VisFeature: a standalone program for features extraction and visualization of biological sequence

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. In this paper, we propose VisFeature, a stand-alone program for features extraction and visualization of DNA, RNA and protein sequences. It provides a function for downloading a number of biological sequences automatically and integrates 29 sequence representation modes for extracting the features of DNA, RNA and protein sequences rapidly. It can also convert sequences into sequence graphs and probability density maps based on their physicochemical properties or feature vectors. Through the feature visualization, the differences between different sequences or groups can be observed visually, which is convenient and valid for sequence analysis.

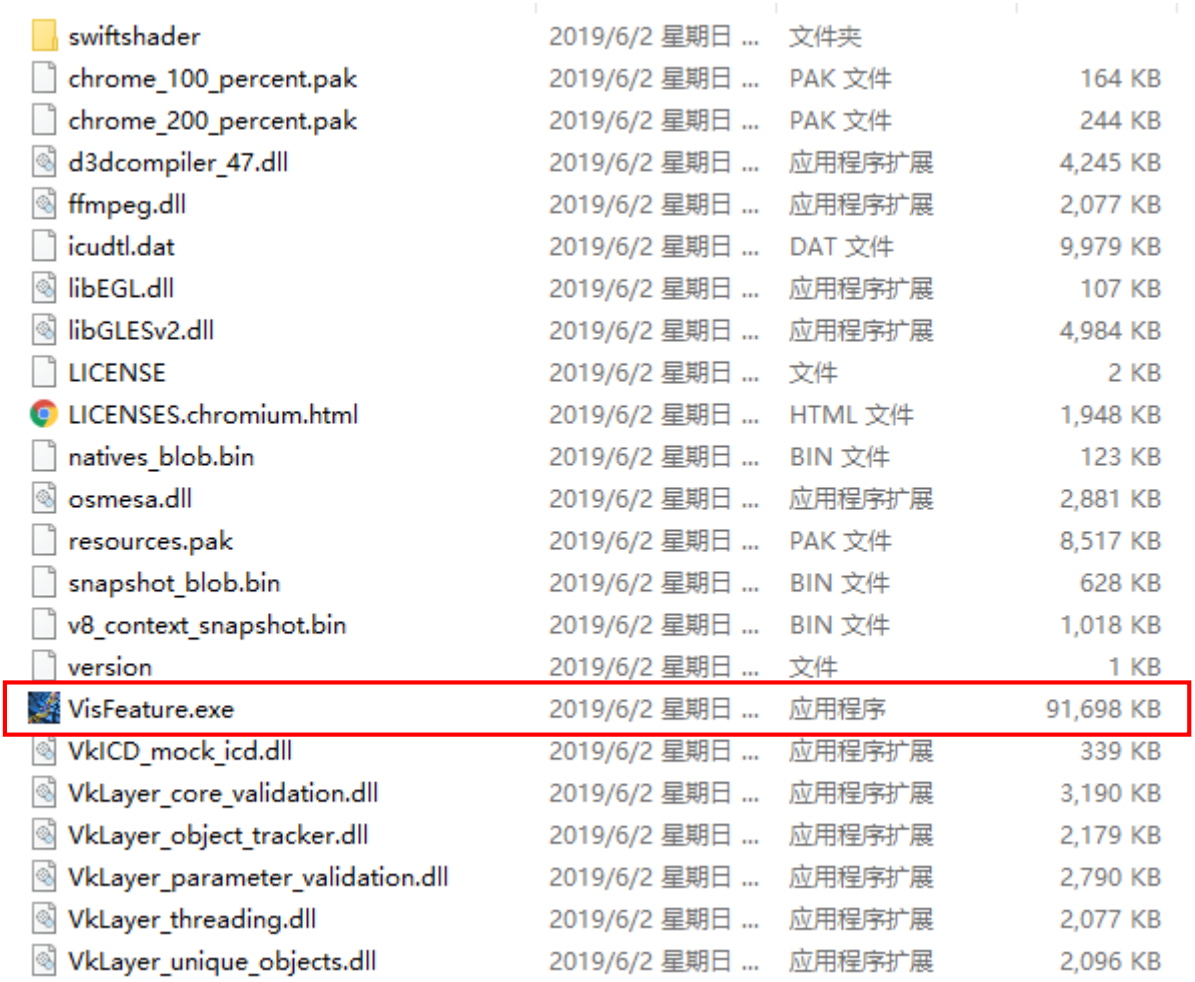
There are four main functions in VisFeature, which are “Fetch data”,“Visualization Mode 1 (Single sequence)”, “Visualization Mode 2 (Multiple sequences)” and “Compute And Visualization”.

You can use “Fetch data” to download a large number of DNA, RNA and protein sequences by identifier or query expression. You can use “Visualization Mode 1 (Single sequence)” to visualize a DNA, RNA or protein sequence by physicochemical properties. You can use “Visualization Mode 2 (Multiple sequences)” to visualize multiple DNA, RNA or protein sequences and do multiple sequence alignment then visualize the alignment result by physicochemical properties. You can use “Compute And Visualization” to calculate feature vectors from DNA, RNA and protein sequences. It allows the users to define their own representation mode, their own physicochemical properties, or even their own types of biological sequences. After the calculation, you can also use this function to visualize sequences by their feature vectors. Details will be given below.

2 Usage

Step 1. Installation

VisFeature has no dependency on Windows. You only need to download the VisFeature package and unpack it in your favorite location then open “VisFeature.exe”.

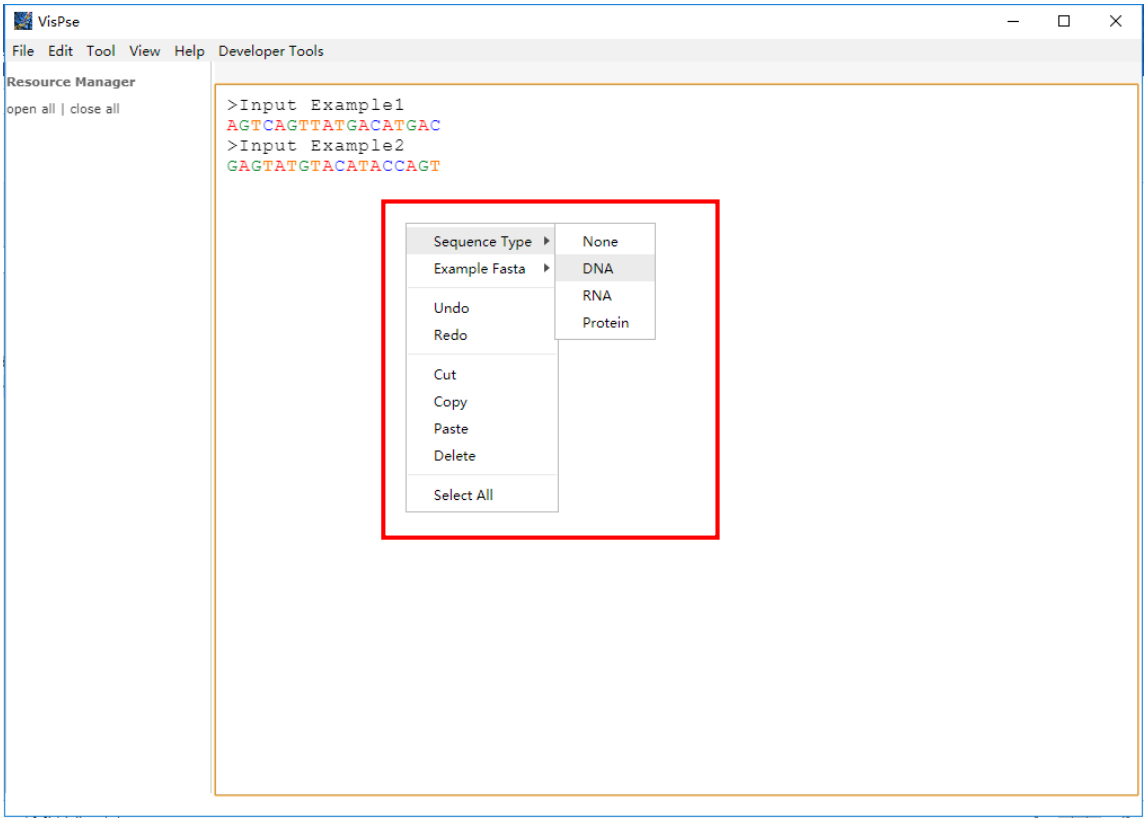


Step 2. Input

There are three ways to input into VisFeature, type, paste, and open file, respectively. Details will be given below.

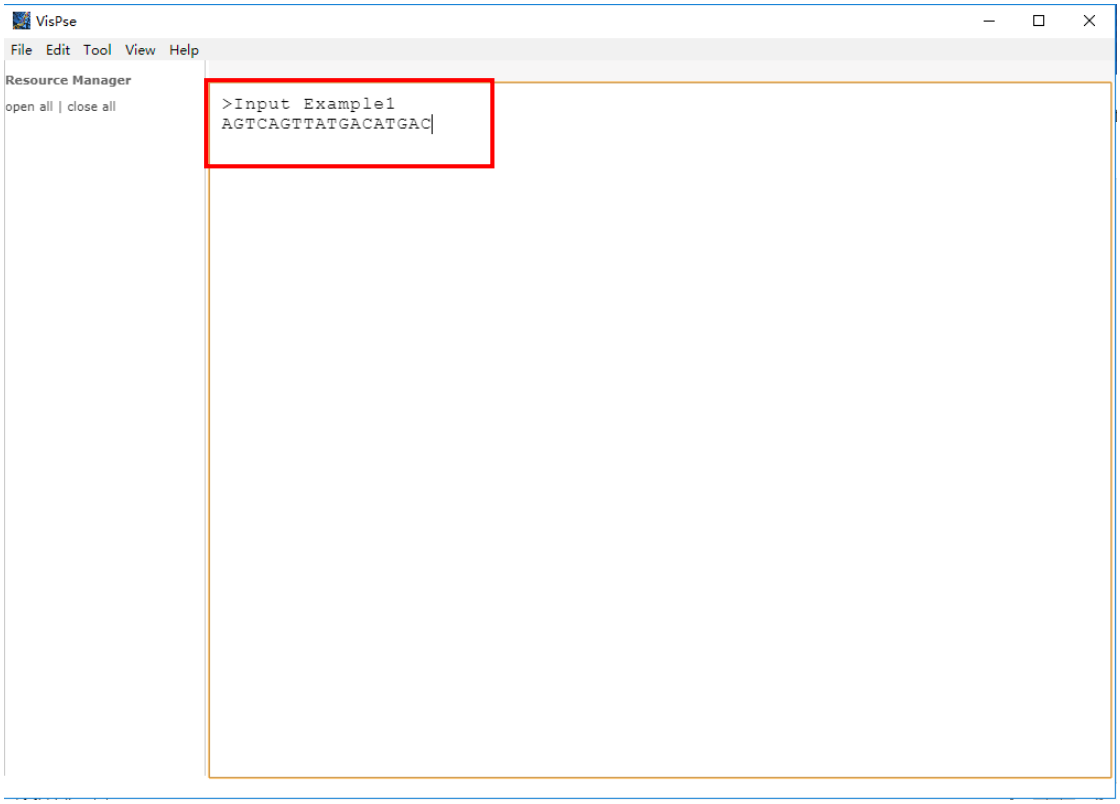
Please note:

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.



1) Type by keyboard

You can type content in the editing area of the main page by the keyboard.

