S2SNet - Sequence to Star Network

< Tutorial >

Cristian R. Munteanu & Humberto González-Díaz

muntisa@gmail.com

https://www.box.com/s/kd07m9kgbd0ate2uqmsr

What is S2SNet

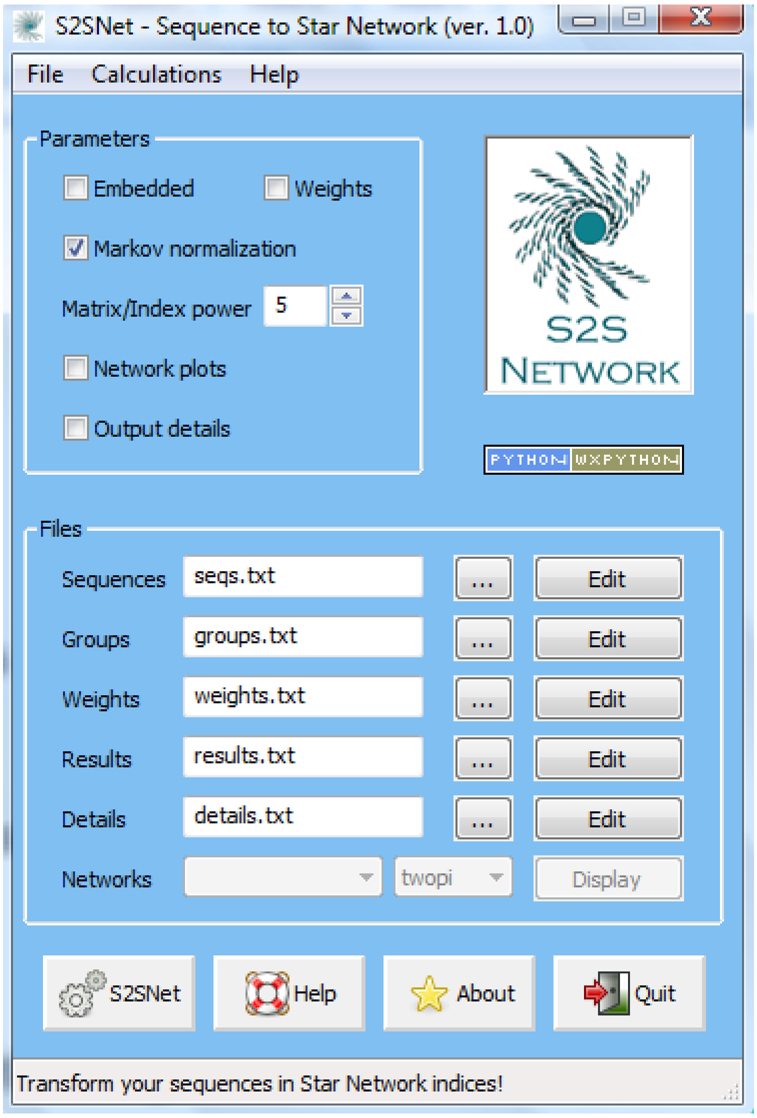
Sequence to Star Network (S2SNet) is a free Python application using wxPython for the GUI and Graphviz as plotting back-end. S2SNet helps you to transform the character sequences/strings into Star Network (SN) topological indices (TIs) and visualize the resulted graphs. These indices are the input data for the statistical analysis. Some examples of sequences are the protein amino acid chains, the DNA/RNA strands or the mass spectra results. This desktop version is working under Microsoft Windows XP/Vista operating system. The application is available for free at http://miaja.tic.udc.es/software/S2SNet/.

What S2SNet can do

- **Transform** 1-character **sequences** in topological **Star Networks indices**;
- **Transform number** data in 1-character sequences (No2Seq);
- **Transform N-character** sequence in 1-character sequence by changing the codification;
- **Edit/View** your input and output **TXT** files;
- Create **DOT** language files;
- Plot and **display networks** as PNG images.

