This is the full code description of protein function prediction based on sequence feature clustering.

1. Download the data

First of all, from the official website of UniProt database (http://www.uniprot.org/) Download Homo sapiens(human, [9606]) protein sequence (version 2020.08.03), the specific download method is: first, select the manually reviewed and annotated protein, i.e reviewed:yes Then select entry name, gene names, sequence, protein names, and gene ontology from the options list (go), click to download the information contained in the protein sequence, select to download all 20375 sequences, save the format as FASTA file, we save it as **origin data** file.

After downloading the protein sequence data, we continue to download the go number of Homo sapiens protein sequence on UniProt, and select the same reviewed:yes In the options list, select gene ontology IDs and download all 20375 sequence data. We save them as **GO-BP-CC-MF.csv**.

2. We are in GitHub(https://github.com/pufengdu/UltraPse) download the latest version of ultrapse, the folder is **Windows**. In the command line interface, we use the help document of ultrapse to select parameters.

The TDFS parameter input is classic-pseaac.lua This is the standard pseudo amino acid component transformation task. For -type,2 is selected here, -l is 2-15, step size is 1, a total of 14 parameter values. -w is 0.05-0.80, step size is 0.05, a total of 16 parameter values, where w is the ω parameter of pseudo amino acid components, that is, the proportion of sequence order utility. In this way, a 14*16 output format parameter - F is selected as SVM, that is, libsvm format. Using ultrapse's own command parameter -v to remove untransformed protein sequences, 20375 protein sequences are transformed according to the corresponding parameters. The program is **step.1**, and the results are saved as **ultrapse-file-origin**.

3. We shuffled these protein sequences with shuffle. After shuffling, we cut out the sample set (18302) and test set (2034), and selected 10% of the total number of test sets. These proteins are used to make the final prediction data, code is **step2**, and the results are saved as **training-ultrapse-file**(18302 protein) and **test-ultrapse-file**(2034 protein). And we have changed the content of the two experimental results. According to each protein, we save each set of parameter vector results as **train-protein-list&train-protein-to-vector** and **test-protein-list&test-protein-to-vector**. The execution code is **step5**.

4. In order to store the clustering results and facilitate the enrichment in the next step, we label the SP number of the cluster proteins. We use spectral clustering and other related methods in the python scientific computing library sklearn to implement spectral clustering. The function has two main parameters, numbers of clusters, which refers to the number of clusters of data output structure Neighborhood refers to the number of adjacent nodes reserved in the process of building a graph. We choose 7 parameter values of 10-40 and step size of 5. Considering the time complexity of full connection algorithm, we choose k-nearest neighbor method when constructing spectral clustering graph, in which the number of adjacent points of parameter k is fixed to 150, and we choose NCUT cutting method. The execution code is **step3**. the results are saved as **spectral-cluster-file.**

5. We choose to use the enrich go (gene, keytype, orgdb, ont, pvalue, cutoff, readable) function in the cluster profiler Library of R language to realize enrichment operation. The specific code is **step4**, and the results are saved as **bp-go-p-result，mf-go-p-result，cc-go-p-result**.

6. We use the vector value as the center point of the clustering result, that is, the value of the representative vector of each cluster. For each go enrichment result, we extract all the enriched GO terms and their corresponding P values as their tag values. The format is as follows:

"cluster_name" : "2-0.05-10-150-2"

"vector" :

"data" :

{ "go_id": "GO:0048018", "p_value": "4.30412994772681e-19"}

Among them, cluster_ Name refers to the parameter values of the clustering result, vector is the center value under the clustering result, that is, the representative value of the clustering result, and data refers to all the enriched GO terms and their corresponding p values, go_ ID refers to the ID value of the corresponding go item in the database, p_value is the p value of the enrichment result. So far, our go enrichment matching library has been established. The execution code is **step6**, and the results are saved as **bp-data,mf-data,cc-data**. We compare the real go of the predicted protein with the go library, and remove the go that is not in the go library. The execution code is **step8**, and the result is **real-protein-go-ranklist**. Because there will be very small values in enrichment and distance calculation, it will be displayed as 0 when it is implemented. According to the programming tools, we uniformly mark the point with

the minimum value of 0 as 10^{-308}.

7. The vector value of each protein to be predicted under each set of pseudo amino acid component parameters, the center point and corresponding p value of each corresponding parameter in go library were calculated according to our model score, and then the score table of the protein to be predicted was constructed by arranging the score from high to low. The execution code is **step9**, and the saved results are **protein-predict-bp-list-all, protein-predict-mf-list-all, protein-predict-cc-list-all**.

8. Our goal is to make the protein in the data set to be predicted pass through our prediction algorithm, and make its go item appear high score in our database. By observing the proportion of the real go item of the protein to be predicted, which appears in the top 100(<=10%) of our scoring table, we can evaluate whether our algorithm is effective. The number of go database tables in MF, CC and BP are 1195, 777 and 6418 respectively.

Results: in terms of molecular function, we found that the data of the top 100 proteins were as follows:

Among the 846 proteins to be predicted, there are no go terms related to molecular function, so we do not include 846 proteins, and each number is reserved to two decimal places. Among the remaining proteins to be predicted, 30.47%(362/1188) of the proteins have all gone up in the ranking list, and 50% of the 47.56%(565/1188) of the predicted proteins have been ranked up.

In terms of cell composition, we found that the data of the top 100 proteins were as follows:

Among the 951 proteins to be predicted, there were no go terms related to cell composition, so we did not include 951 proteins. Among the remaining proteins to be predicted, 35.92%(389/1083) of all go terms rose in the ranking list, and 52.35%(567/1083) of the predicted proteins had 50% GO terms.

In terms of biological processes, we found that the data statistics of the top 100 proteins were as follows:

Among the 551 proteins to be predicted, there are no go terms related to biological process, so we do not include 551 proteins. Among the remaining proteins to be predicted, 6.81%(101/1483) of all go terms have risen in the ranking list, and 50% of the predicted proteins in 14.90%(221/1483) are ranked higher.