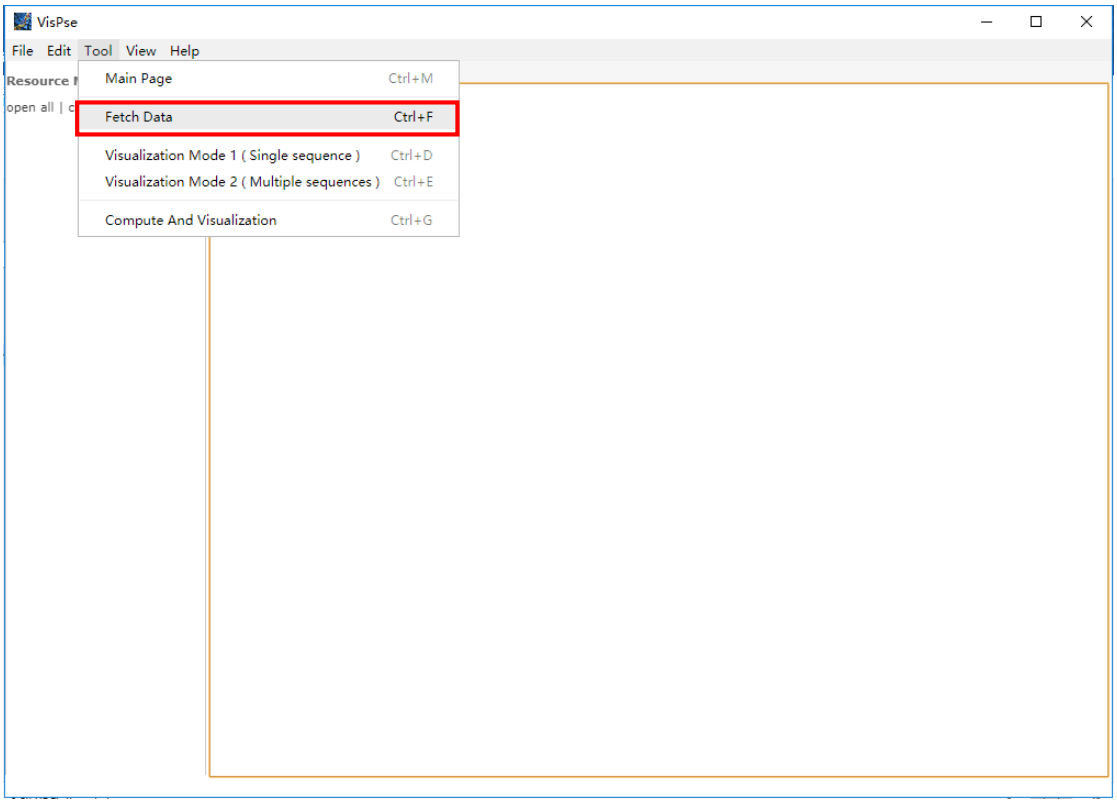


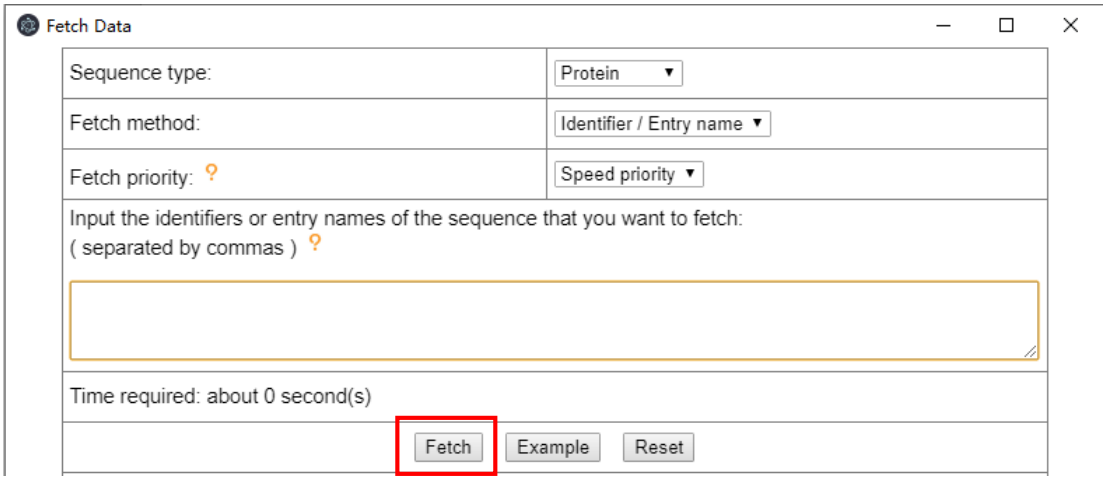
2) Copy and paste from "Fetch Result"

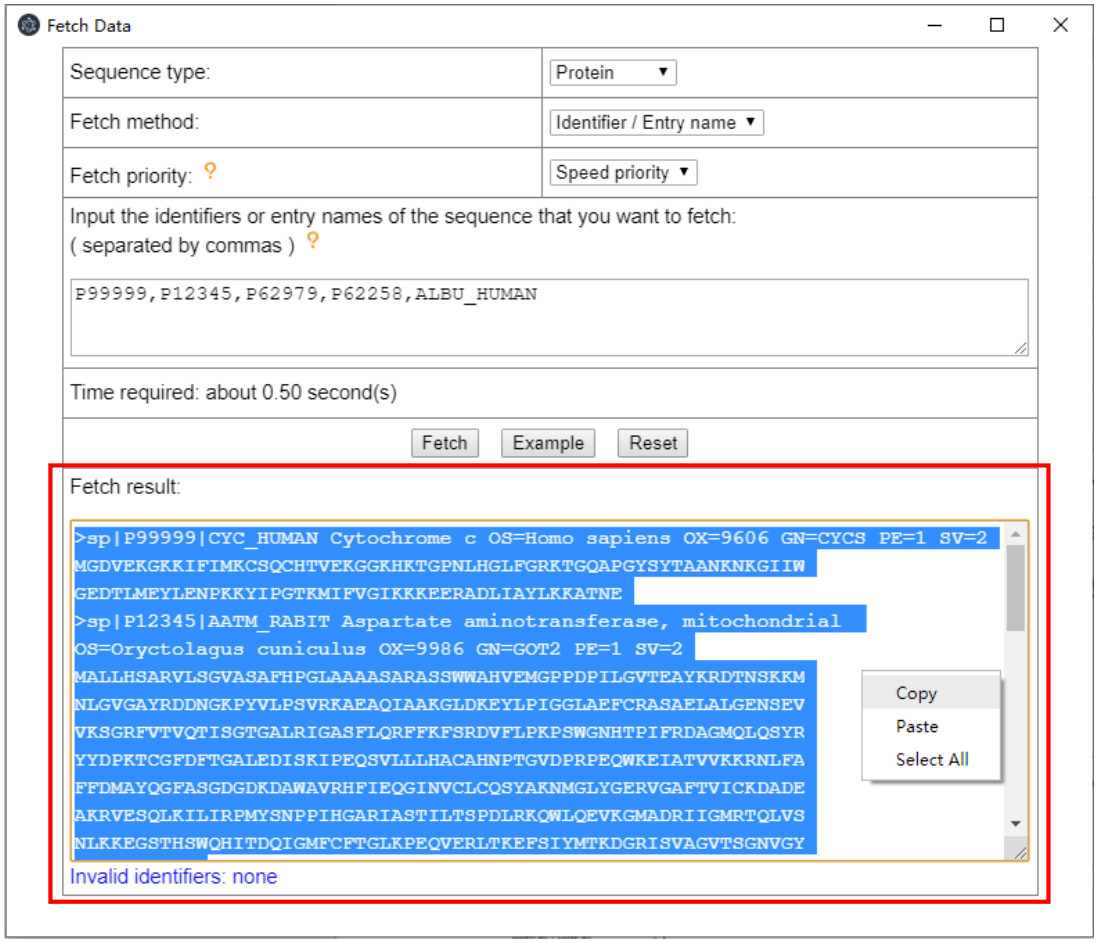
Click "Fetch Data" from "Tool" in the menu bar, then complete all parameters. Finally, input identifiers or query expression and click “Fetch” to fetch data of sequence. You can right-click on the fetch result area and select "Copy", then right-click on the input area of the main page and select "Paste". You can also use the corresponding shortcut keys.

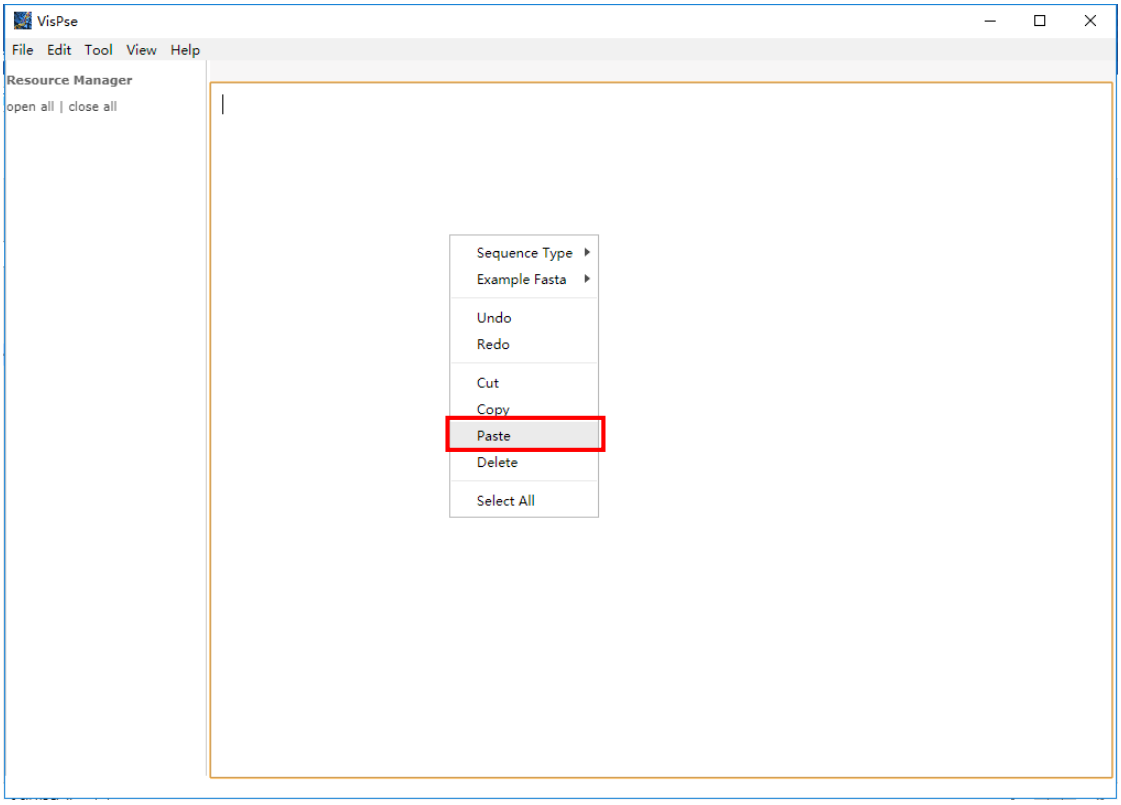
Please note:

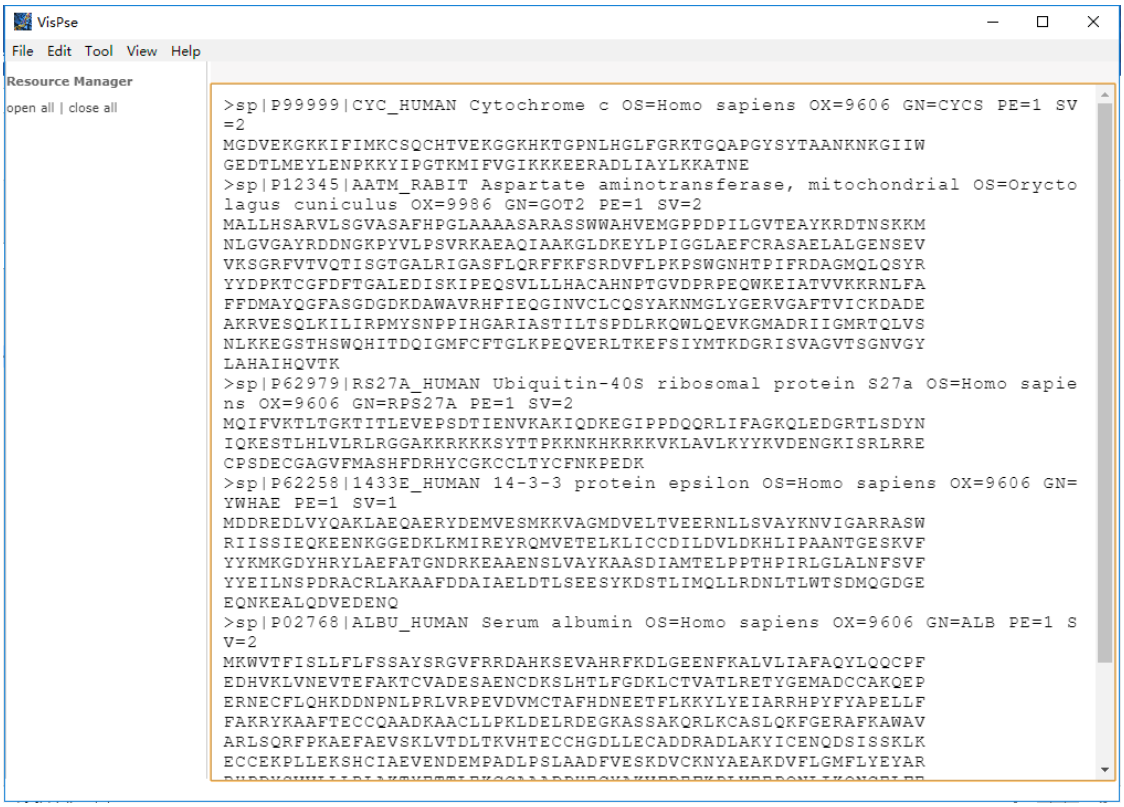
There are two options in "Fetch priority". They are "Speed priority" and "Order priority". If you select “Speed priority”, this will make the speed of fetch faster. There is **no guarantee** that the order of sequences of the fetch result will be the same as the order of input. If you select "Order priority",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


VisPse
File Edit Tool View Help
Main Page Ctrl+M
Fetch Data Ctrl+F
Visualization Mode 1 (Single sequence) Ctrl+D
Visualization Mode 2 (Multiple sequences) Ctrl+E
Compute And Visualization Ctrl+G


