A brief tutorial of feature generation to use in the AlzGenPred tool.

The feature generation has three main steps:

1. Prepare a list of proteins or sequences and then use that data for the classification of AD and non-AD genes. Open the STRING database (https://string-db.org/) and paste the protein names in the given window represented in **Figure 1**. If you have multiple sequences so, please select the Multiple sequences tab and then paste your sequences.

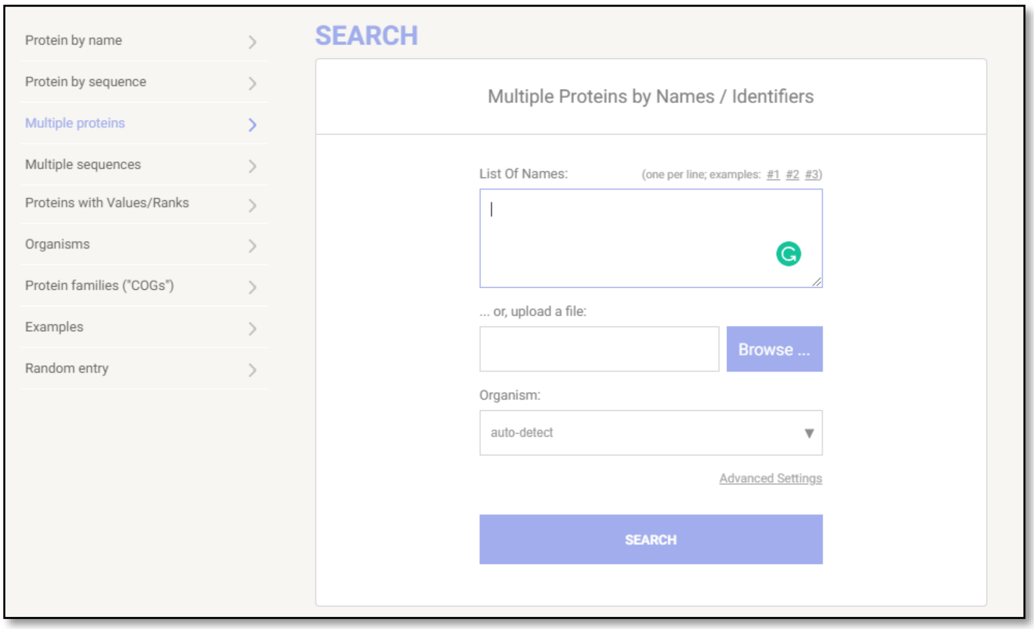
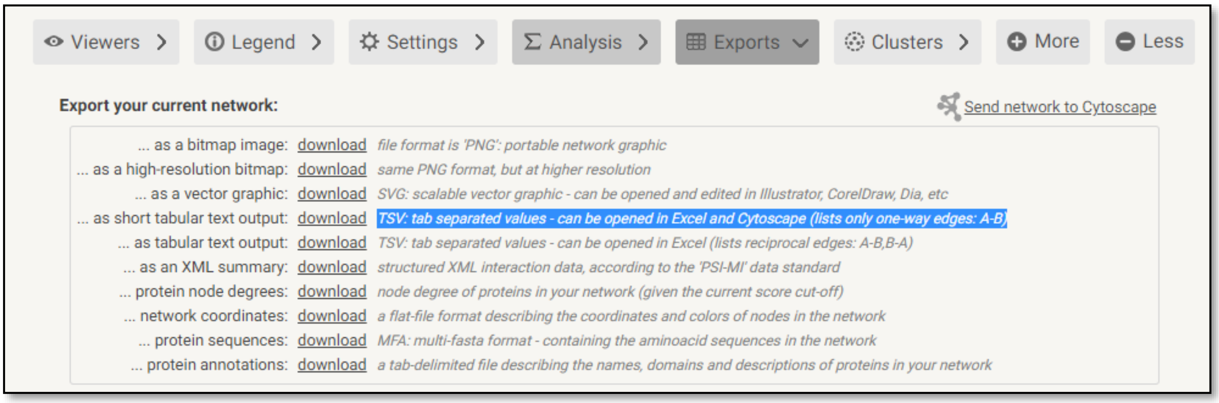


Figure 1. The input window of STRING DB.

