PROFEAT 2016

User Guide (Input & Output)

Table of Contents

In this user manual, we will illustrate the input & output file format for calculating the descriptors for: (1) protein, (2) small molecule, (3) protein-protein interaction pair, (4) protein-ligand interaction pair, and (5) protein network respectively.

1. Calculation of Protein Descriptors

Input File Format:

This file should contain the protein sequences (FASTA format) to be calculated. It is required for the calculation of proteins or protein-protein interactions or protein-ligand interactions.

Output File Format:

"output-protein.dat" is the output file for values of the descriptors of one or more protein sequences. The first line for each protein begins with ">" followed by the protein name. The second line is the number of descriptors, and the rest lines contains the values of the descriptors.

Example of output-protein.dat:

```
>SYC1_MYCTU
1437
0.1237E+02    0.1066E+01    0.7463E+01    0.6183E+01    0.2559E+01
0.9808E+01    0.4051E+01    0.4051E+01    0.2559E+01    0.8742E+01
0.2772E+01    0.1493E+01    0.4904E+01    0.2985E+01    0.8742E+01
0.4478E+01    0.4051E+01    0.6183E+01    0.2772E+01    0.2772E+01
0.2137E+01    0.0000E+00    0.4274E+00    0.1282E+01    0.6410E+00
0.1496E+01    0.2137E+00    0.2137E+00    0.2137E+00    0.1496E+01
0.8547E+00    0.0000E+00    0.0000E+00    0.2137E+00    0.1068E+01
```

In "output-protein.nam", each line contains the name of one descriptor for a protein sequence. The order and the number of the descriptors are in accordance with "output-protein.dat".

2. Calculation of Ligand (Small Molecule) Descriptors

Input File Format:

This file contains the information of ligands to be calculated and is in the SDF format. It is required for the calculation of ligands or protein-ligand interactions.

Output File Format:

"output-ligand.dat" is the output file for the descriptors values of one or more ligands. The meaning and the format are similar with "output-protein.dat". "output-ligand.nam" is the output file for names of the ligand descriptors, and its format is similar with "output-protein.nam".

3. Calculation of Protein-Protein Interaction Pair Descriptors

Input File Format:

This file contains the names of the interacting proteins and it is required for the calculation for protein-protein interactions. Each line of the file contains the names of the two interacting proteins separated by a "+" sign in a free format. For a pair of proteins, only one line of input is needed and the order of the names is optional.

Note: the protein sequences must be present in "input-protein.dat" and the names are consistent.

Output File Format:

This is the output file for the descriptors of protein-protein interaction. For each interacting pair, the first line begins with ">" followed by the two names of the two interacting proteins separated by "+" and the second line is the number of descriptors and the rest lines are the values of the descriptors.

4. Calculation of Protein-Ligand Interaction Pair Descriptors

Input File Format:

This file contains the names of the interacting protein and ligand and it is required for the calculation of protein-ligand interactions. Each line of the file contains the protein name and the ligand name separated by a "+" sign in a free format.

Note: the protein sequences must be present in "input-protein.dat" and the ligand must present in "input-ligand.sdf". Again, the names should be consistent.

Output File Format:

In the output file of protein-ligand interaction descriptors, each interacting pair has its first line begins with ">" followed by the names of the interacting protein and ligand separated by "+", the second line is the number of descriptors, and the rest lines are the values of the descriptors.

5. Calculation of Biological Network Descriptors

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:

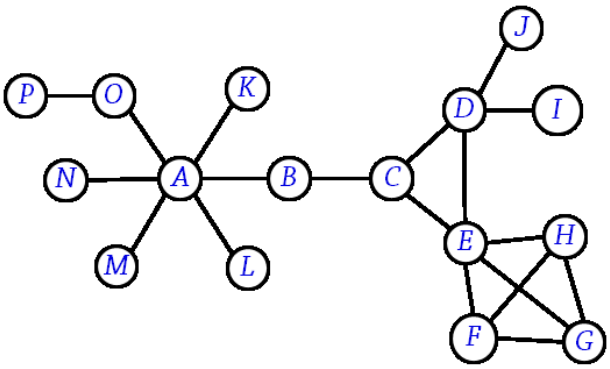
*[**node A**] tab [**relationship type**] tab [**node B**]*

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.