Fetch Data
Sequence type:
Protein
Fetch method:
Identifier / Entry name
Fetch priority:
Speed priority
Input the identifiers or entry names of the sequence that you want to fetch:
(separated by commas)
Time required: about 0 second(s)
Fetch
Example
Reset

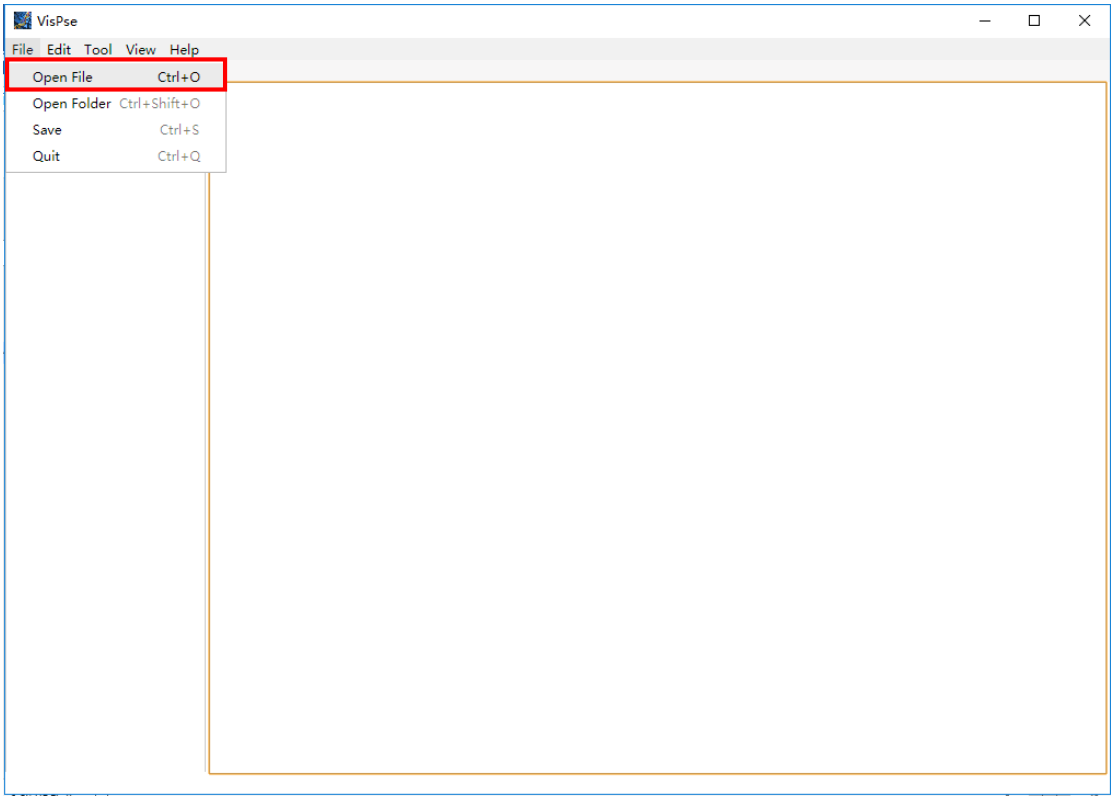

Fetch Data
Sequence type:
Protein
Fetch method:
Identifier / Entry name
Fetch priority:
Speed priority
Input the identifiers or entry names of the sequence that you want to fetch:
(separated by commas)
P99999,P12345,P62979,P62258,ALBU_HUMAN
Time required: about 0.50 second(s)
Fetch
Example
Reset
Fetch result:
>sp|P99999|CYC_HUMAN Cytochrome c OS=Homo sapiens OX=9606 GN=CYCS PE=1 SV=2
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIW
GEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
>sp|P12345|AATM_RABIT Aspartate aminotransferase, mitochondrial
OS=Oryctolagus cuniculus OX=9986 GN=GOT2 PE=1 SV=2
MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKM
NLGVGAYRDDNGKPYVLPSVRKAEAQIAAKGLDKEYLPIGGLAEFCRASAELALGENSEV
VKSGRFVTVQTISGTGALRIGASFLQRFFKFSRDVFLPKPSWGNHTPIFRDAGMQLQSYR
YYDPKTCGFDFTGALEDISKIPEQSVLLLHACAHNPTGVDPRPEQWKEIATVVKKRNLFA
FFDMAYQGFASGDGDKDAWAVRHFIEQGINVCLCQSYAKNMGLYGERVGAFTVICKDADE
AKRVESQLKILIRPMYSNPPIHGARIASTILTSPDLRKQWLQEVKGMADRIIGMRTQLVS
NLKKEGSTHSWQHITDQIGMFCFTGLKPEQVERLTKEFSIYMTKDGRISVAGVTSGNVGY
Copy
Paste
Select All
Invalid identifiers: none

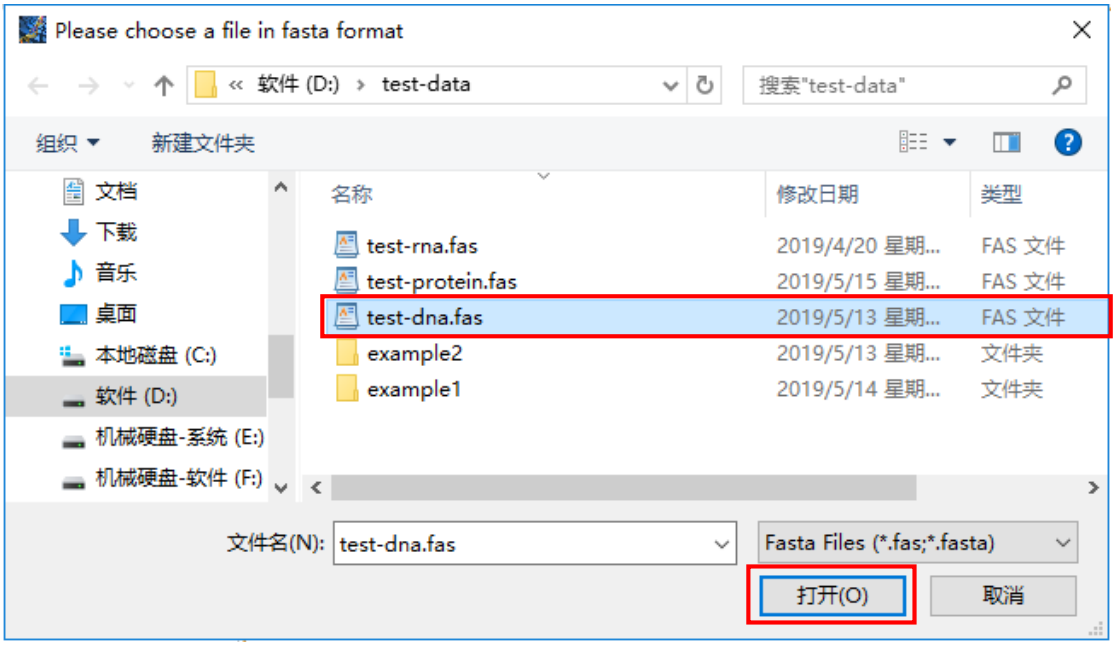

VisPse
File Edit Tool View Help
Resource Manager
open all | close all
Sequence Type
Example Fasta
Undo
Redo
Cut
Copy
Paste
Delete
Select All

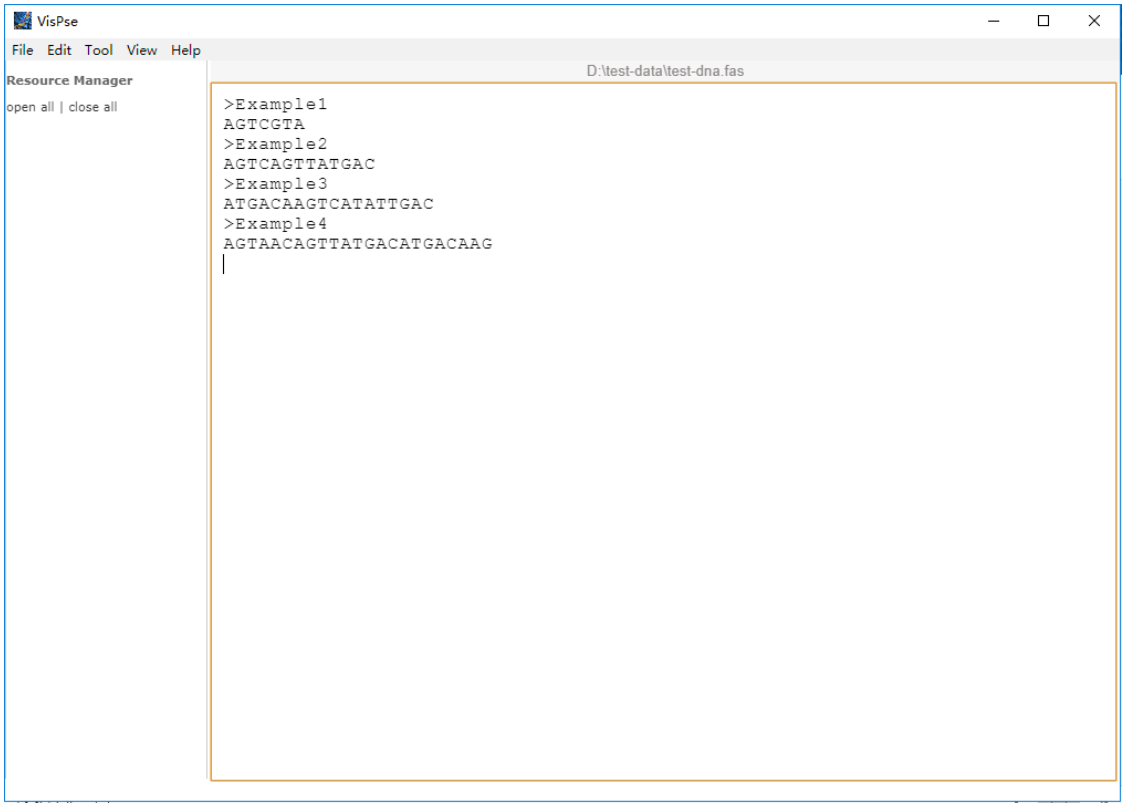


3) Open a file

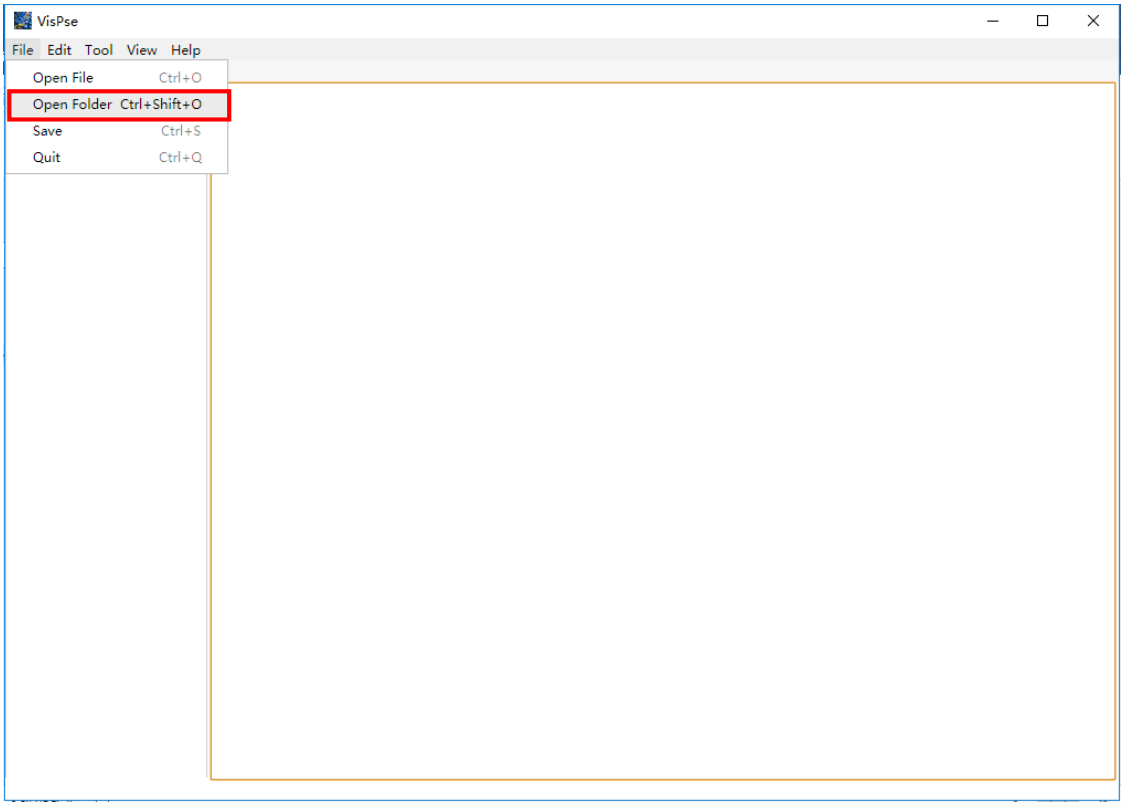
- Method 1. Open a file by clicking "Open File" from "File" in the menu bar.

