PROFEAT 2016

Example Demonstration

Table of Contents

Example 1: Feature computation for a single protein sequence 2

Example 2: Feature computation for multiple protein entries 7

Example 3: Feature computation for ligands 8

Example 4: Feature computation for protein-protein interaction 11

Example 5: Feature computation for protein-ligand interaction 13

Example 6: Feature computation for biological networks 15

PROFEAT (2016) is designed to calculate feature vector for (1) Protein, (2) Protein-protein interaction, (3) Small molecule, (4) Protein-ligand interaction, and (5) Biological network.

In the following section, the usage of those functions will be demonstrated in detail.

Example 1: Feature computation for a single protein sequence

Figure 1 is the page of computing feature vectors for protein and peptide. For the calculation of a single protein or peptide, user can paste protein or peptide sequence in either FASTA or RAW format to the text input box after "Sequence". In Figure 1, the input sequence is "ACM1_HUMAN" in FASTA format.

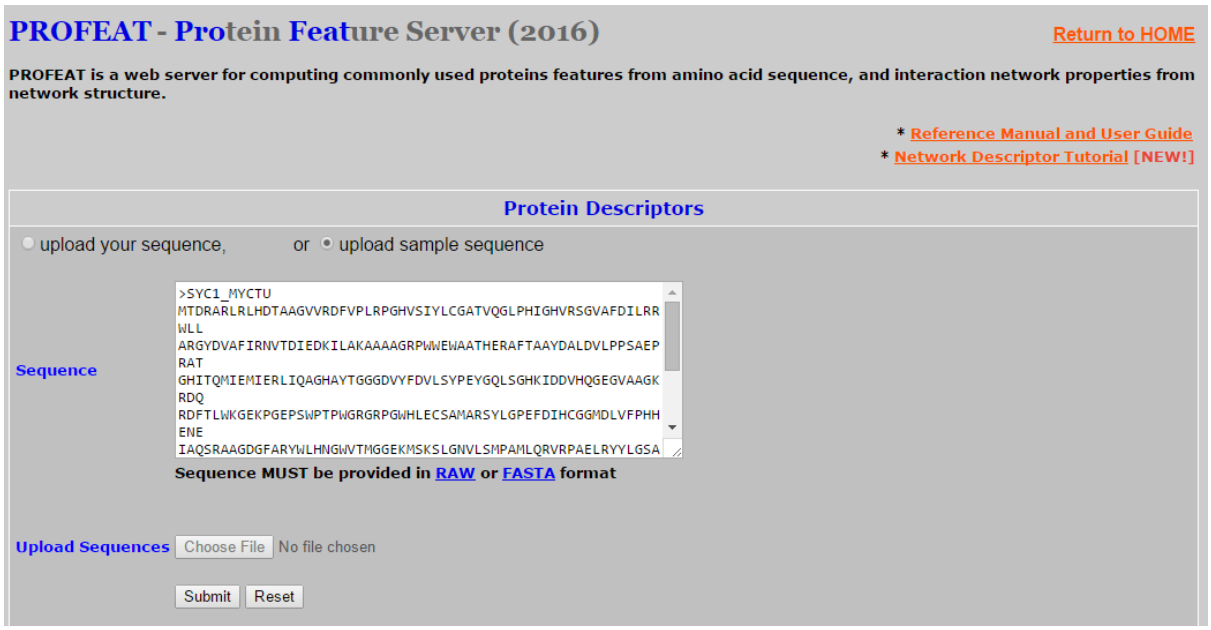


Figure 1: the page of computing feature vectors for protein and peptide

After the submission of protein or peptide sequence, the user will be redirected to the page shown in Figure 2 and 3. In Figure 2, PROFEAT asks the users to select the feature groups they are willing to calculate. There are 10 major feature groups in total, which include (1) Amino acid composition (AAC), (2) Dipeptide composition (DPC), (3) Autocorrelation descriptors (ACD), (4) Composition, transition, distribution (CTD), (5) Quasi-sequence-order descriptors (QSO), (6) Pseudo-amino acid composition (PAAC), (7) Amphiphilic pseudo-amino acid composition (APAAC), (8) Topological descriptors for atom model (TOPD), (9) Total amino acid properties (AAP) and (10) User defined property for calculating autocorrelations.

The first input parameter under the name of each feature group defines whether PROFEAT will calculate this group of feature or not. If the parameter is set as 0, PROFEAT will not calculate this feature group. If the parameter is equal to 1, the descriptor in this feature group will be calculated for

the whole protein or peptide sequence. If the parameter is given as larger than 1, PROFEAT will calculate feature vectors for the segmented sequence and the number here refers to the number of segment.

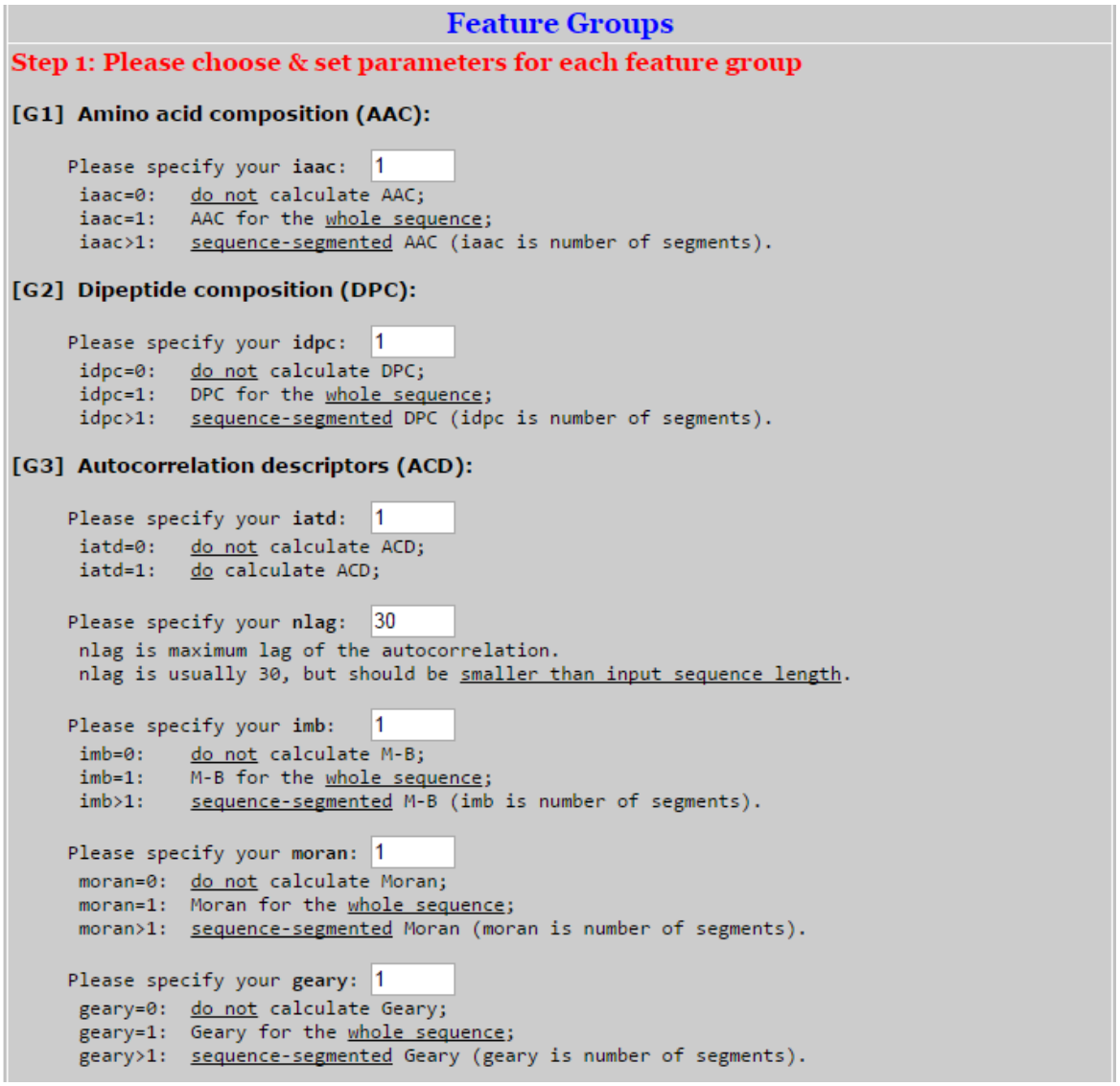


Figure 2: Step 1 for selecting feature groups to calculate descriptors for protein and peptide

Figure 3 shows a set of output format provided in 2011 version of PROFEAT. There are eight output format which are (1) Printer-friendly view of protein feature descriptor, (2) View & download GIST: tab delimited file format, (3) View & download CSV: comma delimited file format, (4) Just download GIST: tab delimited file format, (5) Just download CSV: comma delimited file format, (6) Download file in PROFEAT raw output format, (7) Download protein descriptor values in PROFEAT format and (8) Download protein descriptor names in PROFEAT format.