The STRING database will identify the protein-protein interaction (PPI) for the given sequences. Then download the PPI using the highlighted tab in **Figure 2**.



Export your current network:

... as a bitmap image: download *file format is 'PNG': portable network graphic*
... as a high-resolution bitmap: download *same PNG format, but at higher resolution*
... as a vector graphic: download *SVG: scalable vector graphic - can be opened and edited in Illustrator, CorelDraw, Dia, etc*
... as short tabular text output: download *TSV: tab separated values - can be opened in Excel and Cytoscape (lists only one-way edges: A-B)*
... as tabular text output: download *TSV: tab separated values - can be opened in Excel (lists reciprocal edges: A-B,B-A)*
... as an XML summary: download *structured XML interaction data, according to the 'PSI-MI' data standard*
... protein node degrees: download *node degree of proteins in your network (given the current score cut-off)*
... network coordinates: download *a flat-file format describing the coordinates and colors of nodes in the network*
... protein sequences: download *MFA: multi-fasta format - containing the aminoacid sequences in the network*
... protein annotations: download *a tab-delimited file describing the names, domains and descriptions of proteins in your network*

Figure 2. Download tab of the STRING database. The blue colour represents the file that can be opened into the Cytoscape Software.

2. Now you have the PPI network file *(.tsv file)* which can be directly opened into Cytoscape for the generation of network-based features. If you don't have the Cytoscape so please install it from the given link (https://cytoscape.org/download.html). Then open the downloaded PPI file generated from the STRING database into Cytoscape as shown in **Figure 3**.