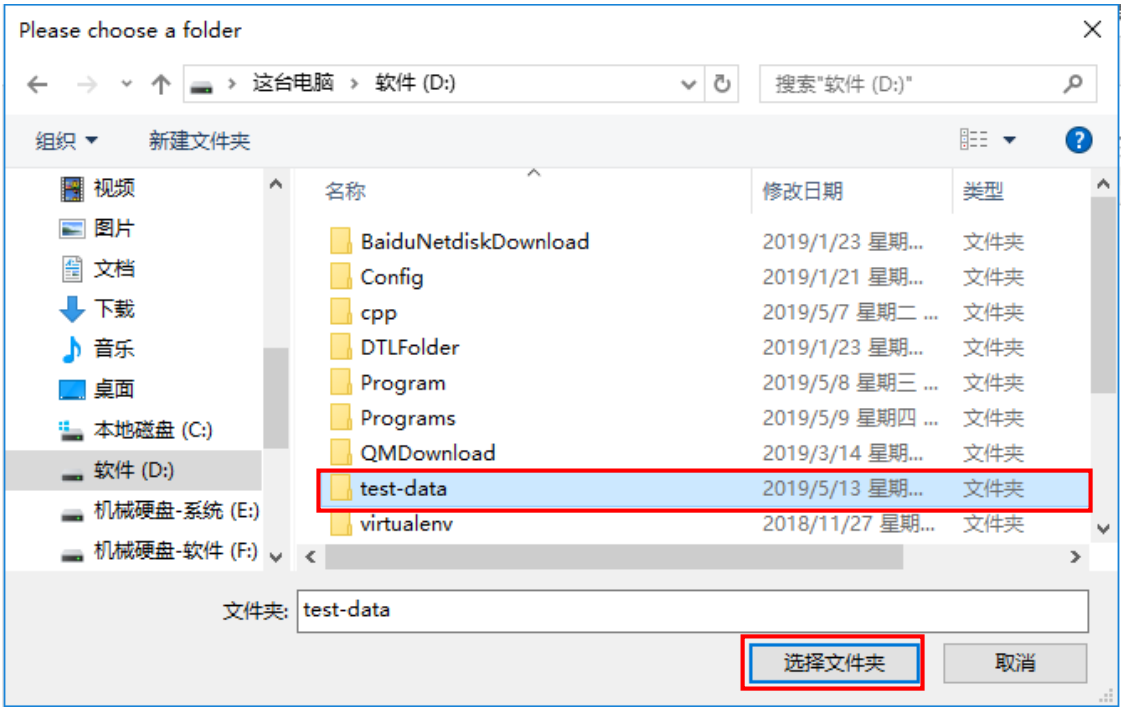


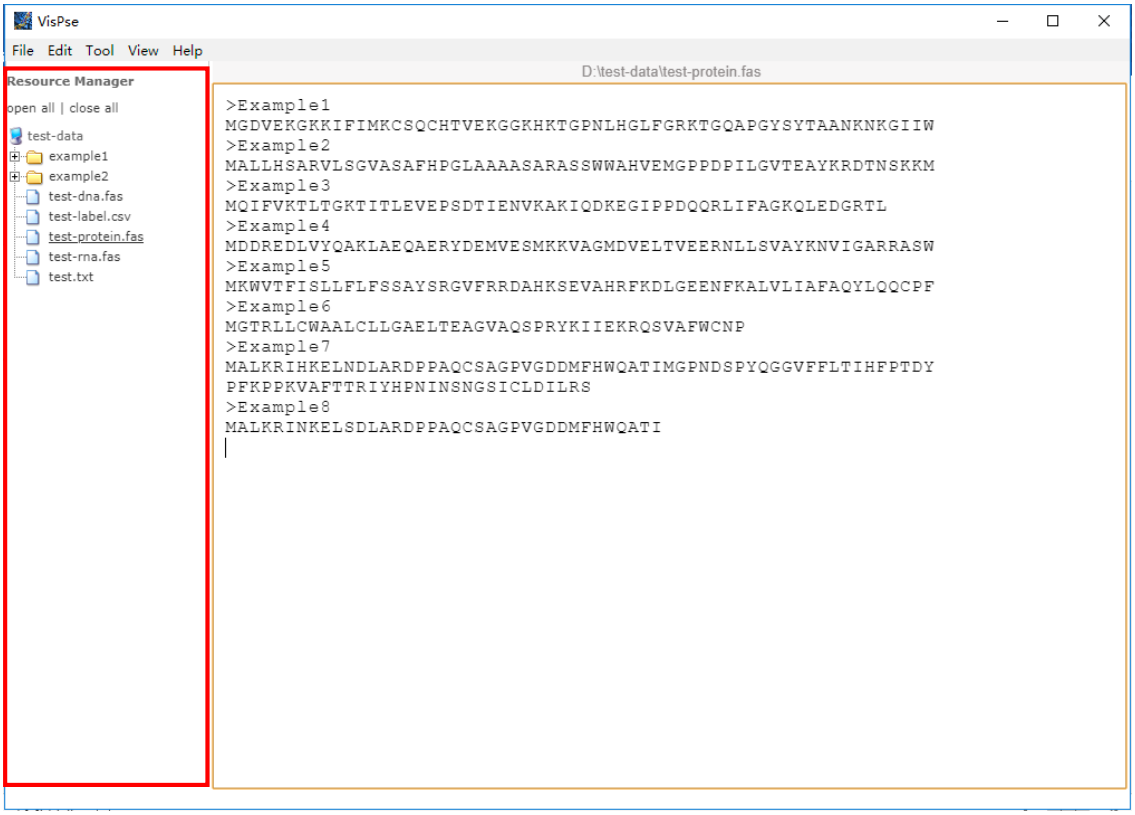

Please choose a file in fasta format
« 软件 (D:) › test-data
搜索"test-data"
组织
新建文件夹
文档
下载
音乐
桌面
本地磁盘 (C:)
软件 (D:)
机械硬盘-系统 (E:)
机械硬盘-软件 (F:)
名称
修改日期
类型
test-rna.fas
2019/4/20 星期...
FAS 文件
test-protein.fas
2019/5/15 星期...
FAS 文件
test-dna.fas
2019/5/13 星期...
FAS 文件
example2
2019/5/13 星期...
文件夹
example1
2019/5/14 星期...
文件夹
文件名(N): test-dna.fas
Fasta Files (*.fas;*.fasta)
打开(O)
取消


VisPse
File Edit Tool View Help
Resource Manager
open all | close all
D:\test-data\test-dna.fas
>Example1
AGTCGTA
>Example2
AGTCAGTTATGAC
>Example3
ATGACAAGTCATATTGAC
>Example4
AGTAACAGTTATGACATGACAAG

- 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.




Please choose a folder
这台电脑 › 软件 (D:)
搜索"软件 (D:)"
组织
新建文件夹
视频
图片
文档
下载
音乐
桌面
本地磁盘 (C:)
软件 (D:)
机械硬盘-系统 (E:)
机械硬盘-软件 (F:)
名称
修改日期
类型
BaiduNetdiskDownload 2019/1/23 星期... 文件夹
Config 2019/1/21 星期... 文件夹
cpp 2019/5/7 星期二 ... 文件夹
DTLFolder 2019/1/23 星期... 文件夹
Program 2019/5/8 星期三 ... 文件夹
Programs 2019/5/9 星期四 ... 文件夹
QMDownload 2019/3/14 星期... 文件夹
test-data 2019/5/13 星期... 文件夹
virtualenv 2018/11/27 星期... 文件夹
文件夹: test-data
选择文件夹
取消

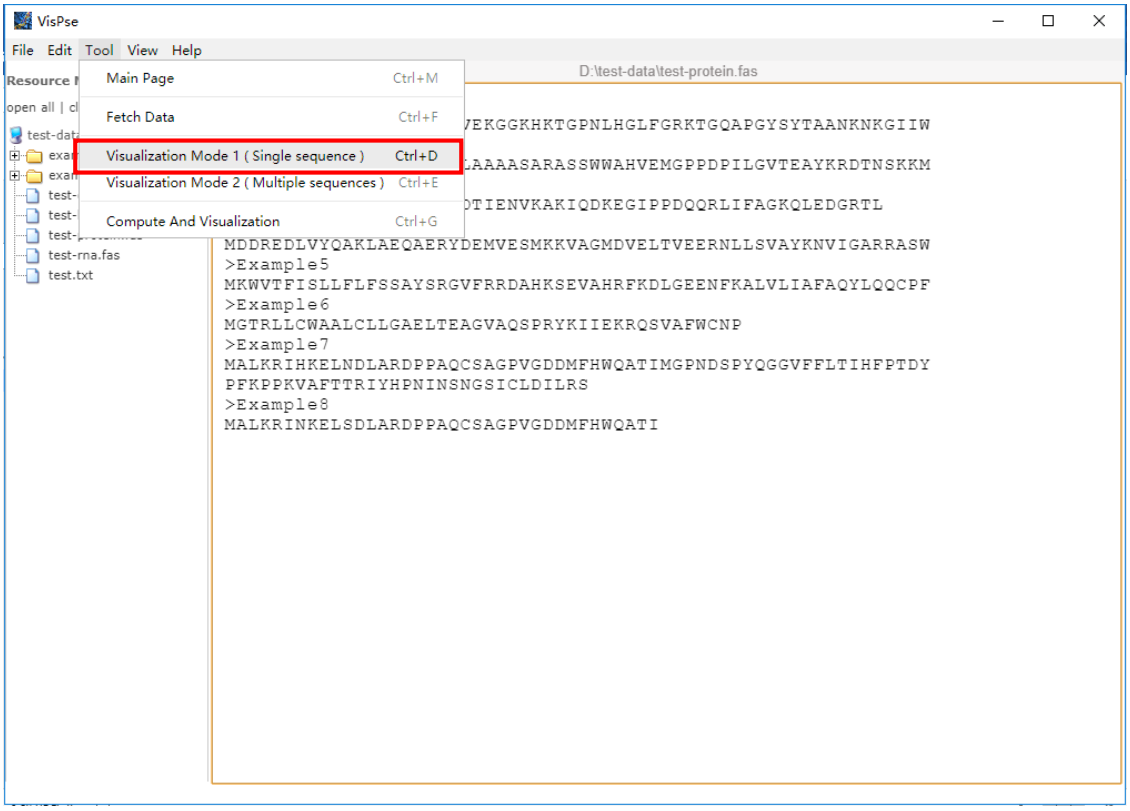

VisPse
File Edit Tool View Help
Resource Manager
open all | close all
test-data
example1
example2
test-dna.fas
test-label.csv
test-protein.fas
test-rna.fas
test.txt
D:\test-data\test-protein.fas
>Example1
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIW
>Example2
MALLHSARVLSGVASAFHPGLAAAASARASSWWAHVEMGPPDPILGVTEAYKRDTNSKKM
>Example3
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTL
>Example4
MDDREDLVYQAKLAEQAERYDEMVESMKKVAGMDVELTVEERNLLSVAYKNVIGARRASW
>Example5
MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPF
>Example6
MGTRLLCWAALCLLGAELTEAGVAQSPRYKIIEKRQSVAFWCNP
>Example7
MALKRIHKELNDLARDPPAQCSAGPVGDDMFHWQATIMGPNDSPYQGGVFFLTIHFPTDY
PFKPPKVAFTTRIYHPNINSNGSICLDILRS
>Example8
MALKRINKELSDLARDPPAQCSAGPVGDDMFHWQATI