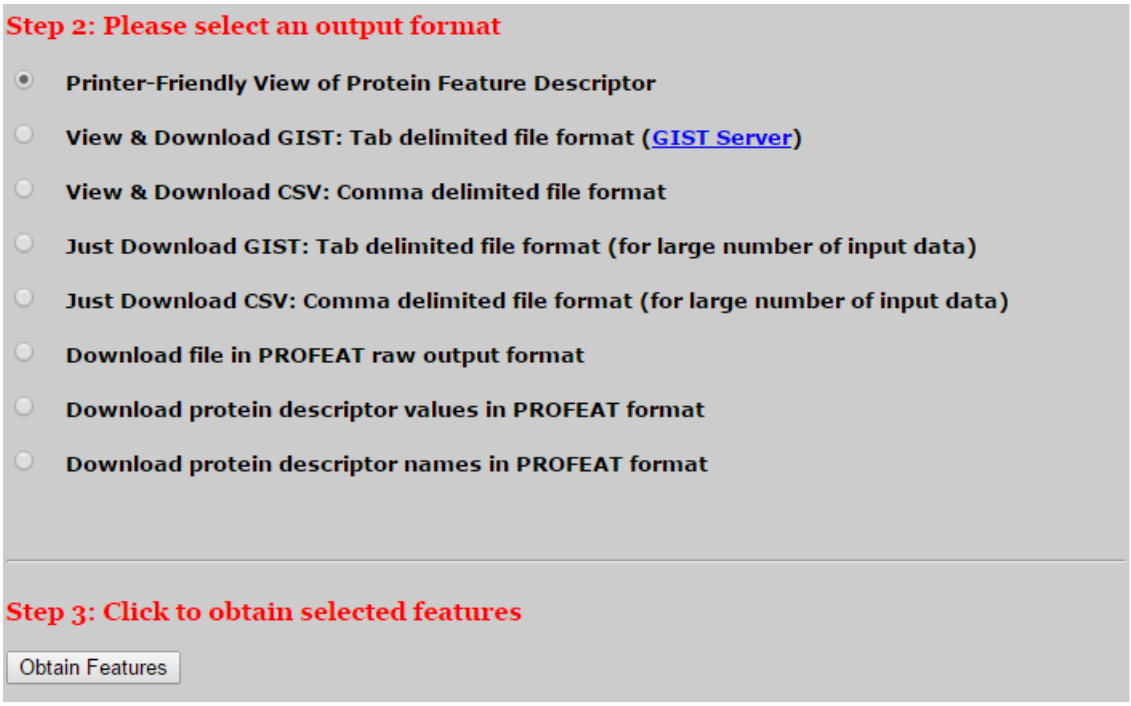
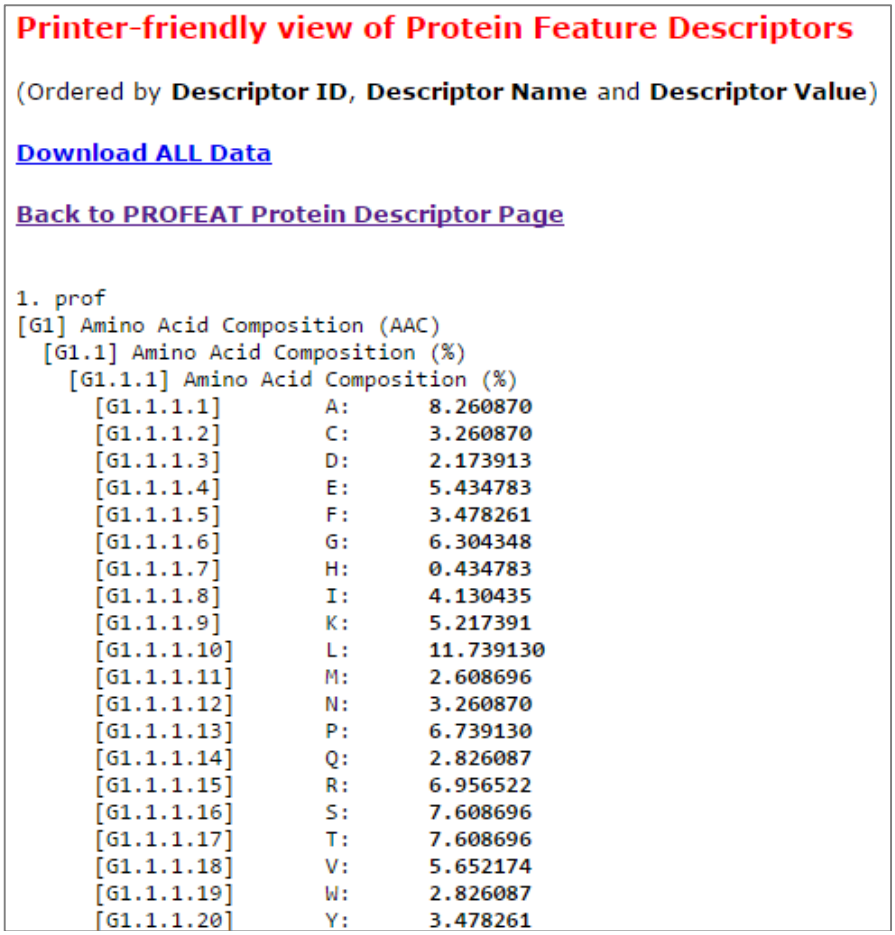


Figure 3: Step 2 & 3 for selecting output format

Figure 4 is the printer-friendly view of protein feature descriptors which calculate only the descriptors for amino acid composition. Feature vectors are listed on the right side of the descriptor name. Figure 5 is the tab delimited view of the protein feature vectors. The line start with "Feature" shows the descriptor ID defined in the printer-friendly view page, and the line under it displays the corresponding feature descriptors for those descriptor IDs. Similar to the tab delimited view, the CSV delimited view of the output data is shown in Figure 6, which separates each protein descriptor by comma. For the output format choices 4 and 5, they are actually similar to the choices 2 and 3. The only difference is that 4 and 5 are used for large data output. If the users found it takes long to get the output information, they can chose choices 4 and 5 and right click on the hyperlink on the page to save the data to their computer.



Printer-friendly view of Protein Feature Descriptors

(Ordered by **Descriptor ID**, **Descriptor Name** and **Descriptor Value**)

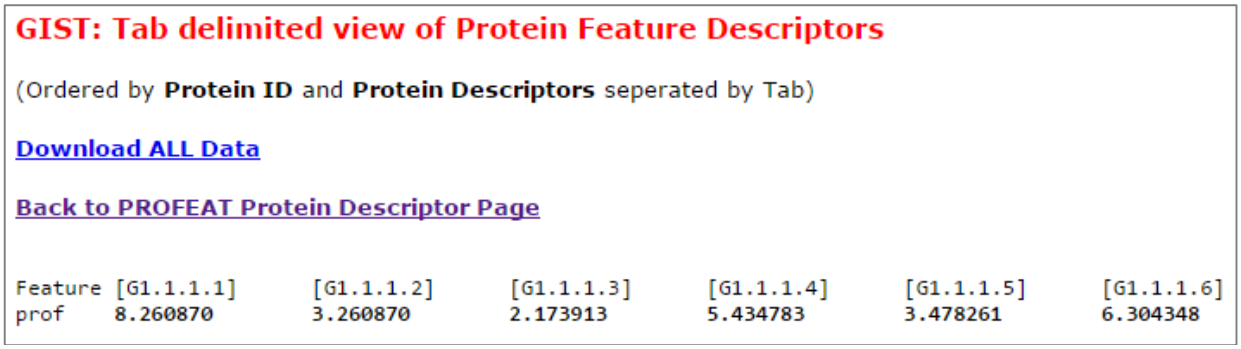
Download ALL Data

Back to PROFEAT Protein Descriptor Page

1. prof
[G1] Amino Acid Composition (AAC)
  [G1.1] Amino Acid Composition (%)
    [G1.1.1] Amino Acid Composition (%)

[G1.1.1.1]	A:	8.260870
[G1.1.1.2]	C:	3.260870
[G1.1.1.3]	D:	2.173913
[G1.1.1.4]	E:	5.434783
[G1.1.1.5]	F:	3.478261
[G1.1.1.6]	G:	6.304348
[G1.1.1.7]	H:	0.434783
[G1.1.1.8]	I:	4.130435
[G1.1.1.9]	K:	5.217391
[G1.1.1.10]	L:	11.739130
[G1.1.1.11]	M:	2.608696
[G1.1.1.12]	N:	3.260870
[G1.1.1.13]	P:	6.739130
[G1.1.1.14]	Q:	2.826087
[G1.1.1.15]	R:	6.956522
[G1.1.1.16]	S:	7.608696
[G1.1.1.17]	T:	7.608696
[G1.1.1.18]	V:	5.652174
[G1.1.1.19]	W:	2.826087
[G1.1.1.20]	Y:	3.478261