S2SNet description

S2SNet is GUI application with two main panels, the principal window and the console output:



Parameters

□ Embedded □ Weights

☑ Markov normalization

Matrix/Index power 5

□ Network plots

□ Output details

Files

Sequences seqs.txt

Groups groups.txt

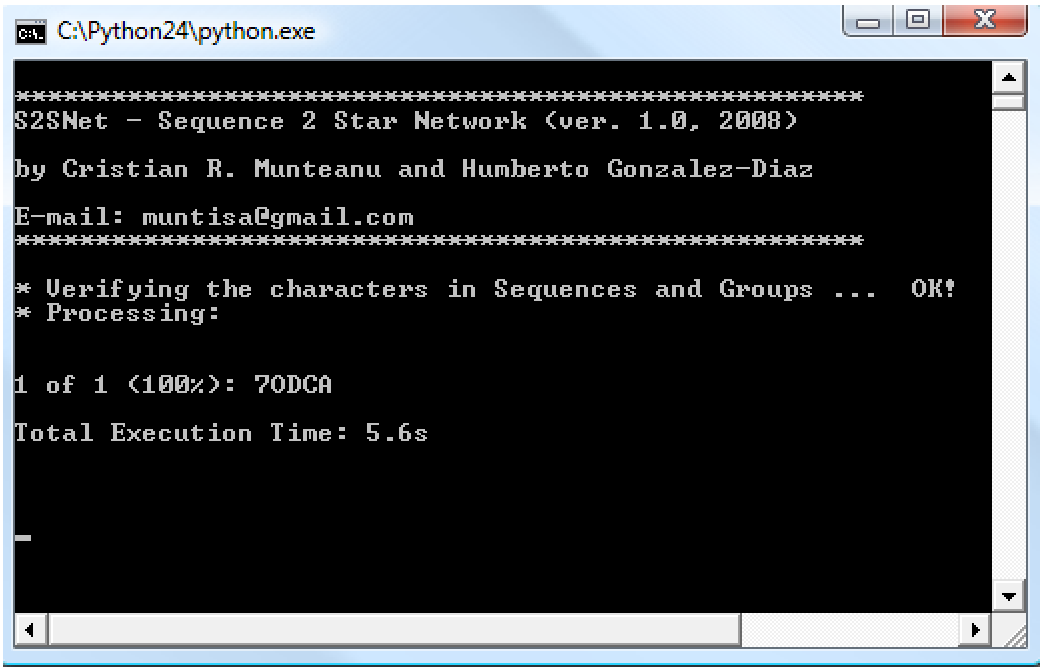
Weights weights.txt

Results results.txt

Details details.txt

Networks twopi

Transform your sequences in Star Network indices!

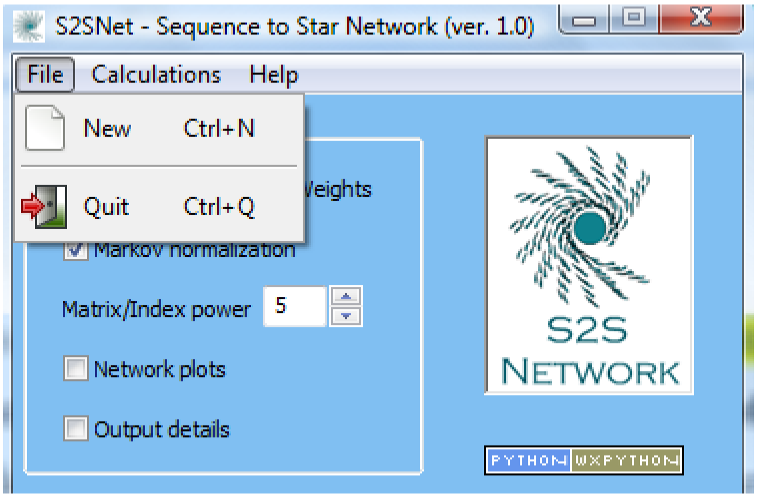


```
*******************************************************
S2SNet - Sequence 2 Star Network (ver. 1.0, 2008)

by Cristian R. Munteanu and Humberto Gonzalez-Diaz

E-mail: muntisa@gmail.com
*******************************************************

* Verifying the characters in Sequences and Groups ...   OK!
* Processing:


1 of 1 (100%): 7ODCA

Total Execution Time: 5.6s



_
```

The main window has buttons for a fast access to the main features of the application:

- **S2SNet -** the transformation of sequences in SN TIs;
- **Help -** a short help page;
- **About -** details about S2SNet and the authors;
- **Quit -** leave the tool.