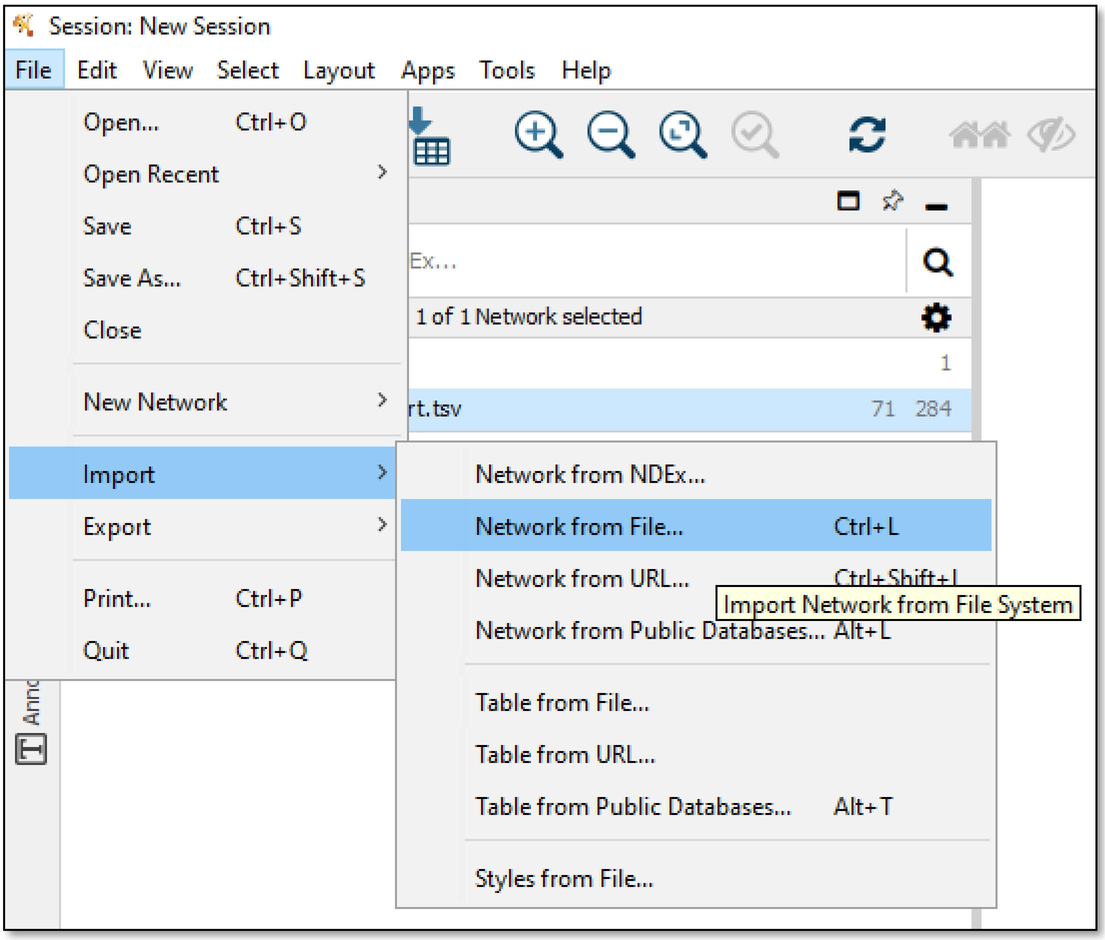


Figure 3. Open the PPI generated from the STRING database using these options.

3. It will show the PPI network. You can change the representation and can visualize the network in a different view. But here our main aim is to identify the topological parameters of the given PPI. So, use the **Figure 4** option to generate the network features. When you will click on the option "*Analyze Network*" then you will see a new window that will ask to treat the network as directed or undirected. In this option, you have to treat the network as indirect and then click on ok.

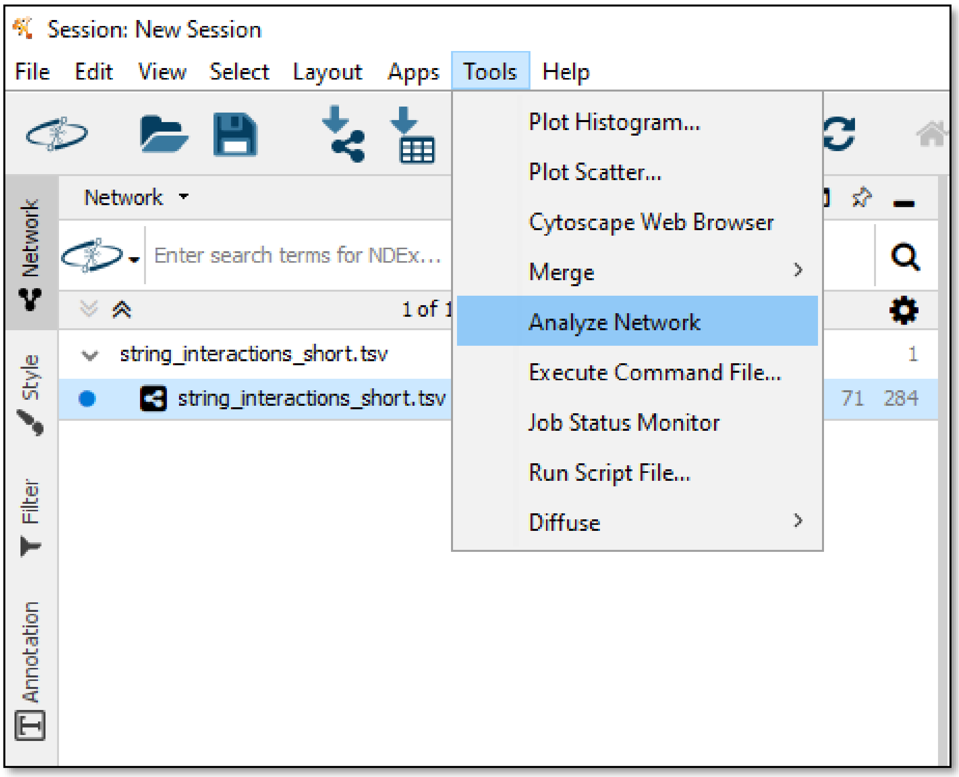
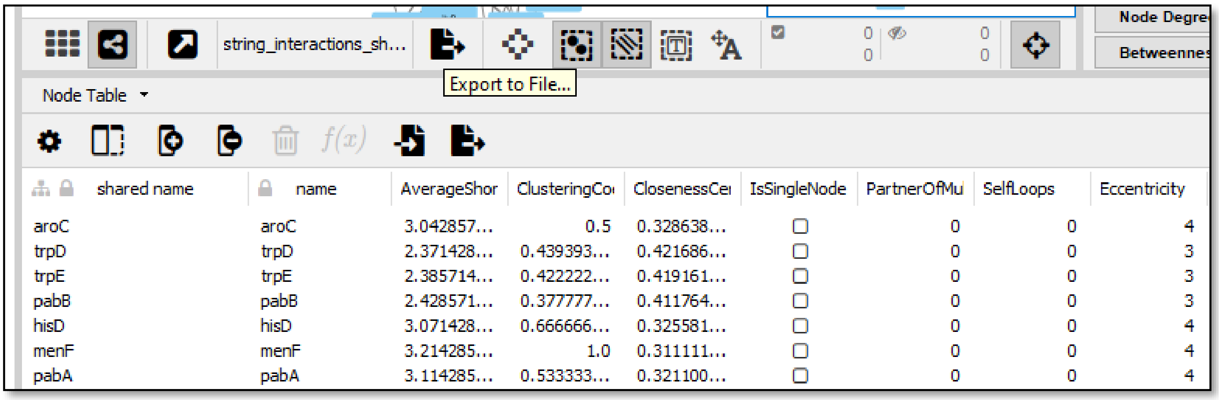