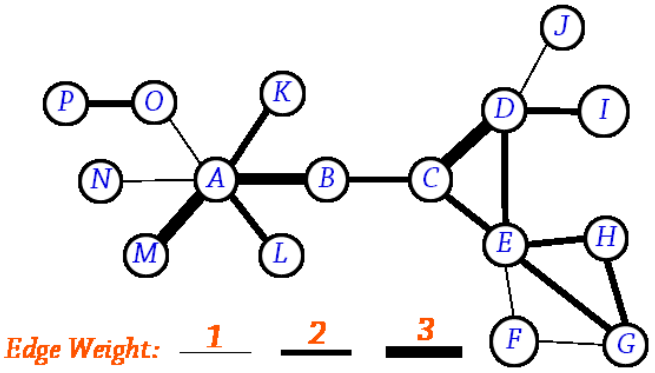
*[**node A**] tab [**relationship type**] tab [**node B**] tab [**edge weight**]*

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.

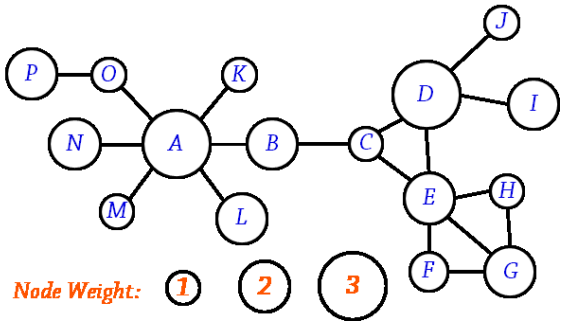
*[**node ID**] tab [**node weight**]*

- **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:**

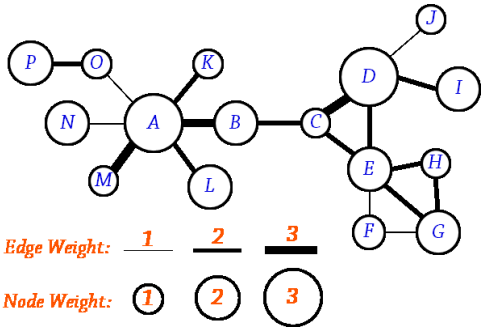
One edge-weighted SIF network file and one node weight TXT file are required here.

- **Sample Input with Graphics:**

"sample_network_edgeweight.sif" "sample_network_nodeweight.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
```

```
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).

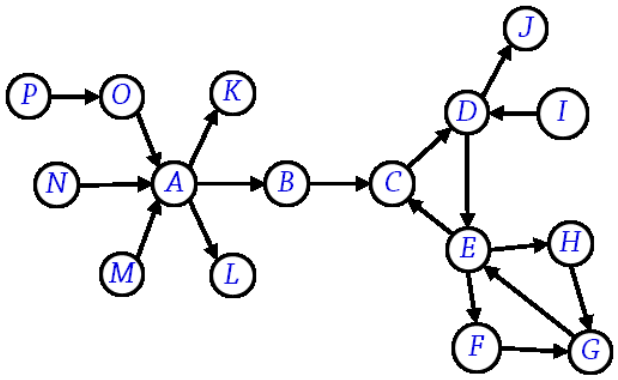
*[**node A**] tab [**relationship type**] tab [**node B**]*

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
```

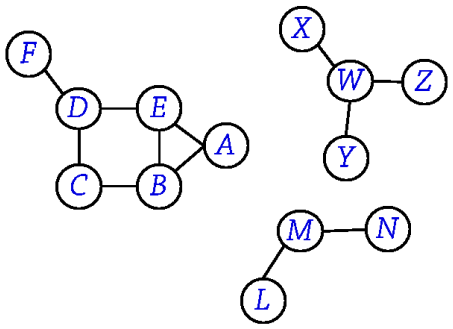
Multiple Networks in Single Input File

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
          ...                            ...                               ...
```