The same options available in the S2SNet *Menu* divided in **File**, **Calculations** and **Help** too:

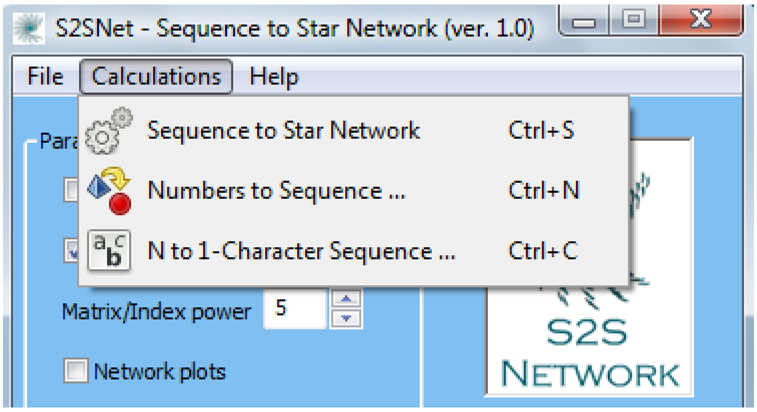


☑ Markov normalization

Matrix/Index power 5

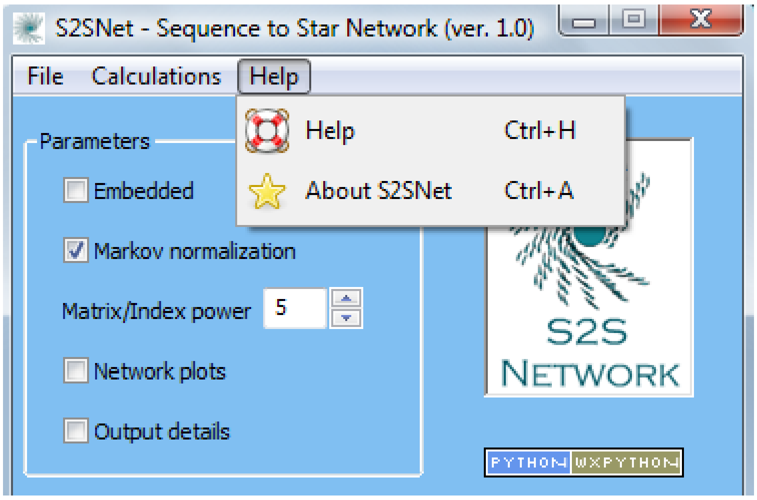
□ Network plots

□ Output details


S2SNet - Sequence to Star Network (ver. 1.0)
File
Calculations
Help
Sequence to Star Network Ctrl+S
Numbers to Sequence ... Ctrl+N
N to 1-Character Sequence ... Ctrl+C
Para
S2S NETWORK

Matrix/Index power 5

□ Network plots



□ Embedded

☑ Markov normalization

Matrix/Index power 5

□ Network plots

□ Output details

In addition, the **File** and **Calculations** menus contain extra options:

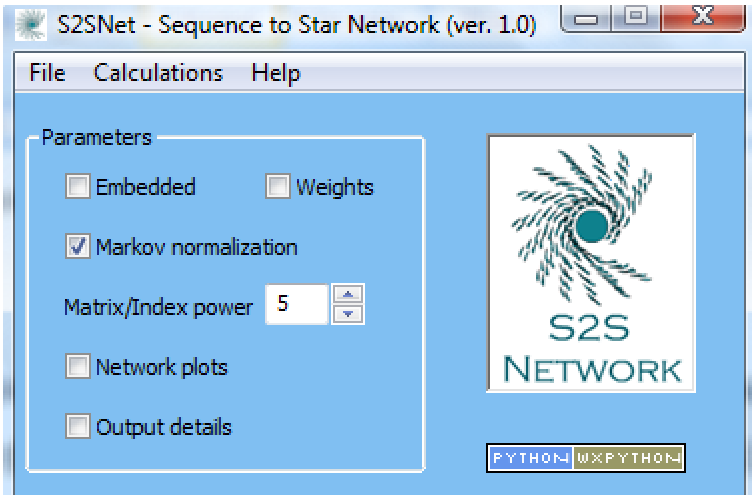
- **New** – create a new Notepad text document;
- **Numbers to Sequence** – transform the numerical input data such as the protein mass spectra in 1-character sequence;
- **N to 1-Character Sequence** – change the sequence codification by transforming N-character groups in 1-character such are the DNA/RNA codon sequences.

In the console panel S2SNet prints details about the stage/errors/results of the calculations.

How to use S2SNet

In the main windows you can choose the TIs calculation parameters, the input/output files and the visualization type of the resulted graphs:

- **Parameters**: *embedded* network, the use of the *weight* for each character; *Markov* normalization of the connectivity matrixes; if you want *details* of the calculations containing all the intermediate matrices and other info; *power* of the connectivity matrices and of some indices (max. 5); networks *plotting* support;



Parameters

☐ Embedded ☐ Weights

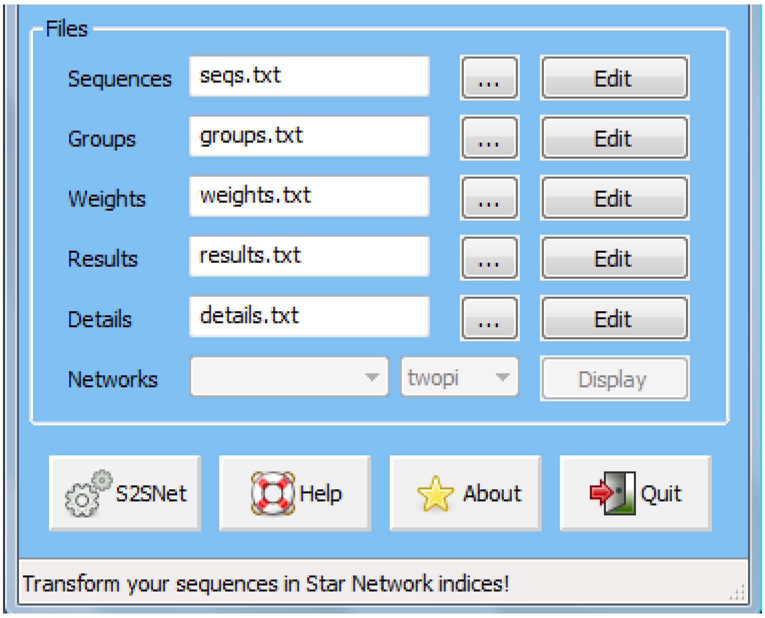
☑ Markov normalization

Matrix/Index power 5

☐ Network plots

☐ Output details

- **Input files**: *sequences*, *groups* and *weights* files;
- **Output files**: *results* and *details* files;
- **Display mode** for the Network plots: the *sequence* to display and the type of drawing *application* (**dot**, **circo**, **twopi**, **neato** and **fdp**; for embedded case use *circo* for better results); extra theoretical graphs are calculated and plotted such as the maximum and the average graphs of the sequences introduced as input.



Files

Sequences: seqs.txt

Groups: groups.txt

Weights: weights.txt

Results: results.txt

Details: details.txt

Networks: twopi

Transform your sequences in Star Network indices!