Figure 4: Printer-friendly view of protein feature descriptor



GIST: Tab delimited view of Protein Feature Descriptors

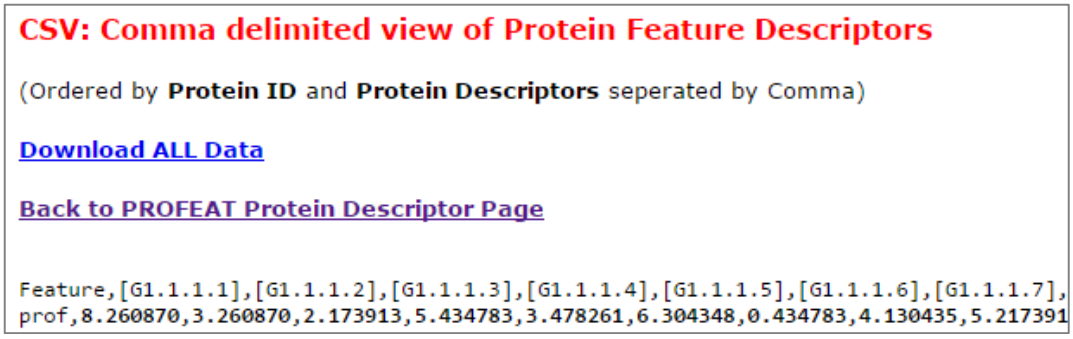
(Ordered by **Protein ID** and **Protein Descriptors** seperated by Tab)

Download ALL Data

Back to PROFEAT Protein Descriptor Page

Feature	[G1.1.1.1]	[G1.1.1.2]	[G1.1.1.3]	[G1.1.1.4]	[G1.1.1.5]	[G1.1.1.6]
prof	8.260870	3.260870	2.173913	5.434783	3.478261	6.304348

Figure 5: Tab delimited view of protein feature descriptor



CSV: Comma delimited view of Protein Feature Descriptors

(Ordered by **Protein ID** and **Protein Descriptors** seperated by Comma)

Download ALL Data

Back to PROFEAT Protein Descriptor Page

```
Feature,[G1.1.1.1],[G1.1.1.2],[G1.1.1.3],[G1.1.1.4],[G1.1.1.5],[G1.1.1.6],[G1.1.1.7],
prof,8.260870,3.260870,2.173913,5.434783,3.478261,6.304348,0.434783,4.130435,5.217391
```

Figure 6: CSV delimited view of protein feature descriptor

PROFEAT also provides information in raw output format (Figure 7), which contains all the information for the feature groups selected by the user. This raw output file is very useful especially when the user want to check the information and retrieve the descriptors into a new format recognized by whatever software. The last two options for descriptor output format are PROFEAT defined descriptors' value and their corresponding name. By using those two file, the user can easily map the output descriptors to their descriptor name.

Protein feature descriptors in PROFEAT raw output format

(Ordered by **Descriptor Names** and **Descriptor Values**)

Download ALL Data

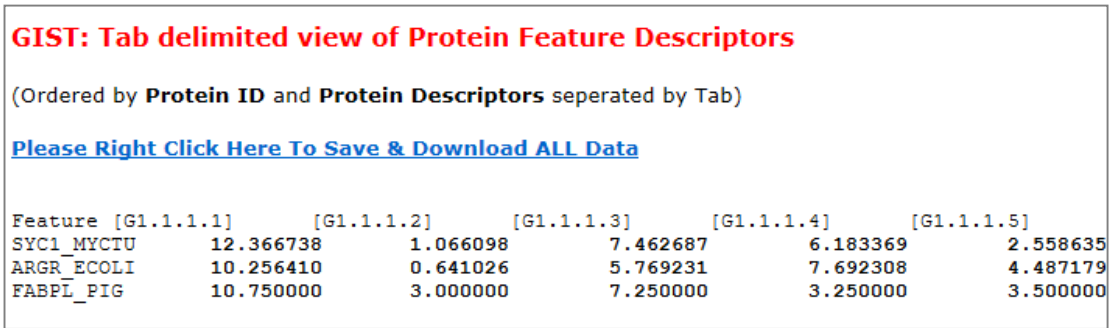
Back to PROFEAT Protein Descriptor Page

```
 ********************************
 For protein sequences

 ************************************************
Sequence     1  >prof
Length    460
 MNTSAPPAVSPNITVLAPGKGPWQVAFIGITTGLLSLATVTGNLLVLISFKVNTELKTVNNYFLLSLACADLIIGTFSMNLYTTYLLMGHWALGTLACDLWLALDYVASNA


 (1) Amino Acid Composition (AAC)

    A         C         D         E         F         G         H         I         K         L
 8.260870  3.260870  2.173913  5.434783  3.478261  6.304348  0.434783  4.130435  5.217391 11.739130
    M         N         P         Q         R         S         T         V         W         Y
 2.608696  3.260870  6.739130  2.826087  6.956522  7.608696  7.608696  5.652174  2.826087  3.478261

 (G2) Dipeptide Composition (DPC)

Residue       A         C         D         E         F         G         H         I         K         L
    A    0.871460  0.435730  0.217865  0.217865  0.871460  0.435730  0.000000  0.435730  0.435730  1.089325
    C    0.217865  0.435730  0.217865  0.000000  0.000000  0.000000  0.000000  0.000000  0.217865  0.000000
    D    0.000000  0.217865  0.000000  0.217865  0.000000  0.000000  0.000000  0.000000  0.217865  0.435730
    E    0.217865  0.000000  0.217865  1.089325  0.000000  0.653595  0.000000  0.000000  0.217865  0.653595
    F    0.000000  0.217865  0.217865  0.000000  0.000000  0.217865  0.000000  0.435730  0.217865  0.435730
    G    0.217865  0.000000  0.000000  0.435730  0.000000  0.435730  0.217865  0.217865  0.871460  0.435730
    H    0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
    I    0.000000  0.000000  0.000000  0.000000  0.000000  0.653595  0.000000  0.435730  0.217865  0.653595
    K    0.435730  0.000000  0.217865  0.653595  0.000000  0.871460  0.000000  0.217865  0.435730  0.000000
    L    2.178649  0.653595  0.217865  0.000000  0.217865  0.435730  0.000000  0.653595  0.217865  2.396514
    M    0.217865  0.435730  0.000000  0.217865  0.000000  0.217865  0.000000  0.217865  0.000000  0.000000
    N    0.217865  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.435730  0.217865  0.653595
    P    0.435730  0.000000  0.000000  0.653595  0.000000  1.307190  0.000000  0.217865  0.217865  0.217865
    Q    0.435730  0.435730  0.000000  0.000000  0.217865  0.217865  0.000000  0.000000  0.217865  0.217865
    R    1.089325  0.435730  0.435730  0.435730  0.000000  0.217865  0.000000  0.217865  0.435730  0.435730
    S    0.435730  0.000000  0.000000  0.871460  0.653595  0.000000  0.000000  0.000000  0.000000  0.871460
    T    0.217865  0.000000  0.000000  0.435730  1.089325  0.435730  0.000000  0.217865  0.435730  0.871460
    V    0.435730  0.000000  0.217865  0.000000  0.000000  0.217865  0.217865  0.217865  0.435730  1.089325
    W    0.435730  0.000000  0.217865  0.217865  0.000000  0.000000  0.000000  0.000000  0.217865  0.653595
    Y    0.217865  0.000000  0.000000  0.000000  0.435730  0.000000  0.000000  0.217865  0.000000  0.653595
```

Figure 7: Protein feature descriptors in PROFEAT raw output format

Example 2: Feature computation for multiple protein entries

Similar to computing feature vectors of single protein or peptide sequence, the calculation of feature descriptors of multiple protein or peptide entries is located at the same page on PROFEAT. In this situation, the users are asked to upload their protein sequence in FASTA format to the Browse input in Figure 1. After submission of the file, users will be redirected to the feature selection page similar to Figure 2 and 3. For the output format of the feature descriptors, there are not so many differences in the output choice 1, 6, 7 and 8. However, because this time the input file contains multiple protein sequences, the output format of choice 2~5 are different from the single situation, which are illustrated in Figure 8.



GIST: Tab delimited view of Protein Feature Descriptors

(Ordered by **Protein ID** and **Protein Descriptors** seperated by Tab)

Please Right Click Here To Save & Download ALL Data

Feature	[G1.1.1.1]	[G1.1.1.2]	[G1.1.1.3]	[G1.1.1.4]	[G1.1.1.5]
SYC1_MYCTU	12.366738	1.066098	7.462687	6.183369	2.558635
ARGR_ECOLI	10.256410	0.641026	5.769231	7.692308	4.487179
FABPL_PIG	10.750000	3.000000	7.250000	3.250000	3.500000

Figure 8: Tab delimited view of protein feature descriptor for multiple protein entries

Example 3: Feature computation for ligands

Figure 9 is the page of computing feature vectors for small molecule. The user can upload the ligand structure in SDF format to the Browse input box after "Upload Ligand SDF". The detail format for SDF file is also given on this page. User can download the SDF example file and use it as a testing data to see the calculation results.