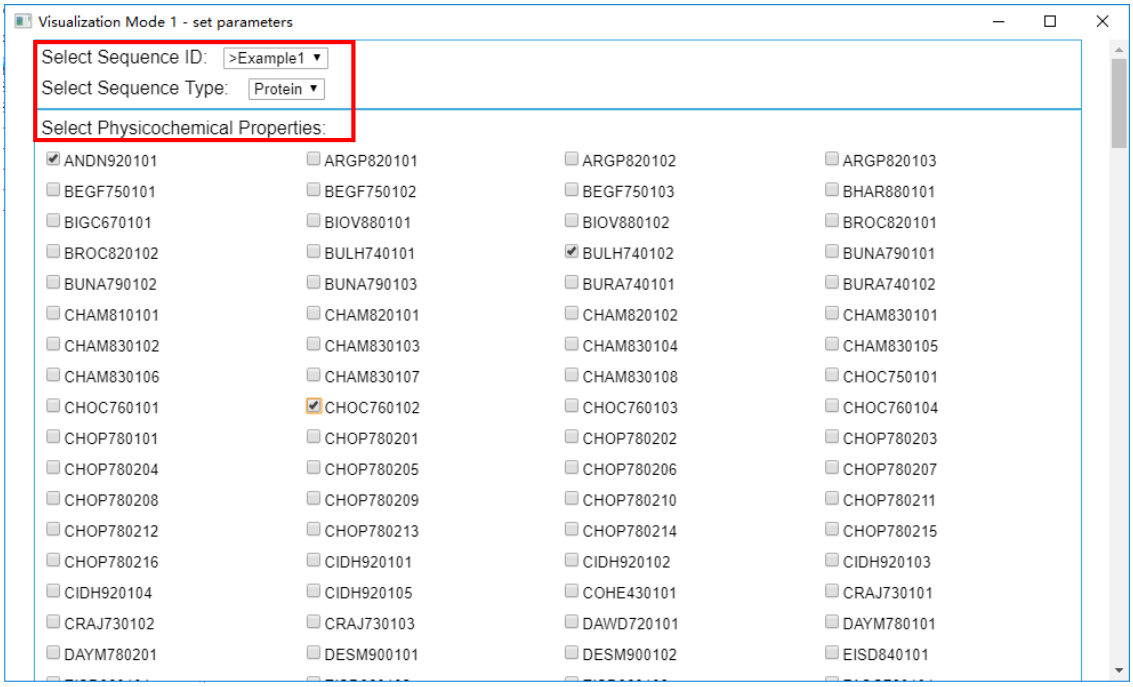
Step 3. Visualization

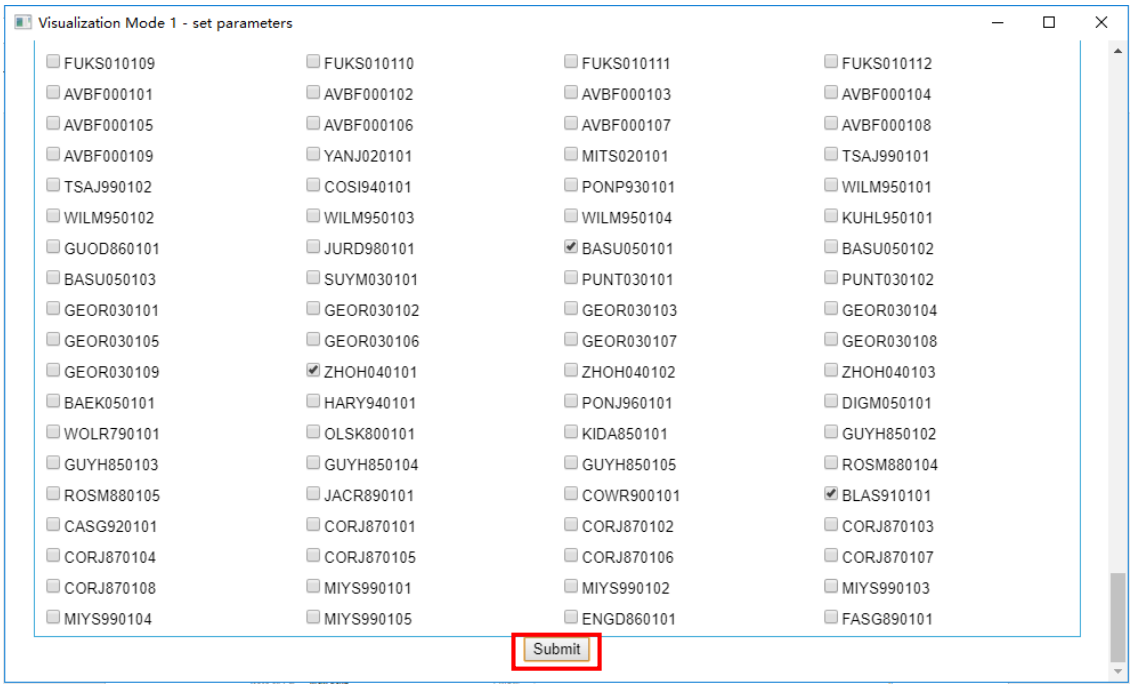
There are three visual modes in VisFeature, which are “Visualization Mode 1 (Single sequence)”, “Visualization Mode 2 (Multiple sequences)” and “Compute And Visualization”. In order to meet different needs, you can use "Zoom In" and "Zoom Out" for scaling from "View" in the menu bar. Of course, you can use them on other pages. Details will be given below.

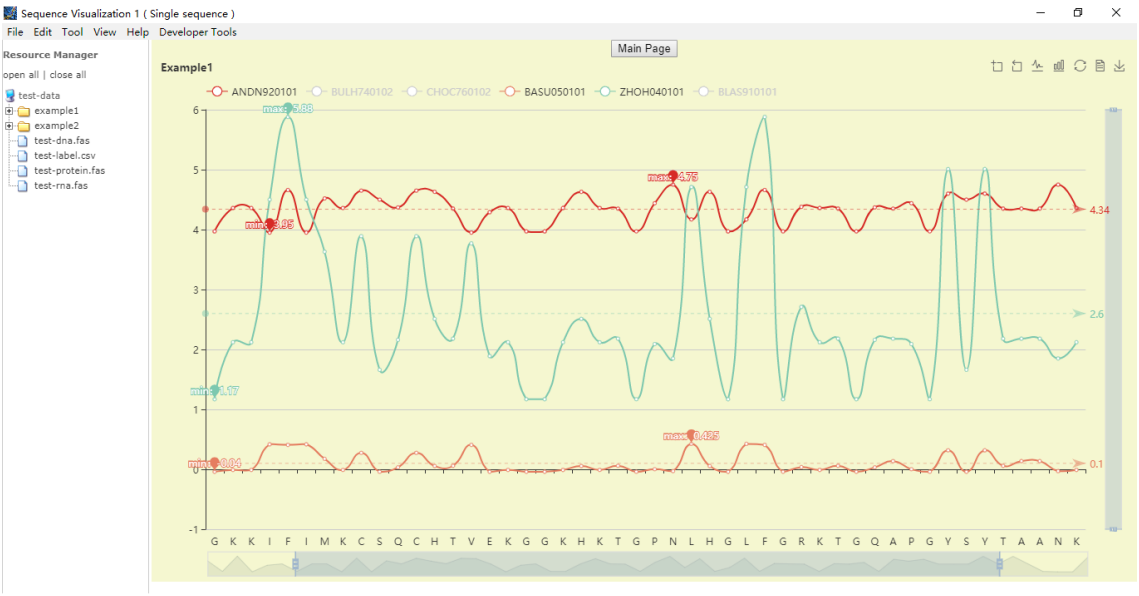
1) Visualization Mode 1 (Single sequence)

First, you need to click "Visualization Mode 1 (Single sequence) " from "Tool" in the menu bar, then complete all parameters and click “Submit”. Finally, 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.



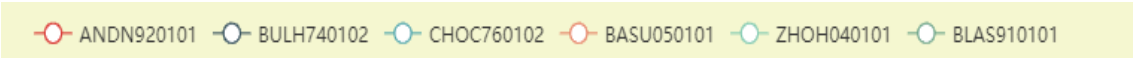

Visualization Mode 1 - set parameters
Select Sequence ID: >Example1
Select Sequence Type: Protein
Select Physicochemical Properties:
ANDN920101 ARGP820101 ARGP820102 ARGP820103
BEGF750101 BEGF750102 BEGF750103 BHAR880101
BIGC670101 BIOV880101 BIOV880102 BROC820101
BROC820102 BULH740101 BULH740102 BUNA790101
BUNA790102 BUNA790103 BURA740101 BURA740102
CHAM810101 CHAM820101 CHAM820102 CHAM830101
CHAM830102 CHAM830103 CHAM830104 CHAM830105
CHAM830106 CHAM830107 CHAM830108 CHOC750101
CHOC760101 CHOC760102 CHOC760103 CHOC760104
CHOP780101 CHOP780201 CHOP780202 CHOP780203
CHOP780204 CHOP780205 CHOP780206 CHOP780207
CHOP780208 CHOP780209 CHOP780210 CHOP780211
CHOP780212 CHOP780213 CHOP780214 CHOP780215
CHOP780216 CIDH920101 CIDH920102 CIDH920103
CIDH920104 CIDH920105 COHE430101 CRAJ730101
CRAJ730102 CRAJ730103 DAWD720101 DAYM780101
DAYM780201 DESM900101 DESM900102 EISD840101


Visualization Mode 1 - set parameters
FUKS010109 FUKS010110 FUKS010111 FUKS010112
AVBF000101 AVBF000102 AVBF000103 AVBF000104
AVBF000105 AVBF000106 AVBF000107 AVBF000108
AVBF000109 YANJ020101 MITS020101 TSAJ990101
TSAJ990102 COSI940101 PONP930101 WILM950101
WILM950102 WILM950103 WILM950104 KUHL950101
GUOD860101 JURD980101 BASU050101 BASU050102
BASU050103 SUYM030101 PUNT030101 PUNT030102
GEOR030101 GEOR030102 GEOR030103 GEOR030104
GEOR030105 GEOR030106 GEOR030107 GEOR030108
GEOR030109 ZHOH040101 ZHOH040102 ZHOH040103
BAEK050101 HARY940101 PONJ960101 DIGM050101
WOLR790101 OLSK800101 KIDA850101 GUYH850102
GUYH850103 GUYH850104 GUYH850105 ROSM880104
ROSM880105 JACR890101 COWR900101 BLAS910101
CASG920101 CORJ870101 CORJ870102 CORJ870103
CORJ870104 CORJ870105 CORJ870106 CORJ870107
CORJ870108 MIYS990101 MIYS990102 MIYS990103
MIYS990104 MIYS990105 ENGD860101 FASG890101
Submit



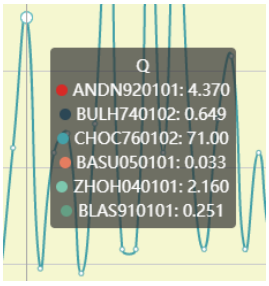
There are many tools on this page. They can help you with further analysis.

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



- You can slide the bar tool of the x-axis and the y-axis or move 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 the bar tool on the x-axis to find the trend of data change.

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



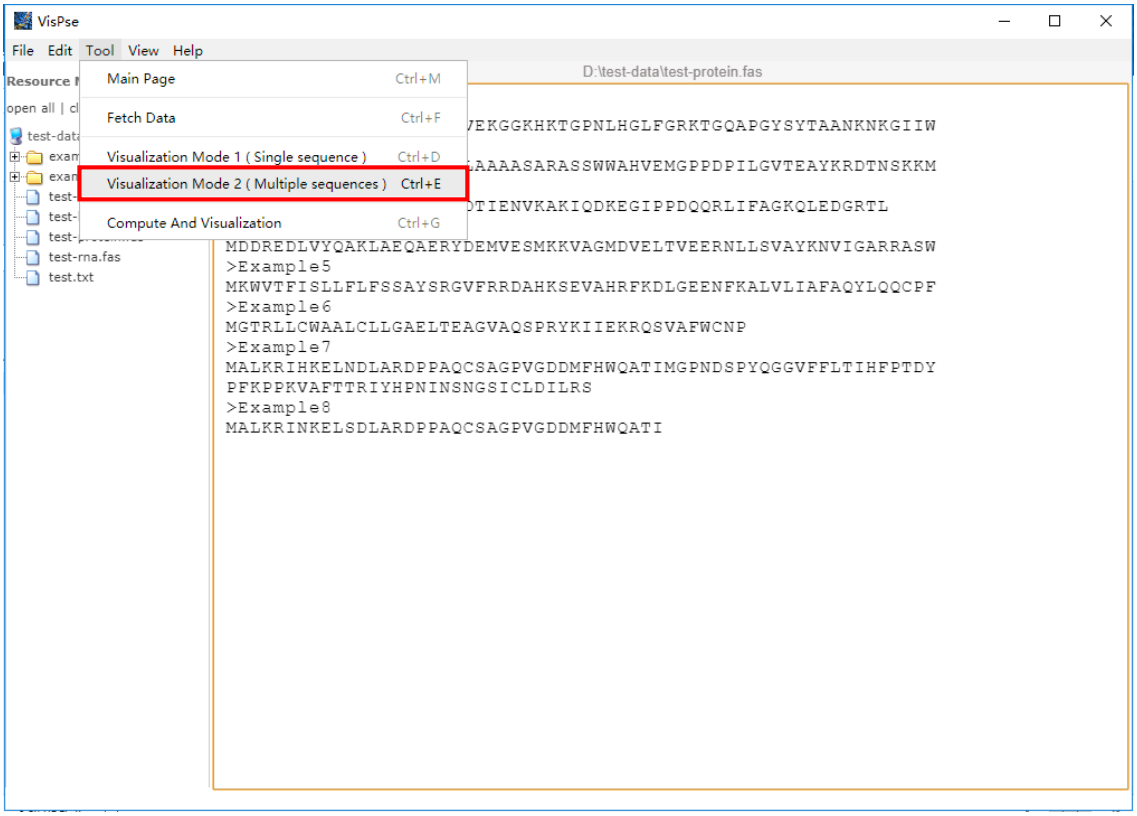
- There are seven buttons in the upper right corner of the page, they are “Zoom In Area”, ”Restore Zoom”, ”Line Chart”, ”Bar Chart”, “Restore”, “Data View” and “Save”. Click “Zoom In Area” to expand the area that you selected. Click “Restore Zoom” to undo