Figure 4. The topological parameters generation option.

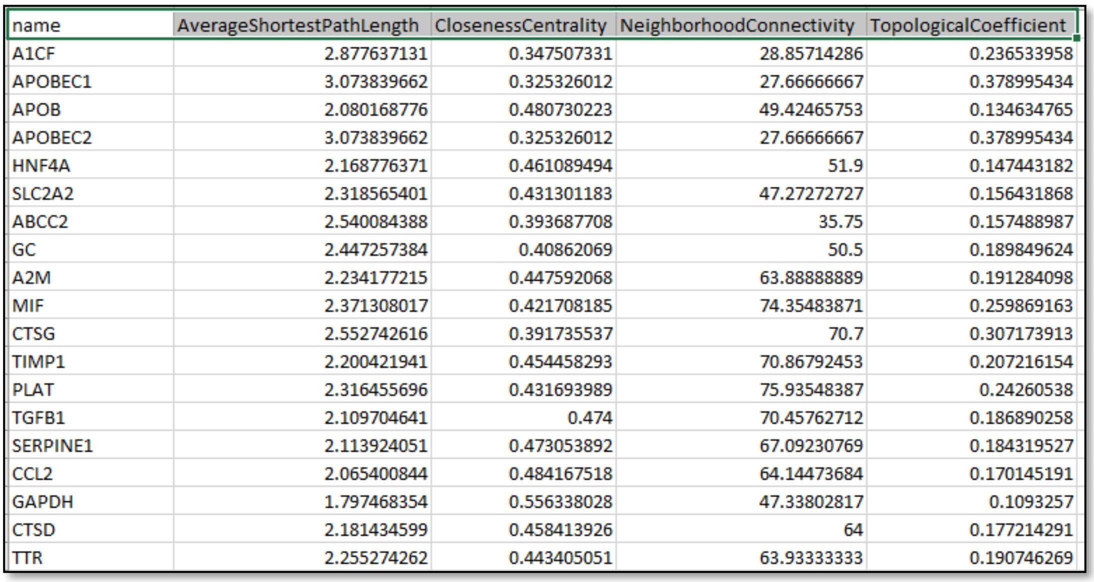
Now you have generated the topological features of the given PPI so please export the csv file of all the features as shown in **Figure 5**. You will get the diverse type of features in this file.