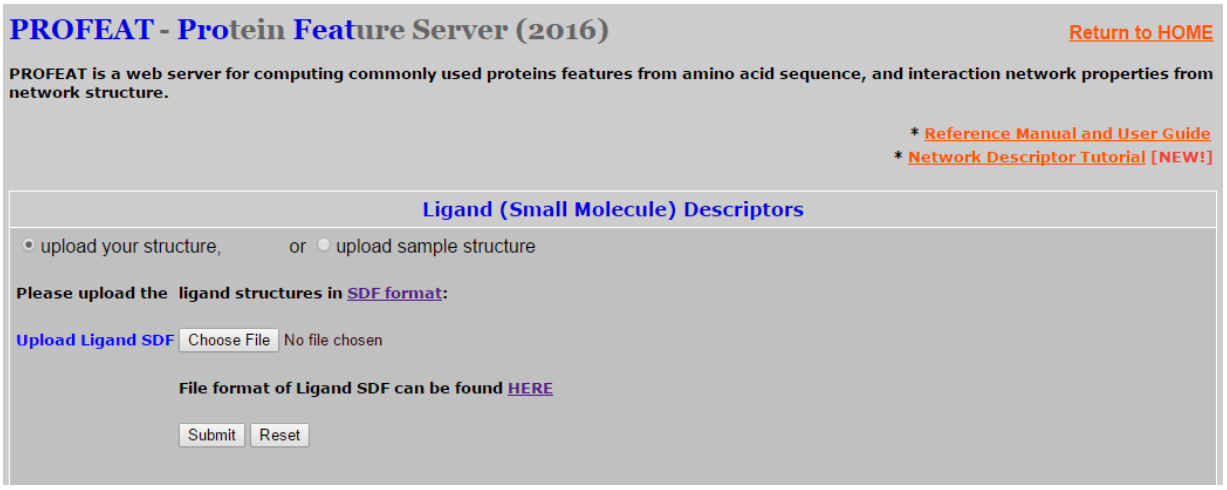


Figure 9: the page of computing feature vectors for small molecule

After the submission of small molecule SDF file, the user will be redirected to the page shown in Figure 10. In Figure 10, PROFEAT asks the users to select the output formats from seven options: (1) Download file in PROFEAT raw output format, (2) Download small molecule descriptor values in PROFEAT format, (3) Download small molecule descriptor names in PROFEAT format, (4) View & download GIST: tab delimited file format, (5) View & download CSV: comma delimited file format, (6) Just download GIST: tab delimited file format and (7) Just download CSV: comma delimited file format.

Figure 11 provides information of small molecule feature vectors in PROFEAT raw output format, which contains all the information for the descriptors calculated by PROFEAT. Similar to the protein feature descriptor raw file (Figure 7), this file would be very useful for the users. The next two options for descriptor output format are PROFEAT defined descriptors' value and their corresponding name. By using those two file, the user can easily map the output descriptors to their descriptor name.

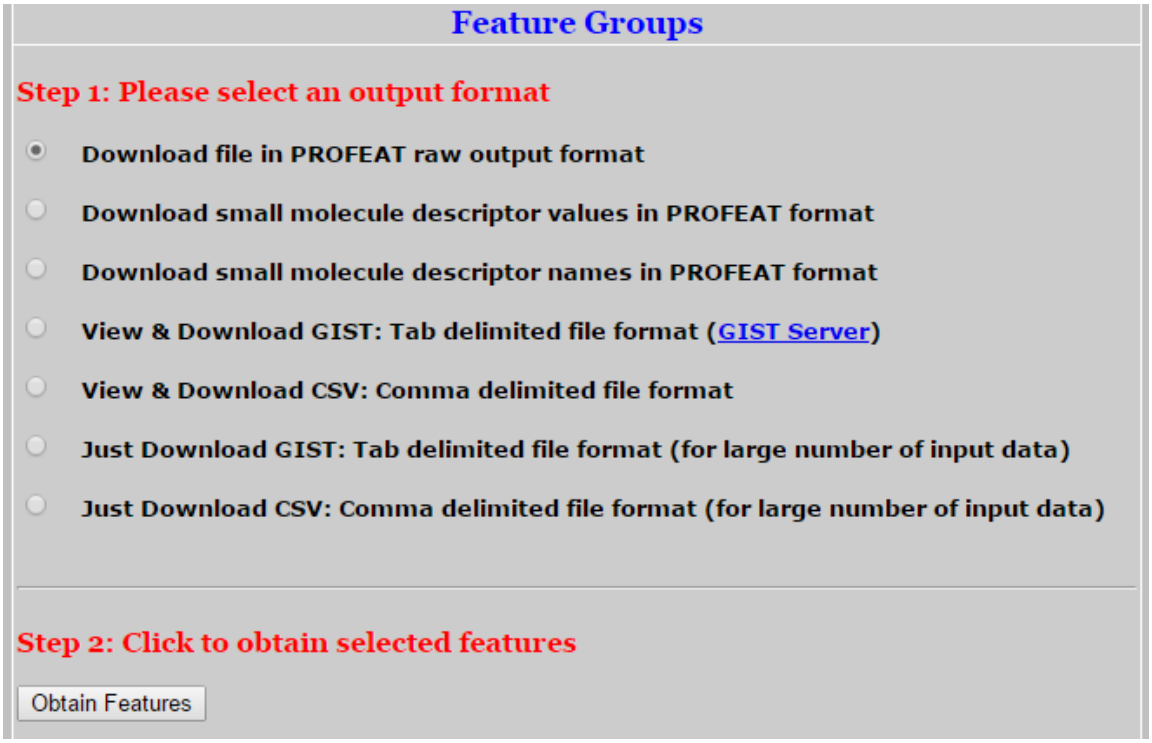


Figure 10: selecting output format for small molecule

Ligand feature descriptors in PROFEAT raw output format

(Ordered by **Descriptor Names** and **Descriptor Values**)

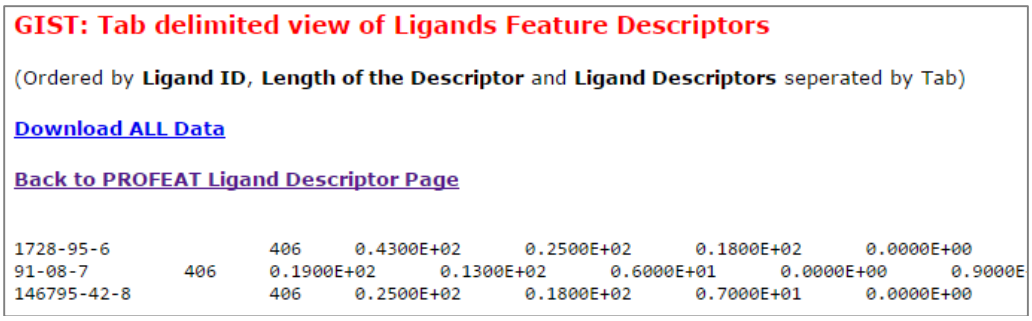
Download ALL Data

Back to PROFEAT Ligand Descriptor Page

```
Descriptors for ligand nmol= 1    1728-95-6

   Constitutational Descriptors
Number of Atoms:  43
Number of Heavy atoms:  25
Number of H atoms:   18
Number of B atoms:    0
Number of C atoms:   22
Number of N atoms:    2
Number of O atoms:    1
Number of F atoms:    0
Number of P atoms:    0
Number of S atoms:    0
Number of Cl atoms:    0
Number of Br atoms:    0
Number of I atoms:    0
Number of Bonds:   46
Number of non-H Bonds:   28
Number of rings:    4
Molecular weight(MW) :         0.3264E+03
Average molecular weight(AMW):  0.1306E+02
Number of H-bond donnor :      1
Number of H-bond acceptor:     3
```

Figure 11: Small molecule feature descriptors in PROFEAT raw output format

Figure 12 is the tab delimited view of the small molecule vectors. Each line start with small molecule ID followed by the number of feature vectors and then is the feature descriptor. Similar to the tab delimited view, the CSV delimited view of the output data is shown in Figure 13, which separates each small molecule descriptor by a comma. For the output format choices 6 and 7, they are actually similar to the choices 4 and 5. The only difference is that 6 and 7 are used for large data output.



GIST: Tab delimited view of Ligands Feature Descriptors

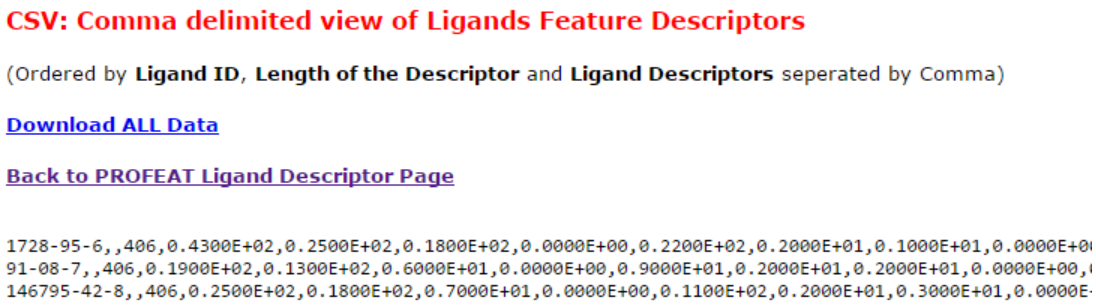
(Ordered by **Ligand ID**, **Length of the Descriptor** and **Ligand Descriptors** seperated by Tab)

Download ALL Data

Back to PROFEAT Ligand Descriptor Page

```
1728-95-6               406     0.4300E+02      0.2500E+02      0.1800E+02      0.0000E+00
91-08-7         406     0.1900E+02      0.1300E+02      0.6000E+01      0.0000E+00      0.9000E
146795-42-8             406     0.2500E+02      0.1800E+02      0.7000E+01      0.0000E+00
```

Figure 12: Tab delimited view of small molecule feature descriptor



CSV: Comma delimited view of Ligands Feature Descriptors

(Ordered by **Ligand ID**, **Length of the Descriptor** and **Ligand Descriptors** seperated by Comma)

