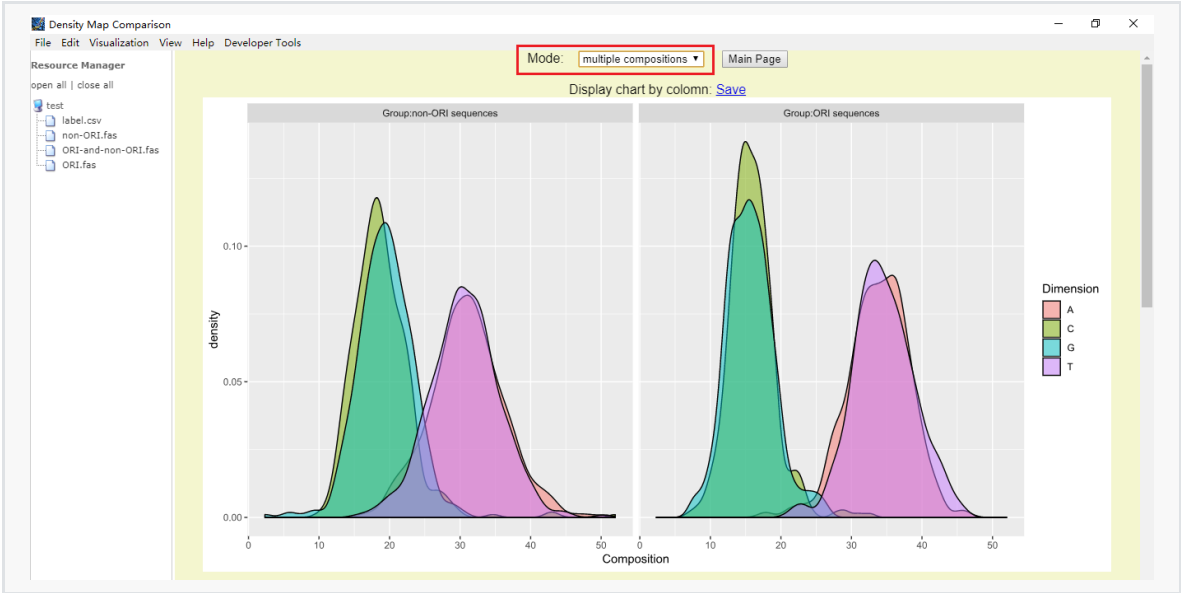

Figure 40 - Density map: group comparison

3 Other Information

3.1 Development

If you want to run these codes in the development environment, you should install **Node.js** first. Please go to https://nodejs.org/en/download/ download Node.js environment and install that on your machine.

After your Node.js environment is ready, find out the location of the source code of VisFeature that you unpack and enter this directory in command line program. Then type and execute the command: `npm start`. After a few seconds, VisFeature will start.

If you want to **package** application, you should install **electron** and **electron-packager** additionally by executing the command `npm install electron -g` and `npm install electron-packager -g` in command line program. Then, find out the location of the source code of VisFeature that you unpack and enter this directory. Finally, type and execute the command: `npm run windows` or `npm run linux` to get corresponding binary release.

3.2 Authors correspondence

If you have any questions or bug reports about VisFeature, please feel free to contact JunWang, Email: wj0708@tju.edu.cn.

3.3 Submitting your add-ons

If you want to contribute plugins for VisFeature, just join the project on GitHub. The address of VisFeature is https://github.com/wangjun1996/VisFeature.

3.4 Resources

The following resources may be useful when you use VisFeature.

- UltraPse

https://github.com/pufengdu/UltraPse

- How to fetch data by Uniprot API

https://www.uniprot.org/help/api

- How to fetch data by NCBI API

https://www.ncbi.nlm.nih.gov/home/develop/api/

- Clustalw2

https://www.ebi.ac.uk/Tools/msa/clustalw2/

- Electron

https://electronjs.org/

- Echarts

https://github.com/apache/incubator-echarts

- R

https://www.r-project.org/

References

- Li,W.-C. et al. (2015) iORI-PseKNC: A predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. Chemometrics and Intelligent Laboratory Systems, 141, 100–106.