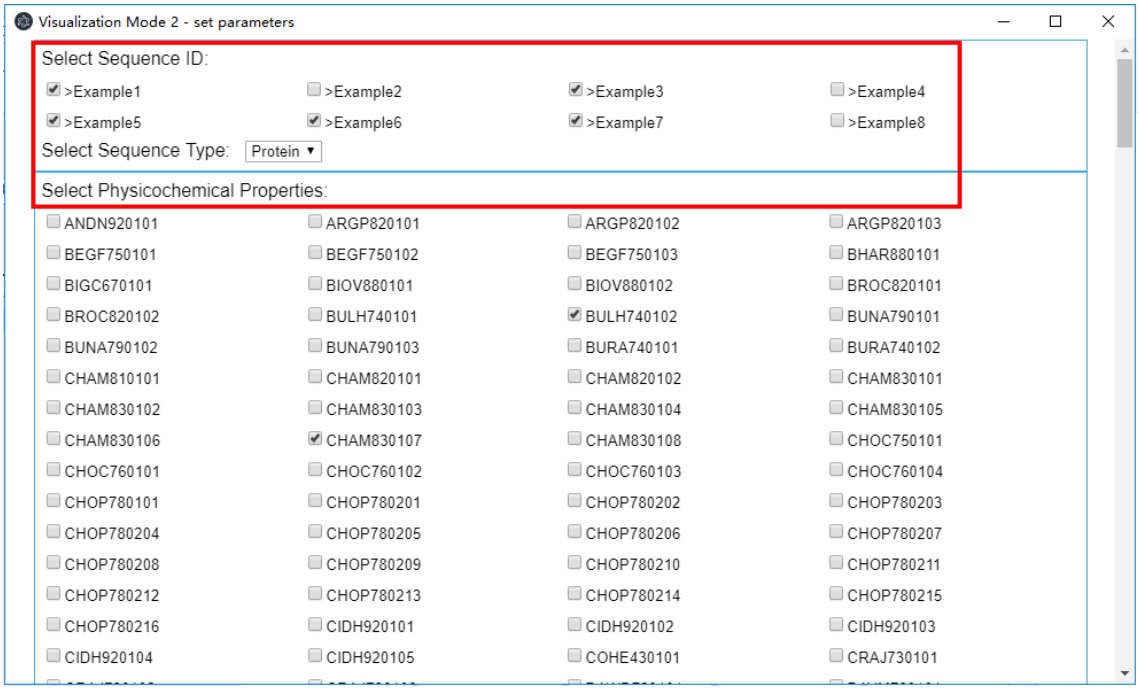
expand. Click “Line Chart” to display the line chart. Click “Bar Chart” to display the bar chart. Click “Restore” to restore the chart to its original state. Click “Data View” to view or edit data for all curves. After editing the data, you can click "Update" to update the chart. Click "Save" to save the chart to your computer hard drive.

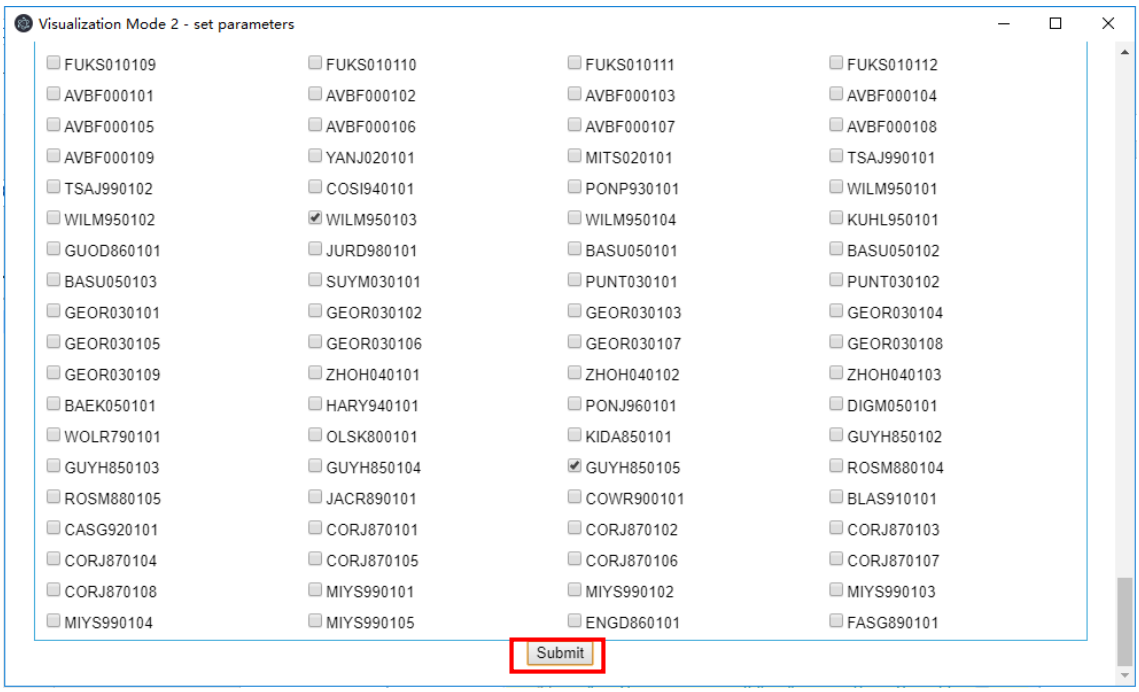
2) Visualization Mode 2 (Multiple sequences)

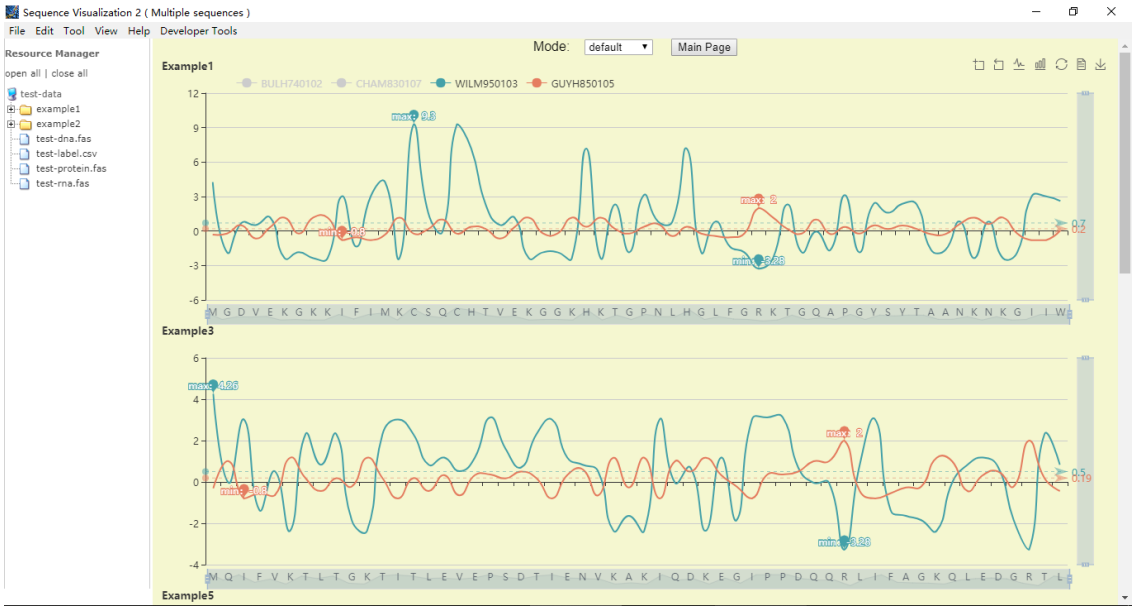
First, you need to click "Visualization Mode 2 (Multiple sequences) " from "Tool" in the menu bar, then complete all parameters and click “Submit”. Finally, the program will jump to the visual page. The visual page will display the curves according to the sequences and physicochemical properties that 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 Visualization Mode 1 (Single sequence).

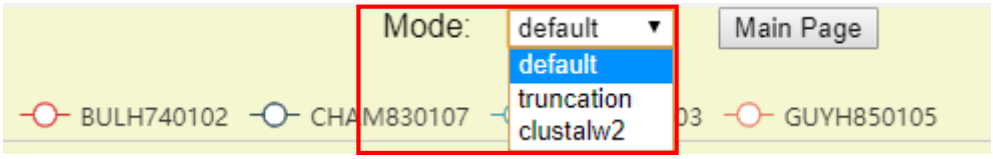
- If you select " truncation ", the program will truncate the sequences according to the shortest sequence of all the sequences you select, which means that the length of all sequences in this sub-mode is the same.
- If you select " clustalw2", the program will use clustalw2 for multiple sequence alignment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

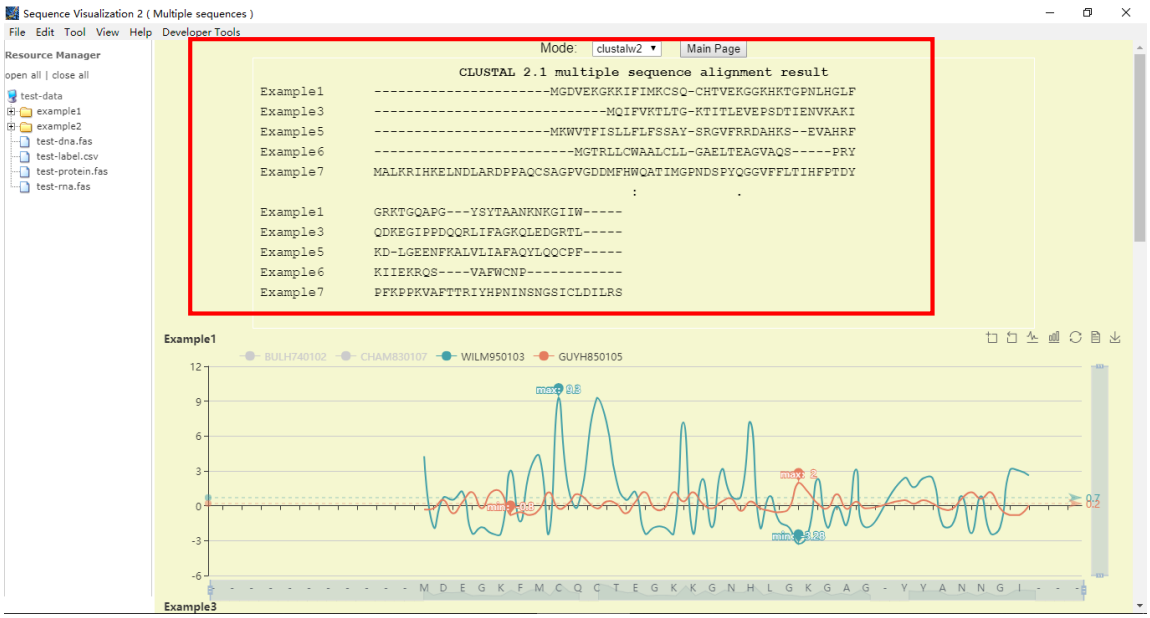

VisPse
File Edit Tool View Help
Main Page Ctrl+M
Fetch Data Ctrl+F
Visualization Mode 1 (Single sequence) Ctrl+D
Visualization Mode 2 (Multiple sequences) Ctrl+E
Compute And Visualization Ctrl+G
D:\test-data\test-protein.fas
test-rna.fas
test.txt
MDDREDLVYQAKLAEQAERYDEMVESMKKVAGMDVELTVEERNLLSVAYKNVIGARRASW
>Example5
MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPF
>Example6
MGTRLLCWAALCLLGAELTEAGVAQSPRYKIIEKRQSVAFWCNP
>Example7
MALKRIHKELNDLARDPPAQCSAGPVGDDMFHWQATIMGPNDSPYQGGVFFLTIHFPTDY
PFKPPKVAFTTRIYHPNINSNGSICLDILRS
>Example8
MALKRINKELSDLARDPPAQCSAGPVGDDMFHWQATI


Visualization Mode 2 - set parameters
Select Sequence ID:
>Example1
>Example2
>Example3
>Example4
>Example5
>Example6
>Example7
>Example8
Select Sequence Type: Protein
Select Physicochemical Properties:
ANDN920101
ARGP820101
ARGP820102
ARGP820103
BEGF750101
BEGF750102
BEGF750103
BHAR880101
BIGC670101
BIOV880101
BIOV880102
BROC820101
BROC820102
BULH740101
BULH740102
BUNA790101
BUNA790102
BUNA790103
BURA740101
BURA740102
CHAM810101
CHAM820101
CHAM820102
CHAM830101
CHAM830102
CHAM830103
CHAM830104
CHAM830105
CHAM830106
CHAM830107
CHAM830108
CHOC750101
CHOC760101
CHOC760102
CHOC760103
CHOC760104
CHOP780101
CHOP780201
CHOP780202
CHOP780203
CHOP780204
CHOP780205
CHOP780206
CHOP780207
CHOP780208
CHOP780209
CHOP780210
CHOP780211
CHOP780212
CHOP780213
CHOP780214
CHOP780215
CHOP780216
CIDH920101
CIDH920102
CIDH920103
CIDH920104
CIDH920105
COHE430101
CRAJ730101


Visualization Mode 2 - set parameters
FUKS010109
FUKS010110
FUKS010111
FUKS010112
AVBF000101
AVBF000102
AVBF000103
AVBF000104
AVBF000105
AVBF000106
AVBF000107
AVBF000108
AVBF000109
YANJ020101
MITS020101
TSAJ990101
TSAJ990102
COSI940101
PONP930101
WILM950101
WILM950102
WILM950103
WILM950104
KUHL950101
GUOD860101
JURD980101
BASU050101
BASU050102
BASU050103
SUYM030101
PUNT030101
PUNT030102
GEOR030101
GEOR030102
GEOR030103
GEOR030104
GEOR030105
GEOR030106
GEOR030107
GEOR030108
GEOR030109
ZHOH040101
ZHOH040102
ZHOH040103
BAEK050101
HARY940101
PONJ960101
DIGM050101
WOLR790101
OLSK800101
KIDA850101
GUYH850102
GUYH850103
GUYH850104
GUYH850105
ROSM880104
ROSM880105
JACR890101
COWR900101
BLAS910101
CASG920101
CORJ870101
CORJ870102
CORJ870103
CORJ870104
CORJ870105
CORJ870106
CORJ870107
CORJ870108
MIYS990101
MIYS990102
MIYS990103
MIYS990104
MIYS990105
ENGD860101
FASG890101
Submit


Sequence Visualization 2 (Multiple sequences)
File Edit Tool View Help Developer Tools
Resource Manager
open all | close all
test-data
example1
example2
test-dna.fas
test-label.csv
test-protein.fas
test-rna.fas
Mode: default
Main Page
Example1
BULH740102
CHAM830107
WILM950103
GUYH850105
Example3
Example5


Mode:
default
Main Page
default
truncation
clustalw2
BULH740102
CHAM830107
GUYH850105



3) Compute And Visualization

First, you need to click " Compute And Visualization " from "Tool" 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 upload a label file and 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 probability density maps of the feature vectors. There are two sub-modes in this mode, which are “single composition” and “multiple compositions”. You can save the chart to your computer hard disk by clicking "save".