Download ALL Data

Back to PROFEAT Ligand Descriptor Page

```
1728-95-6,,406,0.4300E+02,0.2500E+02,0.1800E+02,0.0000E+00,0.2200E+02,0.2000E+01,0.1000E+01,0.0000E+0
91-08-7,,406,0.1900E+02,0.1300E+02,0.6000E+01,0.0000E+00,0.9000E+01,0.2000E+01,0.2000E+01,0.0000E+00,
146795-42-8,,406,0.2500E+02,0.1800E+02,0.7000E+01,0.0000E+00,0.1100E+02,0.2000E+01,0.3000E+01,0.0000E
```

Figure 13: CSV delimited view of small molecule feature descriptor

Example 4: Feature computation for protein-protein interaction

Figure 14 is the page of computing feature vectors for protein-protein interaction pairs. The user can either paste a pair of protein sequences into the text input box beside the text "Sequence A" and "Sequence B" to calculate the feature vectors for a single pair of proteins or upload two files containing protein-protein interaction information to get feature descriptors for multiple protein pairs. One of the file need to be uploaded should contains the protein sequences information in FASTA format and each sequence is start with ">protein ID". Another uploaded file should provide information of protein-protein pairs in the format of "protein ID1 + protein ID2". The detail examples of the format of those two file are provide in the HOME page of computing features for protein-protein interaction.

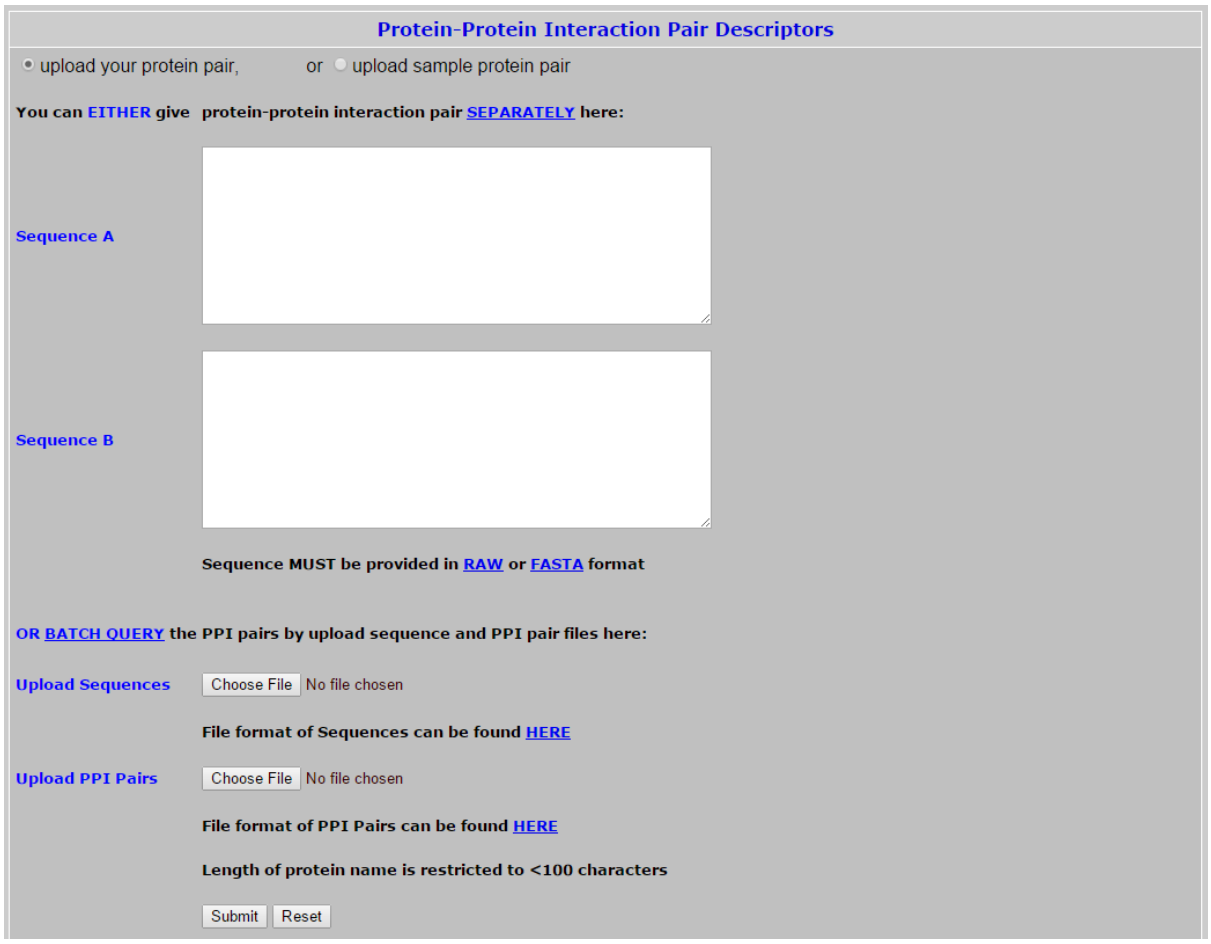


Figure 14: the page of computing feature vectors for protein-protein interaction pairs

In this Demo, we choose calculating a single pair of proteins as an example. After the submission of protein sequences of protein A and B, the user will be redirected to the feature selection page similar to Figure 2. However, the difference in this situation comparing to Figure 2 lines in that this time PROFEAT asks the users to select their methods for computing protein-protein interaction pairs (Figure 15). There are three methods.

Step 2: Please select a method to construct feature vector V for PPI pair from vectors of protein A and B: Va and Vb

- (•) **Method 1: two vectors Vab and Vba with dimension of 2n are constructed:**

  Vab=(Va,Vb) for interaction between protein A and protein B
  Vba=(Vb,Va) for interaction between protein B and protein A

- ( ) **Method 2: one vector V with dimension of 2n is constructed:**

  V={Va(i)+Vb(i), Va(i)xVb(i), i=1,2, i, n}

- ( ) **Method 3: one vector V with dimension of n2 is constructed:**

  V={Va(i)xVb(j), i=1,2, i, n, j=1,2,i,n}

Step 3: Click to obtain selected features

Obtain Features

Figure 15: selecting methods used for calculating protein-protein interaction

Feature descriptor values for PPI pair(s) in PROFEAT format

(Ordered by **PPI pair(s) Name**, **Length of the vector** and **Descriptors**)

Download ALL Data

Back to PROFEAT Protein-Protein Interaction Pair Descriptor Page

```
>PROF  +    PROF2
840
   0.1237E+02    0.1066E+01    0.7463E+01    0.6183E+01    0.2559E+01    0.9808E+01
   0.2772E+01    0.1493E+01    0.4904E+01    0.2985E+01    0.8742E+01    0.4478E+01
   0.2137E+01    0.0000E+00    0.4274E+00    0.1282E+01    0.6410E+00    0.1496E+01
   0.8547E+00    0.0000E+00    0.0000E+00    0.2137E+00    0.1068E+01    0.2137E+00
   0.0000E+00    0.0000E+00    0.2137E+00    0.0000E+00    0.0000E+00    0.4274E+00
   0.0000E+00    0.0000E+00    0.2137E+00    0.0000E+00    0.0000E+00    0.2137E+00
   0.6410E+00    0.0000E+00    0.6410E+00    0.4274E+00    0.6410E+00    0.6410E+00
   0.0000E+00    0.0000E+00    0.4274E+00    0.4274E+00    0.2137E+00    0.0000E+00
   0.2137E+00    0.2137E+00    0.4274E+00    0.0000E+00    0.4274E+00    0.4274E+00
   0.2137E+00    0.2137E+00    0.4274E+00    0.2137E+00    0.4274E+00    0.2137E+00
```

Figure 16: Feature descriptor value for PPI pair(s) in PROFEAT format

Example 5: Feature computation for protein-ligand interaction

Figure 17 is the page of computing feature vectors for protein-ligand interaction pairs. The users are required to upload three files for this function. One of the file need to be uploaded should contains the protein sequences information in FASTA format and each sequence is start with ">protein ID". The second uploaded file should provide ligand structure information in SDF format and the SDF structure for each ligand should start with their ID. The third file is required to provide information of protein-ligand pairs in the format of "protein ID + ligand ID". The detail examples of the format of those three file are provide in the HOME page of computing features for protein-ligand interaction.