All the options have default values for a calculation characterized by a non-embedded graph, a Markov normalizations of the matrices, a power of 5 for the matrices, no weights for the nodes, no network plot support and no detail file.

The processing of the sequences can be seen in a **console** window. If you will close it, all the application windows will close too. In the displayed graphs, each group has a different colour. If you need to obtain modified plots, you can find the **DOT** files (one for each sequence) and the **Graphviz** executables (***dot***, ***circo***, ***twopi***, ***neato***, ***fdp***) in the "**dot**" folder (if you enabled the **Network plot** option).

The **Calculations** menu allows you to transform your data in the S2SNet format (1-character string):

- **Numbers to Sequence -** Transforms your numbers in a sequence (the numbers must be TAB separated); you can choose the following:
  - **Parameters**: *minimum* and *maximum* values of your data, *number of groups* you need (maximum 80); you have a GET button if you want to use the minimum and maximum calculated from your entire data;
  - **Input files**: *number* (data) file;
  - **Output files**: *sequence* files, *group* file and *interval* file (description of the group range).

  Note: this can be used to transform the *protein mass spectra* numbers in sequences in order to calculate Star Networks indices.

- **N to 1-Character Sequence -** Transform your N-character sequences in 1-character sequences; you can set the following:
  - **Input files**: *N-character* file (initial file), *code* file (the equivalence between N-character and 1-character; ex: ALA=A);
  - **Output files**: *1-character* files for S2SNet (final file) and *group* file (one item groups).

  Note: this calculation can be used to transform sequences that contain items described by *3-letter codes of the amino acids* into sequences of 1-letter code; another example is the analyse of the *DNA sequences* with the Star Networks (the codon 3-letter nucleotides are transformed in the correspondent/translated amino acid 1-letter codes).

The application is coming with example files for any input and output files. If S2SNet will raise any error, you can see it in the same console and please send us in order to fix it.

S2SNet topological indices

For each sequence S2SNet will calculate the following star network indices:

✓ **Shannon Entropy** of the *n* Markov Matrices (*Sh*):

$$Sh_n = -\sum_i p_i * log(p_i)$$

where p_i are the n_i elements of the vector *p* resulted from the matrix multiplication of the powered Markov normalized matrix (n_i x n_i) and a vector (n_i x 1) with each element equal with 1/n_i;

✓ **Traces** of the *n* connectivity matrices (*Tr*):

$$Tr_n = \sum_i (M^n)_{ii}$$

where *n* = 0 … power, *M* = connectivity matrix (*i***i* dimension); *ii* = *i*th diagonal element;

✓ **Harary** number (*H*):

$$H = \sum_{i<j} \frac{m_{ij}}{d_{ij}} * w_j^{nw}$$

where d_{ij} = elements of the distance matrix, m_{ij} = elements of the *M* connectivity matrix, w_j = weight elements, *nw* = for selection (1) or no selection (0) of weights calculations;

✓ **Wiener** index (*W*):

$$W = \sum_{i<j} d_{ij} * w_j^{nw}$$

✓ **Gutman** topological index (*S6*):

$$S_6 = \sum_{ij} \frac{deg_i * deg_j}{d_{ij}} * w_j^{nw}$$

where deg_i = elements of the degree matrix;

✓ **Schultz** topological index (non-trivial part) (*S*):

$$S = \sum_{i<j} \left(deg_i + deg_j\right) * d_{ij} * w_j^{nw}$$

✓ **Moreau-Broto**, Autocorrelation of Topological Structure (*ATS_n*, *n*=1 … power), only with weights included:

$$ATS_n = \sum_{ij} dp_{ij}^{n} * w_i * w_j$$

where dp^{n}_{ij} = elements of the pair distance matrix when the distance is n;

✓ **Balaban** distance connectivity index (*J*):

$$J = edges \big/ (edges - nodes + 2) * \sum_{i<j} m_{ij} \sqrt{\sum_k d_{ik} * \sum_k d_{kj}} * w_j^{nw}$$

where *nodes*/*edges* = node/edge numbers in the Star Network;