- If you select "single composition", the visual page will display probability density maps of a dimension of all feature vectors in a chart. For example, if the dimension of each vector in the result is λ, then this sub-mode will generate λ charts.
- If you select "multiple compositions", the visual page will display probability density maps of all dimensions of all feature vectors in a chart. 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.

- 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** follow this format:

  ```
  sp|P99999,A
  ```

- 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,B
  ```

- 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:

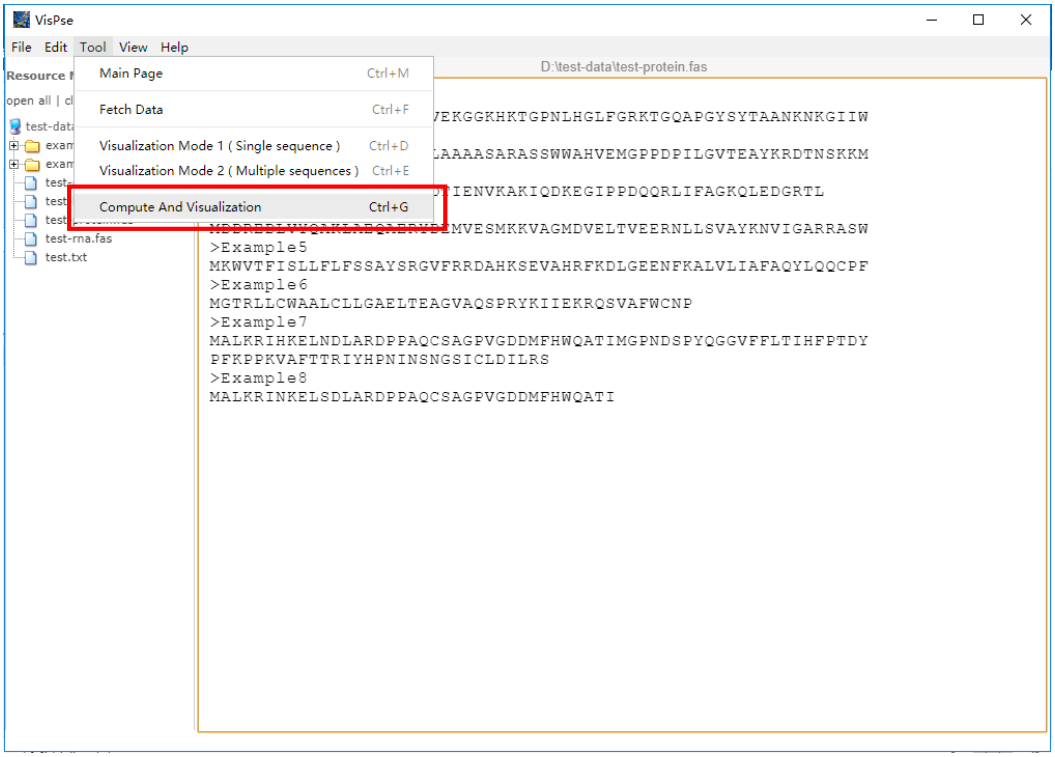
  ```
  0	1:6.564 2:1.544 3:3.475 4:5.405 5:5.019 6:6.564 7:2.703 8:5.019 9:5.792 10:10.811 11:1.544 12:3.861 13:8.494 14:5.019 15:4.247 16:6.178 17:6.564 18:6.564 19:1.158 20:3.475 # sp|Q8N2K1
  ```

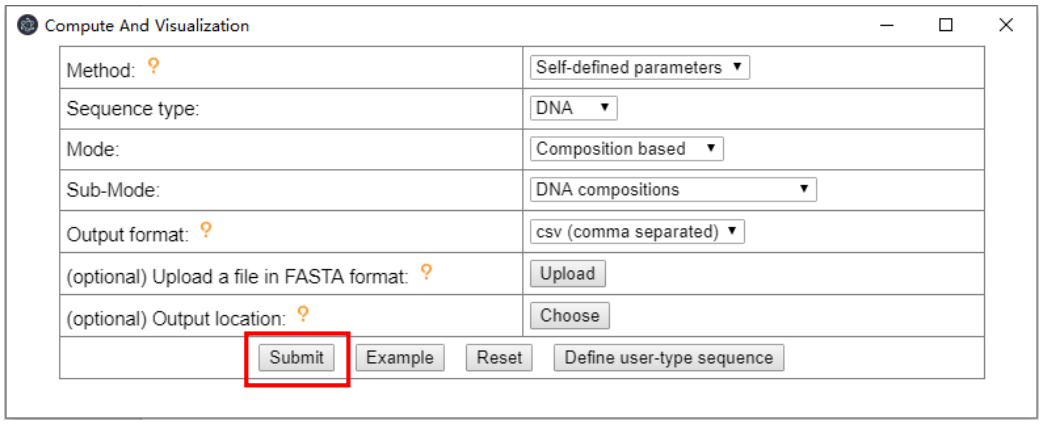
  Your label file **MUST** follow this format:

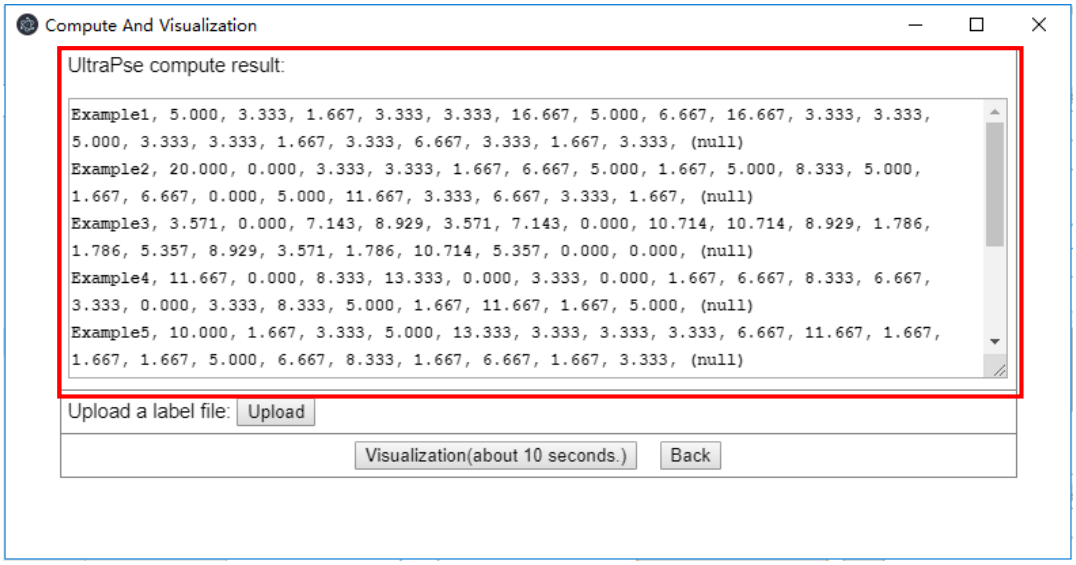
  ```
  sp|Q8N2K1,C
  ```

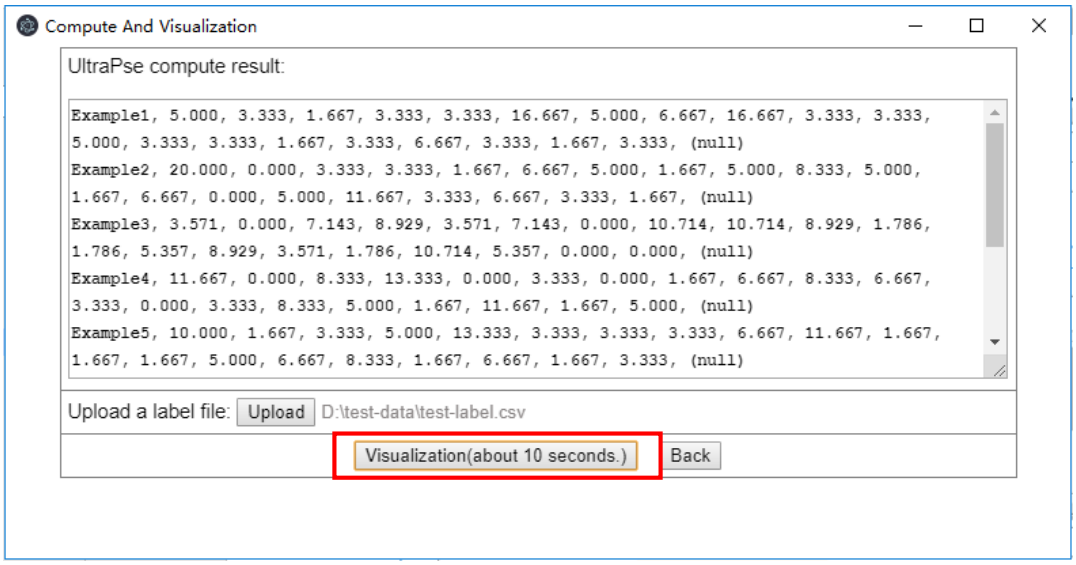
Please note:

- 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.
- The maximum dimension of the visual feature vector is **30**, exceeded parts will be ignored.
- 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**.

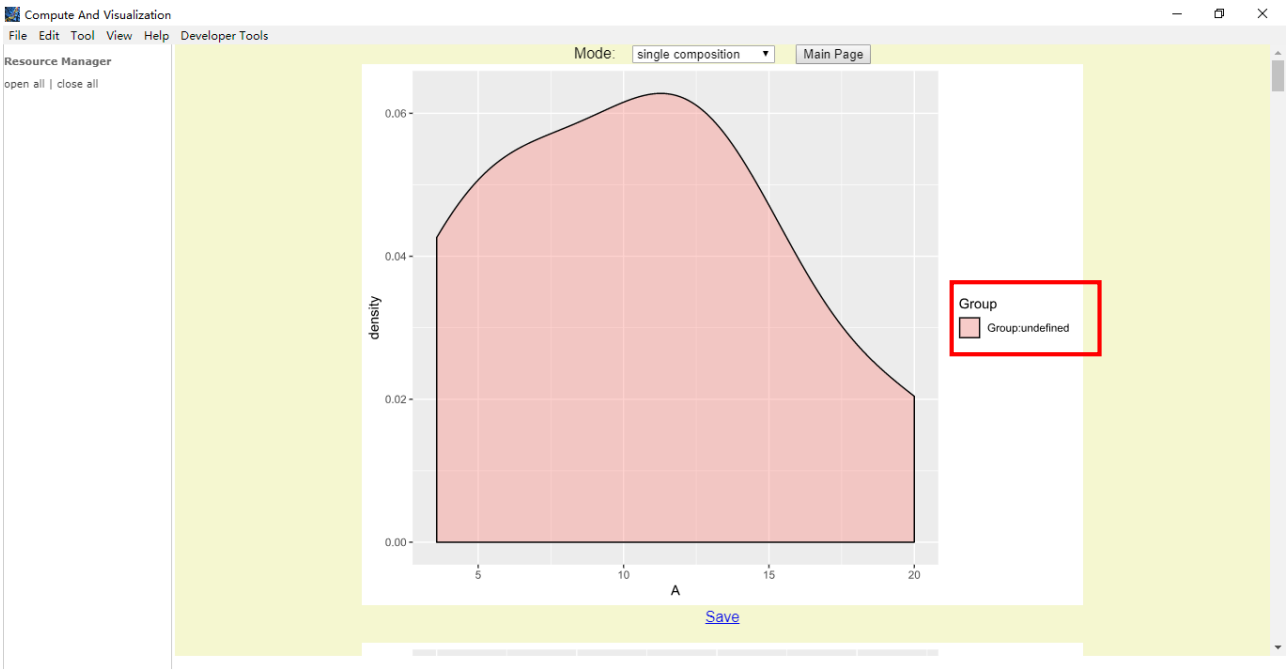



Compute And Visualization
Method:
Self-defined parameters
Sequence type:
DNA
Mode:
Composition based
Sub-Mode:
DNA compositions
Output format:
csv (comma separated)
(optional) Upload a file in FASTA format:
Upload
(optional) Output location:
Choose
Submit
Example
Reset
Define user-type sequence

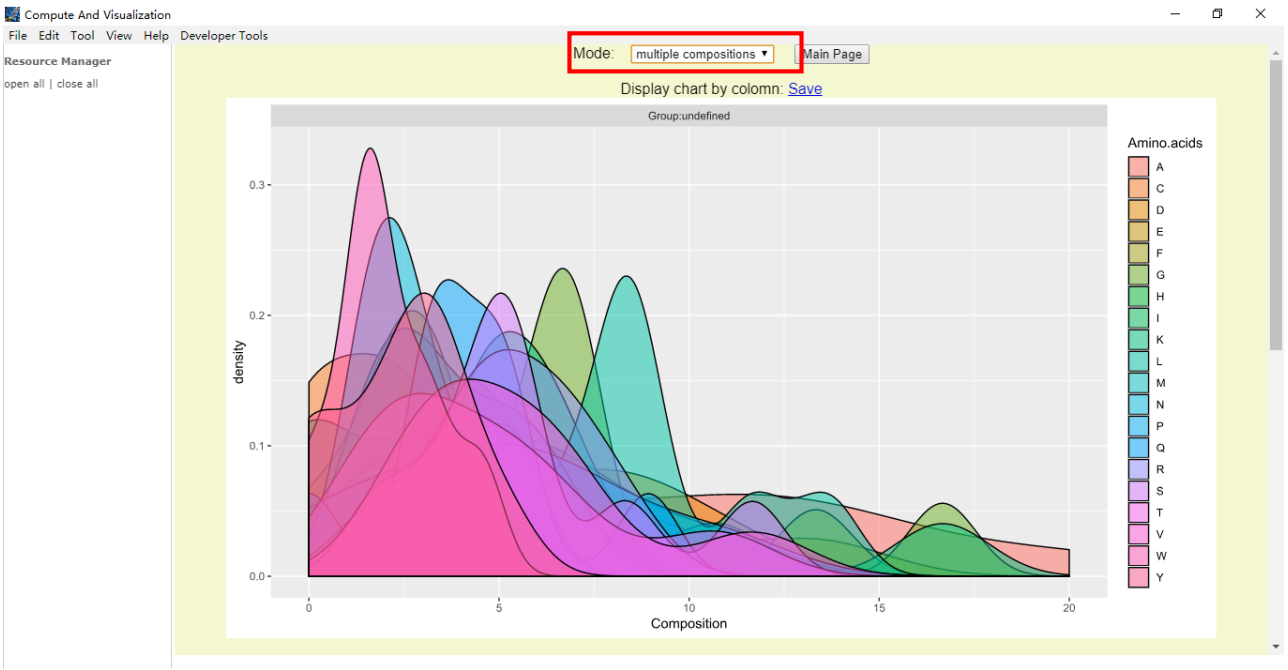

Compute And Visualization
UltraPse compute result:
Example1, 5.000, 3.333, 1.667, 3.333, 3.333, 16.667, 5.000, 6.667, 16.667, 3.333, 3.333, 5.000, 3.333, 3.333, 1.667, 3.333, 6.667, 3.333, 1.667, 3.333, (null)
Example2, 20.000, 0.000, 3.333, 3.333, 1.667, 6.667, 5.000, 1.667, 5.000, 8.333, 5.000, 1.667, 6.667, 0.000, 5.000, 11.667, 3.333, 6.667, 3.333, 1.667, (null)
Example3, 3.571, 0.000, 7.143, 8.929, 3.571, 7.143, 0.000, 10.714, 10.714, 8.929, 1.786, 1.786, 5.357, 8.929, 3.571, 1.786, 10.714, 5.357, 0.000, 0.000, (null)
Example4, 11.667, 0.000, 8.333, 13.333, 0.000, 3.333, 0.000, 1.667, 6.667, 8.333, 6.667, 3.333, 0.000, 3.333, 8.333, 5.000, 1.667, 11.667, 1.667, 5.000, (null)