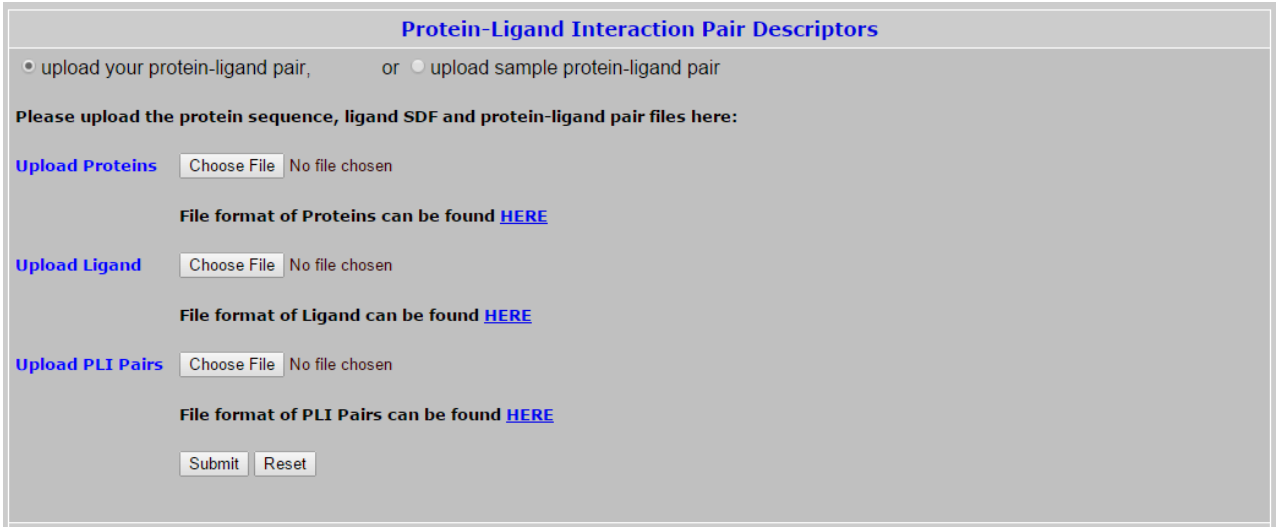


Figure 17: the page of computing feature vectors for protein-ligand interaction

After the submission of protein sequences, ligand structure and protein-ligand pairs, user will be redirected to the feature selection page similar to Figure 2. However, the difference in this situation comparing to Figure2 lines in that this time PROFEAT asks the users to select their methods for computing protein-ligand interaction pairs (Figure 18). There are two methods.

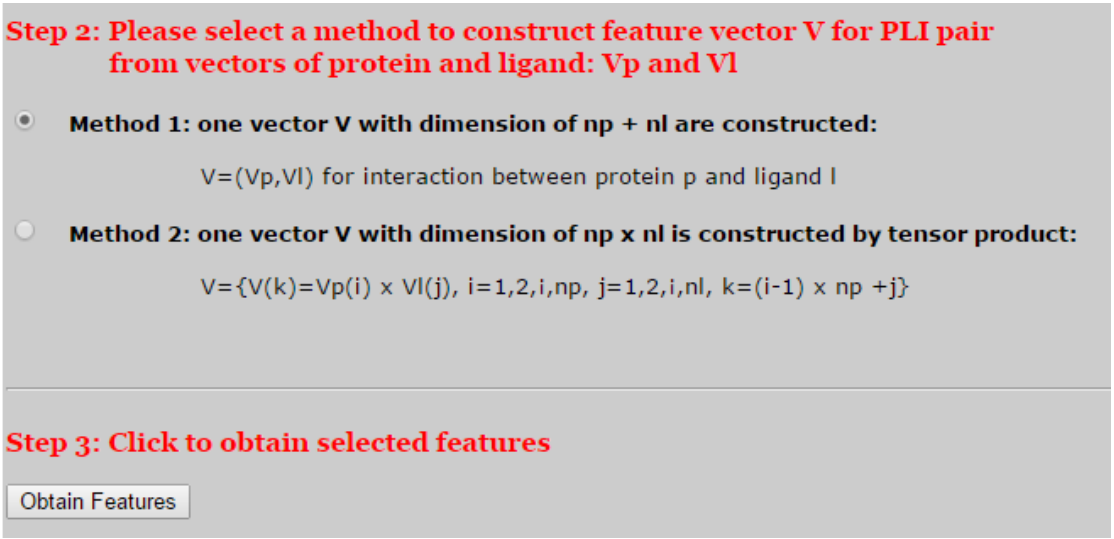


Figure 18: selecting methods used for calculating protein-ligand interaction

Feature descriptor values for PLI pair(s) in PROFEAT format

(Ordered by **PLI pair(s) Name**, **Length of the vector** and **Descriptors**)

Download ALL Data

Back to PROFEAT Protein-Ligand Interaction Pair Descriptor Page

```
>SYC1_MYCTU  +  1728-95-6
426
   0.1237E+02    0.1066E+01    0.7463E+01    0.6183E+01    0.2559E+01    0.9808E+01
   0.2772E+01    0.1493E+01    0.4904E+01    0.2985E+01    0.8742E+01    0.4478E+01
   0.4300E+02    0.2500E+02    0.1800E+02    0.0000E+00    0.2200E+02    0.2000E+01
   0.0000E+00    0.0000E+00    0.0000E+00    0.4600E+02    0.2800E+02    0.4000E+01
   0.0000E+00    0.0000E+00    0.0000E+00    0.1000E+01    0.3000E+01    0.0000E+00
   0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00
   0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00
   0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00
   0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00
   0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00    0.0000E+00
   0.6603E+04    0.1398E+05    0.3500E+01    0.2444E+01    0.1097E+01    0.1234E+01
   0.5144E-01    0.4286E-01    0.3877E+00    0.1493E+04    0.4977E+01    0.9265E+02
   0.3600E+02    0.3600E+02    0.3500E+02    0.3000E+03    0.1041E+02    0.1049E+02
   0.6028E+01    0.5611E+01    0.1900E+02    0.1758E+02    0.1381E+02    0.1461E+02
```

Figure 19: Feature descriptor value for PLI pair(s) in PROFEAT format

Example 6: Feature computation for biological networks

Figure 20 is the page of computing biological networks features (e.g. PPI network, gene regulatory network, etc.). Users have five input network options: an undirected (un-weighted, edge-weighted, node-weighted, edge-node-weighted) network, or a directed unweighted network.

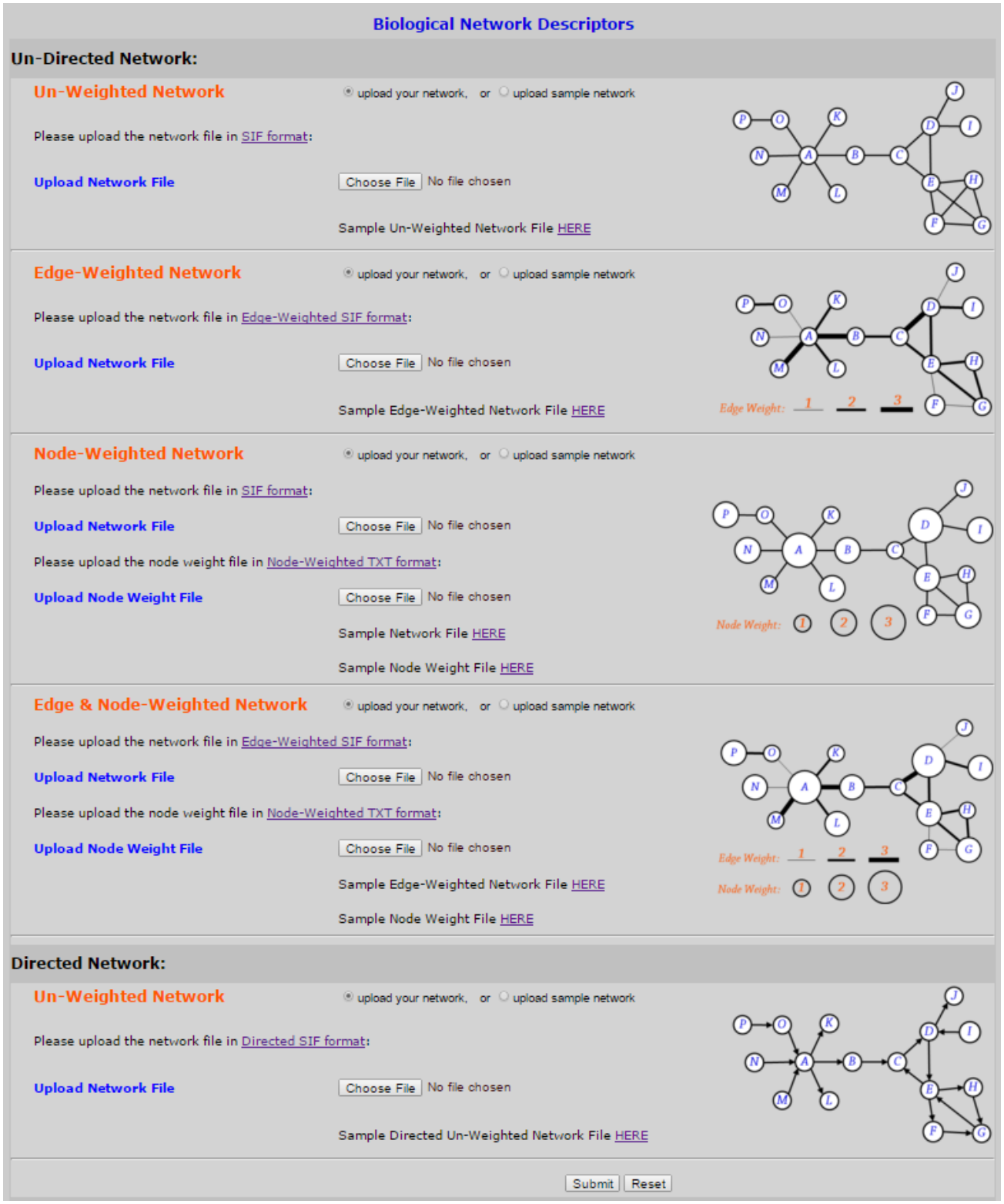


Figure 20: the page of computing features for biological networks

Additionally, there are some embedded functions programmed in PROFEAT: (1) Input network file undergoes a format check. (2) Global adjacency information are read to check if exists multiple separated networks in the single input file. PROFEAT automatically detects each connected network, ranks them by their number of nodes, renames and creates new network files by adding subscripts, and proceeds to calculate descriptors for each individual network respectively. (3) Edge length is inversely related to edge weight, as high weight typically represents strong interaction and close relation. Such that the weighted-distance-related descriptors are calculated based on the reciprocal of the edge weights. (4) Weight normalization is carried out for all edge weight and node weight, such that weighted features will be calculated based on both original weight and normalized weight.