shared name	name	AverageShor	ClusteringCo	ClosenessCe	IsSingleNode	PartnerOfMu	SelfLoops	Eccentricity
aroC	aroC	3.042857...	0.5	0.328638...	□	0	0	4
trpD	trpD	2.371428...	0.439393...	0.421686...	□	0	0	3
trpE	trpE	2.385714...	0.422222...	0.419161...	□	0	0	3
pabB	pabB	2.428571...	0.377777...	0.411764...	□	0	0	3
hisD	hisD	3.071428...	0.666666...	0.325581...	□	0	0	4
menF	menF	3.214285...	1.0	0.311111...	□	0	0	4
pabA	pabA	3.114285...	0.533333...	0.321100...	□	0	0	4

Figure 5. The topological features and export option to export all the features in the CSV file.

Then open this file in Excel or any other editor and keep only four columns as shown in **Figure 6**. It has four features and Gene's name. If you will keep the same four columns and the name of genes so you will get the result. Please have a look at the ***topological features.csv*** so you will get a clear idea about the columns and features.



name	AverageShortestPathLength	ClosenessCentrality	NeighborhoodConnectivity	TopologicalCoefficient
A1CF	2.877637131	0.347507331	28.85714286	0.236533958
APOBEC1	3.073839662	0.325326012	27.66666667	0.378995434
APOB	2.080168776	0.480730223	49.42465753	0.134634765
APOBEC2	3.073839662	0.325326012	27.66666667	0.378995434
HNF4A	2.168776371	0.461089494	51.9	0.147443182
SLC2A2	2.318565401	0.431301183	47.27272727	0.156431868
ABCC2	2.540084388	0.393687708	35.75	0.157488987
GC	2.447257384	0.40862069	50.5	0.189849624
A2M	2.234177215	0.447592068	63.88888889	0.191284098
MIF	2.371308017	0.421708185	74.35483871	0.259869163
CTSG	2.552742616	0.391735537	70.7	0.307173913
TIMP1	2.200421941	0.454458293	70.86792453	0.207216154
PLAT	2.316455696	0.431693989	75.93548387	0.24260538
TGFB1	2.109704641	0.474	70.45762712	0.186890258
SERPINE1	2.113924051	0.473053892	67.09230769	0.184319527
CCL2	2.065400844	0.484167518	64.14473684	0.170145191
GAPDH	1.797468354	0.556338028	47.33802817	0.1093257
CTSD	2.181434599	0.458413926	64	0.177214291
TTR	2.255274262	0.443405051	63.93333333	0.190746269

Figure 6. The topological features.

Finally, you have generated the features so use these features and classify the AD-associated genes by using the ***AlzGenPred.py*** script.

For the execution of the AlzGenPred tool:

Dependencies:

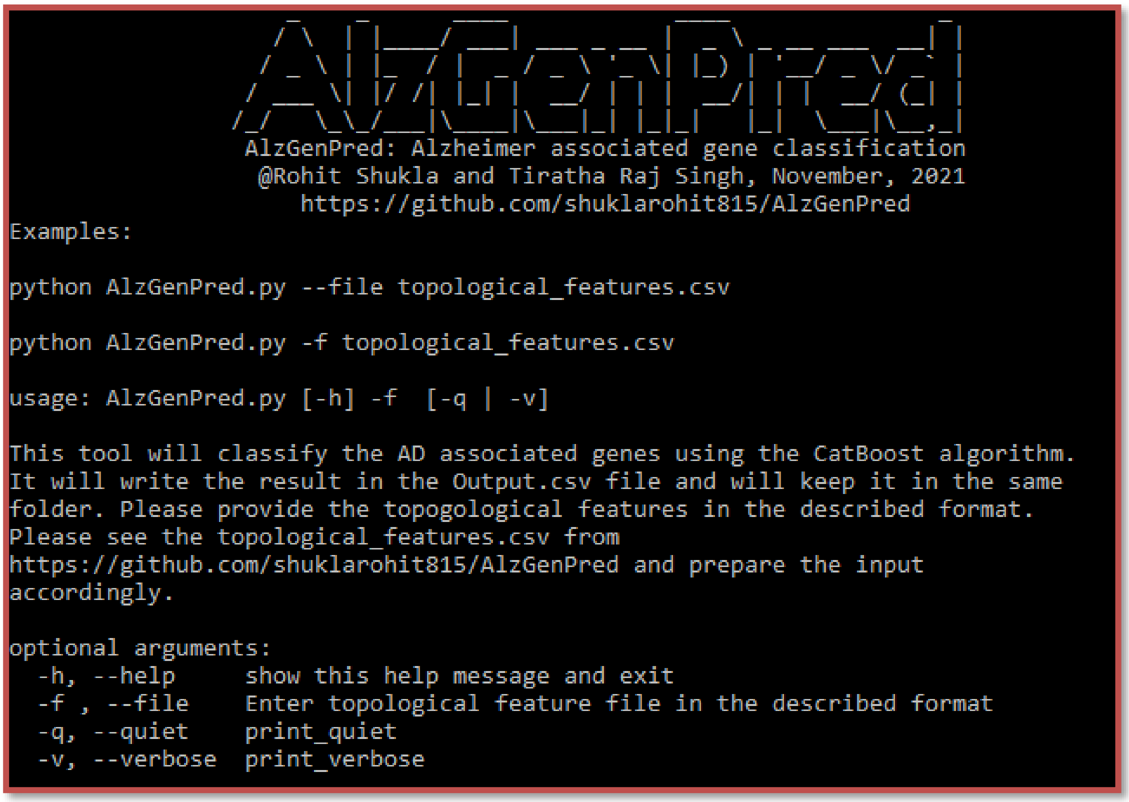
This software requires Python 3.0 or above to be installed on the system. Please make sure that python has been already installed. If not users can download and install the Python from this link https://www.python.org/downloads/. Then install the following dependencies by typing the command “***pip install dependency name***” in the command prompt.

1. Pandas
2. Pickle
3. NumPy
4. Scikit-learn
5. CatBoost

After successful installation of all these packages invokes the ***AlzGenPred.py*** tool using the below-given command. It will show a detailed help page (**Figure 7**).

Type this command in the terminal.

$ python AlzGenePred.py -h



```
                   _    _       ____            ____               _
                  / \  | |____ / ___| ___ _ __ |  _ \ _ __ ___  __| |
                 / _ \ | |_  /| |  _ / _ \ '_ \| |_) | '__/ _ \/ _` |
                / ___ \| |/ / | |_| |  __/ | | |  __/| | |  __/ (_| |
               /_/   \_\_/___|\____|\___|_| |_|_|   |_|  \___|\__,_|
                AlzGenPred: Alzheimer associated gene classification
                 @Rohit Shukla and Tiratha Raj Singh, November, 2021
                    https://github.com/shuklarohit815/AlzGenPred
Examples:

python AlzGenPred.py --file topological_features.csv

python AlzGenPred.py -f topological_features.csv

usage: AlzGenPred.py [-h] -f  [-q | -v]

This tool will classify the AD associated genes using the CatBoost algorithm.
It will write the result in the Output.csv file and will keep it in the same
folder. Please provide the topogological features in the described format.
Please see the topological_features.csv from
https://github.com/shuklarohit815/AlzGenPred and prepare the input
accordingly.

optional arguments:
  -h, --help     show this help message and exit
  -f , --file    Enter topological feature file in the described format
  -q, --quiet    print_quiet
  -v, --verbose  print_verbose
```

Figure 7. Overview of the AlzGenPred.

Then, execute the “AlzGenePred.py” script using the below-given command from any editor. The *AlzGenePred.py* and *topological_features.csv* are located in the GitHub repository (https://github.com/shuklarohit815/AlzGenPred).

$ python AlzGenePred.py --file topological_features.csv

Feel free to write at *shuklarohit815@gmail.com* for any further assistance.