Example5, 10.000, 1.667, 3.333, 5.000, 13.333, 3.333, 3.333, 3.333, 6.667, 11.667, 1.667, 1.667, 1.667, 5.000, 6.667, 8.333, 1.667, 6.667, 1.667, 3.333, (null)
Upload a label file:
Upload
Visualization(about 10 seconds.)
Back


Compute And Visualization
UltraPse compute result:
Example1, 5.000, 3.333, 1.667, 3.333, 3.333, 16.667, 5.000, 6.667, 16.667, 3.333, 3.333, 5.000, 3.333, 3.333, 1.667, 3.333, 6.667, 3.333, 1.667, 3.333, (null)
Example2, 20.000, 0.000, 3.333, 3.333, 1.667, 6.667, 5.000, 1.667, 5.000, 8.333, 5.000, 1.667, 6.667, 0.000, 5.000, 11.667, 3.333, 6.667, 3.333, 1.667, (null)
Example3, 3.571, 0.000, 7.143, 8.929, 3.571, 7.143, 0.000, 10.714, 10.714, 8.929, 1.786, 1.786, 5.357, 8.929, 3.571, 1.786, 10.714, 5.357, 0.000, 0.000, (null)
Example4, 11.667, 0.000, 8.333, 13.333, 0.000, 3.333, 0.000, 1.667, 6.667, 8.333, 6.667, 3.333, 0.000, 3.333, 8.333, 5.000, 1.667, 11.667, 1.667, 5.000, (null)
Example5, 10.000, 1.667, 3.333, 5.000, 13.333, 3.333, 3.333, 3.333, 6.667, 11.667, 1.667, 1.667, 1.667, 5.000, 6.667, 8.333, 1.667, 6.667, 1.667, 3.333, (null)
Upload a label file:
Upload
D:\test-data\test-label.csv
Visualization(about 10 seconds.)
Back

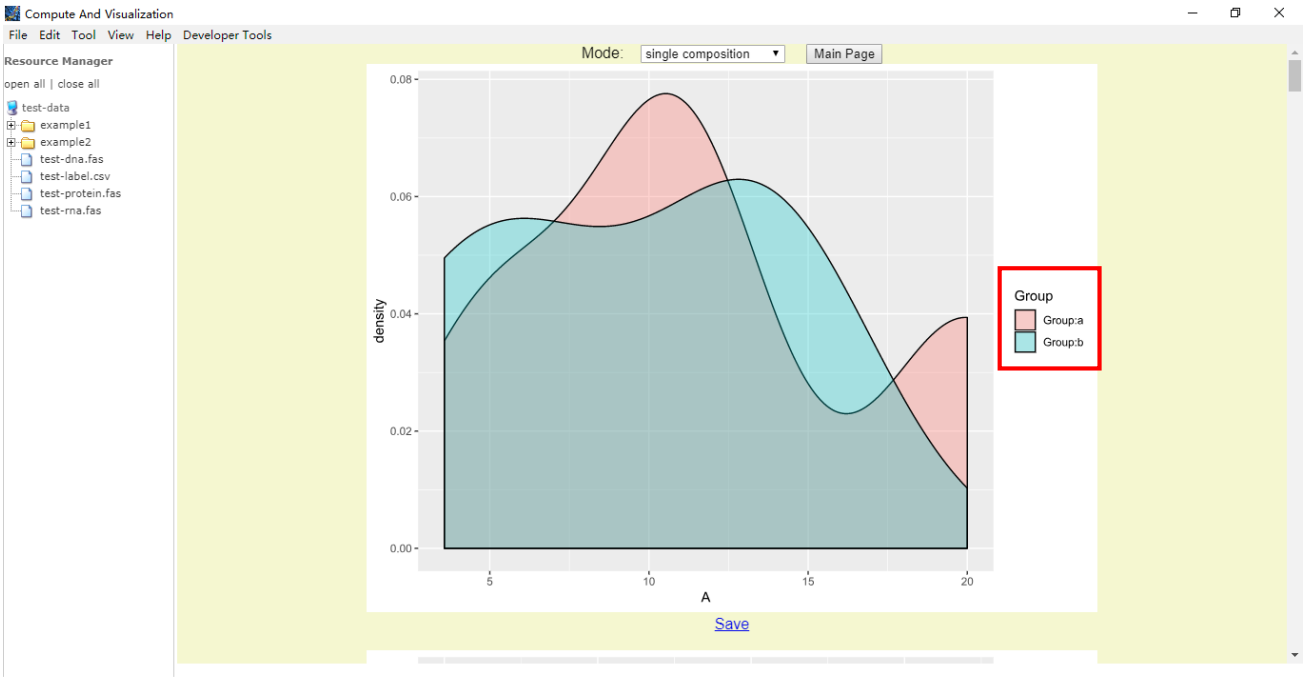
If you don't define groups of sequences, then the groups of these sequences are “Group:undefined”. The following is an example of “single composition” in which groups of all sequences are not defined. The visual page will display probability density maps of a dimension of all feature vectors in a chart.



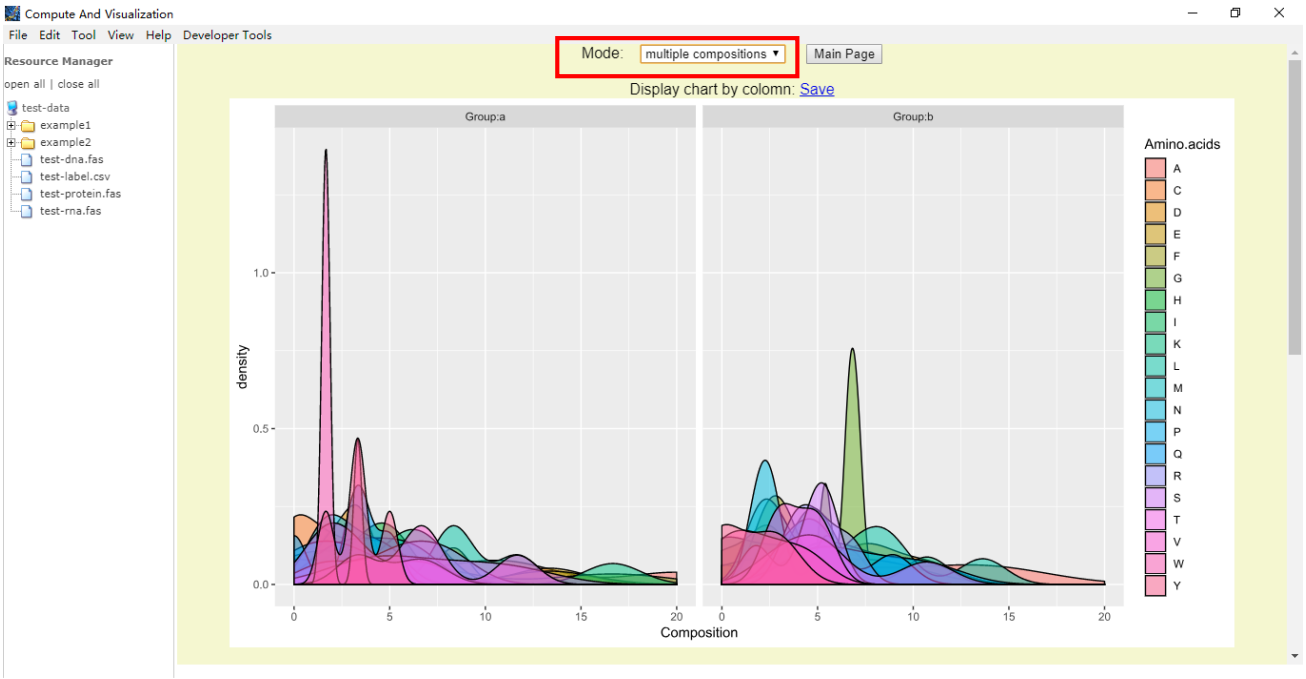
The following is an example of “multiple compositions” in which groups of all sequences are undefined. The visual page will display probability density maps of all dimensions of all feature vectors in a chart.



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



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



3 Other Information

3.1 Authors correspondence

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

3.2 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.

3.3 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/

- Lua official site

  http://www.lua.org/
- PseAAC

  http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/
- PseAAC-General

  https://github.com/pufengdu/PseAAC-General
- Pse-In-One

  http://bioinformatics.hitsz.edu.cn/Pse-in-One/home/