To give a comprehensive demonstration for this newly implemented service for computing network-based descriptors, 4 case studies will be illustrated as follows.

- Undirected Un-Weighted Network
- Undirected Edge-Weighted Network
- Undirected Node-Weighted Network
- Undirected Node-Edge-Weighted Network
- Directed Un-Weighted Network
- Multiple Networks in Single Input File

Undirected Un-Weighted Network

- **Input Format:**

The network file format adopted is SIF format, namely Simple Interaction File. SIF format is tab-delimited, specifying the two linked nodes in each line, with the relationship type in between:

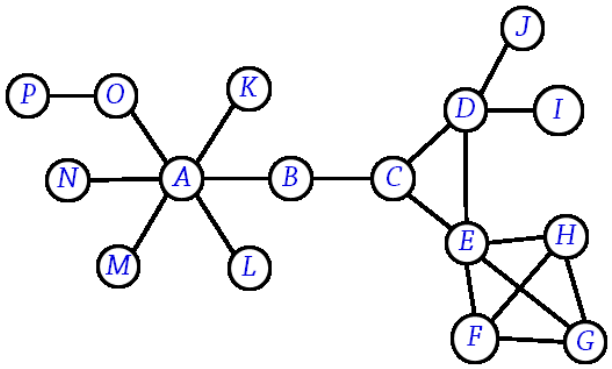
*[**node A**] tab [**relationship type**] tab [**node B**]*

Biologically, the binary interaction network could be protein-protein interaction network, gene co-expression network, gene regulatory network, drug-target network, metabolic network, etc.

- **Sample Input with Graphics:**

"sample_network.sif"

```
P   interact   O
O   interact   A
N   interact   A
M   interact   A
A   interact   B
A   interact   K
A   interact   L
B   interact   C
C   interact   D
D   interact   J
I   interact   D
D   interact   E
E   interact   C
E   interact   H
E   interact   F
F   interact   G
H   interact   G
G   interact   E
```



- **Sample Output:**

```
! Input Network File Name:                sample_network.sif
! Total Number of Networks:               1
! Total Number of Nodes:                  16
! Total Number of Edges:                  18

# Network File:                           sample_network.sif {16 Nodes; 18 Edges}
# # Node-Level Descriptors
        [G10.0.0]     Node ID:                A    B    C    ...  P
        [G10.1]       Un-Weighted Features
        [G10.1.1]     Degree:                 6    2    3    ...  1
           ...        ...
# # Network-Level Descriptors
        [G11.1]       Un-Weighted Features
        [G11.1.1]     Number of Nodes:        16
        [G11.1.2]     Number of Edges:        18
           ...        ...                     ...
```

As shown in the sample output, the header information include the input network file name, total number of networks, total number of nodes, and total number of edges. In the part of the descriptors, each descriptor is indexed, and the output are grouped into node/network-level.

Undirected Edge-Weighted Network

- **Input Format:**

Edge-weighted SIF format is defined based on SIF format, by extending the numerical edge weight for each two connected nodes in each line.

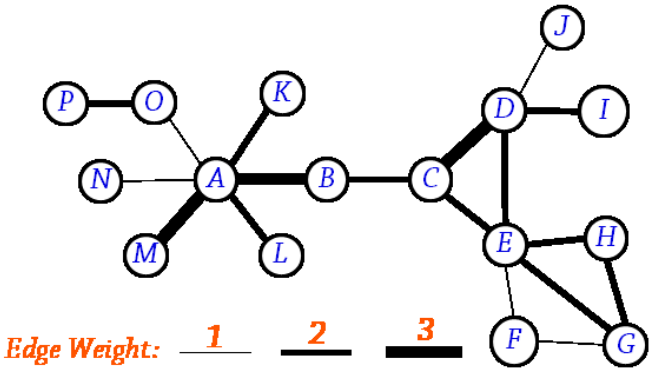
*[**node A**] tab [**relationship type**] tab [**node B**] tab [**edge weight**]*

In biological networks, the edge weight could be PPI kinetics constant, PPI binding affinity, gene co-expression association, interaction confidence level, etc.

- **Sample Input with Graphics:**

"sample_network_edgeweight.sif"

```
P   interact   O   2
O   interact   A   1
N   interact   A   1
M   interact   A   3
A   interact   B   3
A   interact   K   2
A   interact   L   2
B   interact   C   2
C   interact   D   3
D   interact   J   1
I   interact   D   2
D   interact   E   2
E   interact   C   2
E   interact   H   2
E   interact   F   1
F   interact   G   1
H   interact   G   2
G   interact   E   2
```



- **Sample Output:**

! Input Network File Name:	sample_network_edgeweight.sif					
! Total Number of Networks:	1					
! Total Number of Nodes:	16					
! Total Number of Edges:	18					
# Network File:	sample_network_edgeweight.sif {16 Nodes; 18 Edges}					
# # Node-Level Descriptors						
[G10.0.0]	Node ID:	A	B	C	...	P
[G10.1]	Un-Weighted Features					
[G10.1.1]	Degree:	6	2	3	...	1
...	...					
[G10.2]	Orginal Edge-Weighted Features					
[G10.2.11]	Edge-Weight Avg Shortest Path Length:	1.222	1.178	1.178	...	2.489
...	...					
[G10.2N]	Normalized Edge-Weighted Features					
[G10.2N.11]	N. Edge-Weight Avg Shortest Path Length:	0.642	0.641	0.641	...	1.679
...	...					
# # Network-Level Descriptors						
[G11.1]	Un-Weighted Features					
[G11.1.1]	Number of Nodes:	16				
[G11.1.2]	Number of Edges:	18				
...	...	...				
[G11.2]	Original Edge-Weighted Features					
[G11.2.14]	Edge-Weight Total Distance:	207.993				
...	...	...				
[G11.2N]	Normalized Edge-Weighted Features					
[G11.2N.14]	N. Edge-Weight Total Distance:	124.26				
...	...	...				

Undirected Node-Weighted Network

- **Input Format:**

There are 2 separated input files for a node-weighted network. One is the SIF network structure. The other is the node weight in tab-delimited txt format, specifying the node ID and its node weight numerically, while the node ID must be matched with the SIF network structure file. In biological networks, the node weight could be gene expression level, or other molecular level.

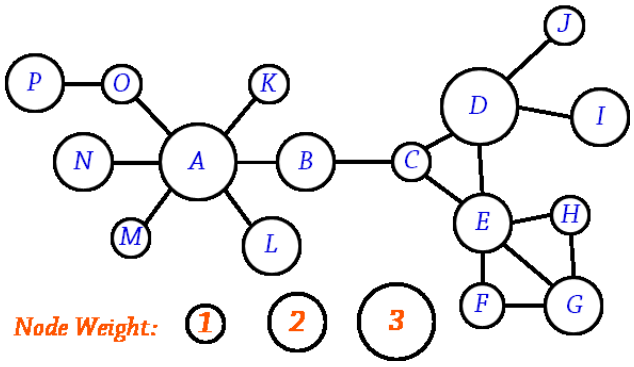
*[**node ID**] tab [**node weight**]*

- **Sample Input with Graphics:**

"sample_network.sif" *"sample_network_nodeweight.sif"*

```
P   interact   O
O   interact   A
N   interact   A
M   interact   A
A   interact   B
A   interact   K
A   interact   L
B   interact   C
C   interact   D
D   interact   J
I   interact   D
D   interact   E
E   interact   C
E   interact   H
E   interact   F
F   interact   G
H   interact   G
G   interact   E
```

```
A   3
B   2
C   1
D   3
E   2
F   1
G   2
H   1
I   2
J   1
K   1
L   2
M   1
N   2
O   1
P   3
```



- **Sample Output:**

! Input Network File Name:	sample_network.sif					
! Input Node Weight File Name:	sample_network_nodeweight.txt					
! Total Number of Networks:	1					
! Total Number of Nodes:	16					
! Total Number of Edges:	18					
# Network File:	sample_network.sif {16 Nodes; 18 Edges}					
# # Node-Level Descriptors						
[G10.0.0]	Node ID:	A	B	C	...	P
[G10.1]	Un-Weighted Features					
[G10.1.1]	Degree:	6	2	3	...	1
...	...					
[G10.3]	Original Node-Weighted Features					
[G10.3.38]	Node Weight:	3	2	1	...	3
...	...					
[G10.3N]	Normalized Node-Weighted Features					
[G10.3N.38]	N. Node Weight:	1	0.502	0.005	...	1
...	...					
# # Network-Level Descriptors						
[G11.1]	Un-Weighted Features					
[G11.1.1]	Number of Nodes:	16				
[G11.1.2]	Number of Edges:	18				
...	...	...				
[G11.3]	Original Node-Weighted Features					
[G11.3.150]	Total Node Weight:	28				
...	...	...				
[G11.3N]	Normalized Node-Weighted Features					
[G11.3N.150]	N. Total Node Weight:	6.05				
...	...	...				

Undirected Edge-Node-Weighted Network

- **Input Format:**

One edge-weighted SIF network file and one node weight TXT file are required here.

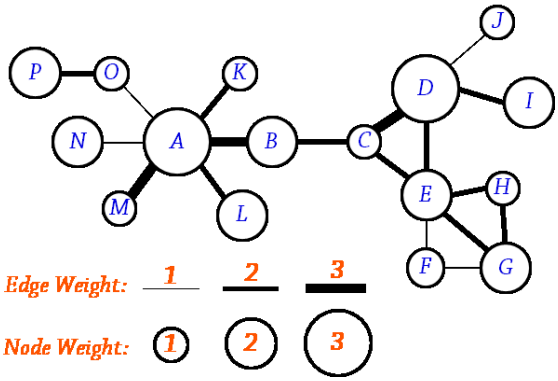
- **Sample Input with Graphics:**

"sample_network_edgeweight.sif"

```
P   interact   O   2
O   interact   A   1
N   interact   A   1
M   interact   A   3
A   interact   B   3
A   interact   K   2
A   interact   L   2
B   interact   C   2
C   interact   D   3
D   interact   J   1
I   interact   D   2
D   interact   E   2
E   interact   C   2
E   interact   H   2
E   interact   F   1
F   interact   G   1
H   interact   G   2
G   interact   E   2
```

"sample_network_nodeweight.sif"

```
A   3
B   2
C   1
D   3
E   2
F   1
G   2
H   1
I   2
J   1
K   1
L   2
M   1
N   2
O   1
P   3
```



- **Sample Output:**

! Input Network File Name:	sample_network_edgeweight.sif					
! Input Node Weight File Name:	sample_network_nodeweight.txt					
! Total Number of Networks:	1					
! Total Number of Nodes:	16					
! Total Number of Edges:	18					
# Network File:	sample_network_edgeweight.sif {16 Nodes; 18 Edges}					
# # Node-Level Descriptors						
[G10.0.0]	Node ID:	A	B	C	...	P
[G10.1]	Un-Weighted Features					
[G10.1.1]	Degree:	6	2	3	...	1
...	...					
[G10.2]	Orginal Edge-Weighted Features					
[G10.2.11]	Edge-Weight Avg Shortest Path Length:	1.222	1.178	1.178	...	2.489
...	...					
[G10.3]	Orginal Node-Weighted Features					
[G10.3.38]	Node Weight:	3	2	1	...	3
...	...					
[G10.2N]	Normalized Edge-Weighted Features					
[G10.2N.11]	N. Edge-Weight Avg Shortest Path Length:	0.642	0.641	0.641	...	1.679
...	...					
[G10.3N]	Normalized Node-Weighted Features					
[G10.3N.38]	N. Node Weight:	1	0.502	0.005	...	1
...	...					
# # Network-Level Descriptors						
[G11.1]	Un-Weighted Features					
[G11.1.1]	Number of Nodes:	16				
[G11.1.2]	Number of Edges:	18				
...	...	...				
[G11.2]	Original Edge-Weighted Features					
[G11.2.14]	Edge-Weight Total Distance:	207.993				
...	...	...				
[G11.3]	Original Node-Weighted Features					
[G11.3.150]	Total Node Weight:	28				
...	...	...				
[G11.2N]	Normalized Edge-Weighted Features					
[G11.2N.14]	N. Edge-Weight Total Distance:	124.26				
...	...	...				
[G11.3N]	Normalized Node-Weighted Features					
[G11.3N.150]	N. Total Node Weight:	6.05				
...	...	...				

Directed Un-Weighted Network

- **Input Format:**

Directed SIF format is similar with the original SIF format, but direction information is added. For the two interacting nodes in each line, the earlier one is pointing to the latter one. In the example below, it means node A points to node B (A→ B).

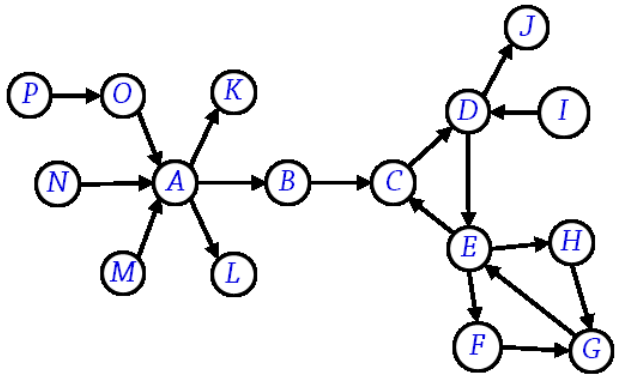
*[**node A**] tab [**relationship type**] tab [**node B**]*

In biological networks, the directed network usually represents the oriented process map (e.g. signalling pathway, metabolic reaction, etc.).

- **Sample Input with Graphics:**

"sample_network_directed.sif"

```
P   point_to    O
O   point_to    A
N   point_to    A
M   point_to    A
A   point_to    B
A   point_to    K
A   point_to    L
B   point_to    C
C   point_to    D
D   point_to    J
I   point_to    D
D   point_to    E
E   point_to    C
E   point_to    H
E   point_to    F
F   point_to    G
H   point_to    G
G   point_to    E
```



- **Sample Output:**

```
! Input Network File Name:                sample_network_directed.sif
! Total Number of Networks:                          1
! Total Number of Nodes:                             16
! Total Number of Edges:                             18

# Network File:                  sample_network_directed.sif {16 Nodes; 18 Edges}
# # Node-Level Descriptors
          [G10.0.0]              Node ID:                    A     B     C     ...   P
          [G10.4]                Directed Features
          [G10.4.41]             In-Degree:                  3     1     2     ...   0
          [G10.4.42]             Out-Degree:                 3     1     1     ...   1
             ...                 ...
# # Network-Level Descriptors
          [G11.4]                Directed Features
          [G11.4.1]              Number of Nodes:            16
          [G11.4.2]              Number of Edges:            18
             ...                 ...                         ...
          [G11.4.158]            Directed Global Clustering Coeff:  0.108
```

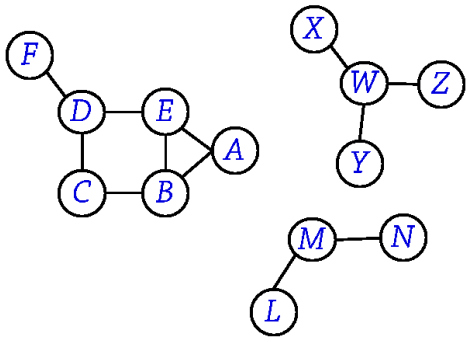
Multiple Networks in Single Input File

Network-based quantitative analysis always gets troubled by having many networks mixed in the downloaded data. Among all the existing tools, there is no one providing the function to split the disconnected network from a single input. We implemented such function in PROFEAT, and it is embedded in all types of network input. To illustrate the function, input "*sample_network_multiple.sif*" is given, which contains 3 separated networks. PROFEAT analyses the global adjacency, splits the raw input file into 3 new files, ranks them based on their number of nodes, and renames them by adding the suffix "*sub_n*". Finally, each network file will be proceed for the descriptor calculation accordingly.

- **Sample Input with Graphics:**

  "sample_network_multiple.sif"

```
A    pp   B
A    pp   E
B    pp   E
D    pp   E
C    pp   B
C    pp   D
D    pp   F
W    pp   X
W    pp   Y
W    pp   Z
L    pp   M
M    pp   N
```



- **Sample Output:**

```
! Input Network File Name:                sample_network_multiple.sif
! Total Number of Networks:               3
! Total Number of Nodes:                  13
! Total Number of Edges:                  12

# Network File:                           sample_network_multiple_sub_1.sif {6 Nodes; 7 Edges}
# # Node-Level Descriptors
        [G10.0.0]                         Node ID:                                  A    B    C    D    E    F
        [G10.1]                           Un-Weighted Features
          ...                             ...
# # Network-Level Descriptors
        [G11.1]                           Un-Weighted Features
          ...                             ...                                       ...

# Network File:                           sample_network_multiple_sub_2.sif {4 Nodes; 3 Edges}
# # Node-Level Descriptors
        [G10.0.0]                         Node ID:                                  W    X    Y    Z
        [G10.1]                           Un-Weighted Features
          ...                             ...
# # Network-Level Descriptors
        [G11.1]                           Un-Weighted Features
          ...                             ...                                       ...

# Network File:                           sample_network_multiple_sub_3.sif {3 Nodes; 2 Edges}
# # Node-Level Descriptors
        [G10.0.0]                         Node ID:                                  L    M    N
        [G10.1]                           Un-Weighted Features
          ...                             ...
# # Network-Level Descriptors
        [G11.1]                           Un-Weighted Features
          ...                             ...                                       ...
```