✓ **Kier-Hall** connectivity indices:

$${}^{0}X = \sum_i \frac{w_i^{nw}}{\sqrt{deg_i}}$$

$${}^{2}X = \sum_{i<j<k} \frac{m_{ij} * m_{jk} * w_k^{nw}}{\sqrt{deg_i * deg_j * deg_k}}$$

$${}^{3}X = \sum_{i<j<k<m} \frac{m_{ij} * m_{jk} * m_{km} * w_m^{nw}}{\sqrt{deg_i * deg_j * deg_k * deg_m}}$$

$${}^{4}X = \sum_{i<j<k<m<o} \frac{m_{ij} * m_{jk} * m_{km} * m_{mo} * w_o^{nw}}{\sqrt{deg_i * deg_j * deg_k * deg_m * deg_o}}$$

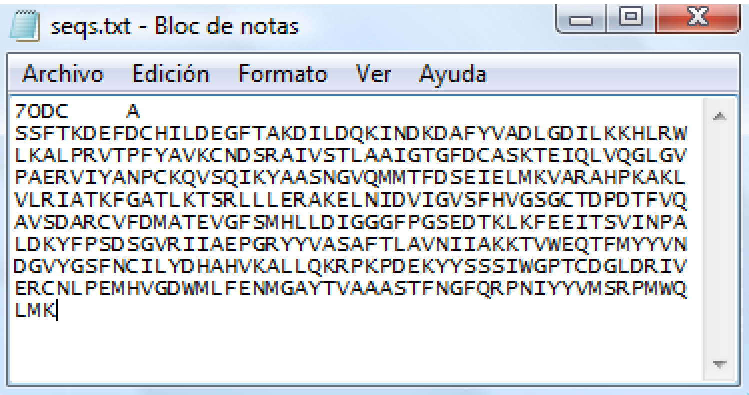
$${}^{5}X = \sum_{i<j<k<m<o<q} \frac{m_{ij} * m_{jk} * m_{km} * m_{mo} * m_{oq} * w_q^{nw}}{\sqrt{deg_i * deg_j * deg_k * deg_m * deg_o * deg_q}}$$

- ✓ **Randic** connectivity index:

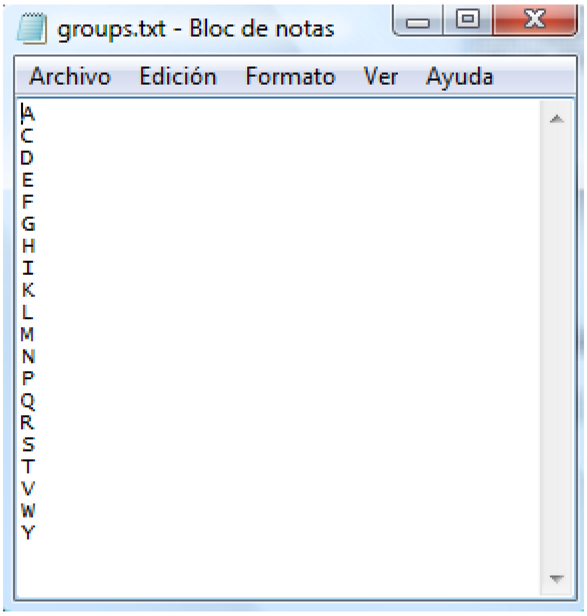
$$ {}^{1}X = \sum_{i<j} \frac{m_{ij} * w_j^{nw}}{\sqrt{deg_i * deg_j}} $$

S2SNet results

The next lines will present the results obtained with S2SNet tool using sequence of the protein chain 7ODCA (from the Protein Data Bank, http://www.rcsb.org/). The sequence file and the group file are the following:

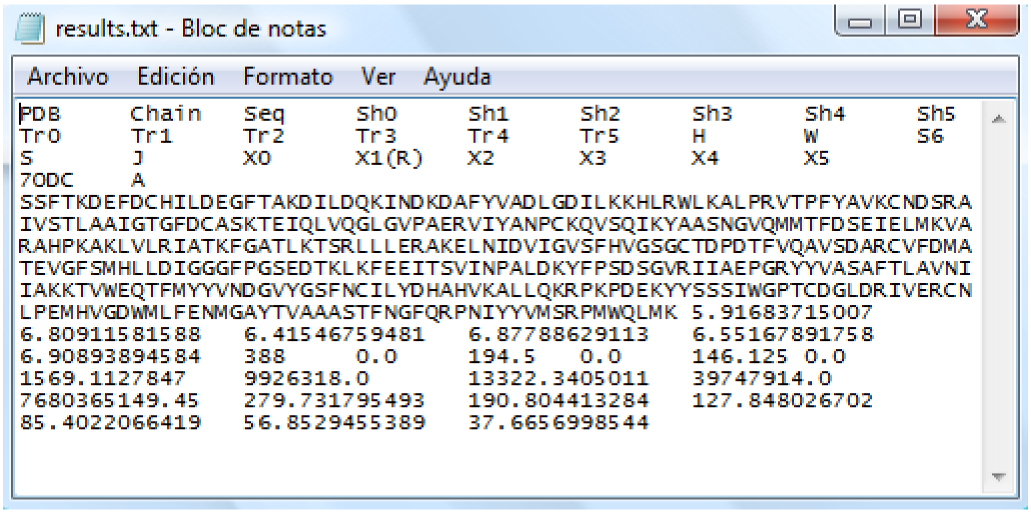


```
7ODC      A
SSFTKDEFDCHILDEGFTAKDILDQKINDKDAFYVADLGDILKKHLRW
LKALPRVTPFYAVKCNDSRAIVSTLAAIGTGFDCASKTEIQLVQGLGV
PAERVIYANPCKQVSQIKYAASNGVQMMTFDSEIELMKVARAHPKAKL
VLRIATKFGATLKTSRLLLERAKELNIDVIGVSFHVGSGCTDPDTFVQ
AVSDARCVFDMATEVGFSMHLLDIGGGFPGSEDTKLKFEEITSVINPA
LDKYFPSDSGVRIIAEPGRYYVASAFTLAVNIIAKKTVWEQTFMYYVN
DGVYGSFNCILYDHAHVKALLQKRPKPDEKYYSSSIWGPTCDGLDRIV
ERCNLPEMHVGDWMLFENMGAYTVAAASTFNGFQRPNIYYVMSRPMWQ
LMK
```

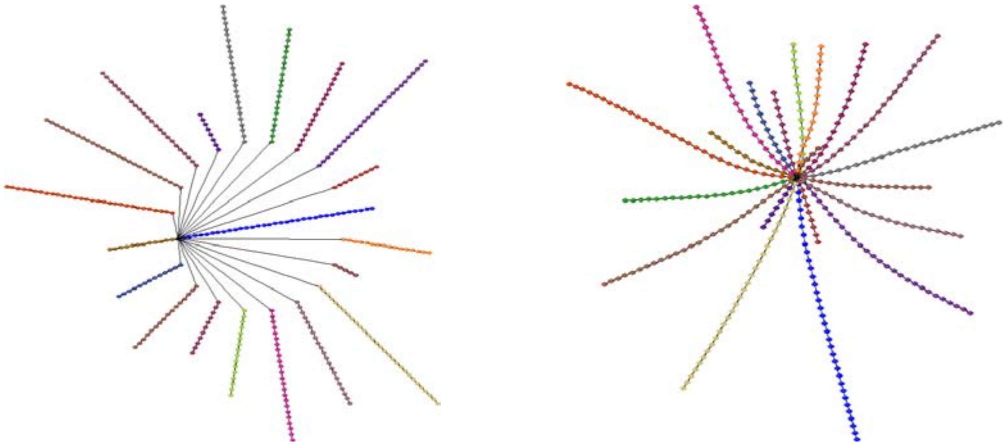


```
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
Y
```

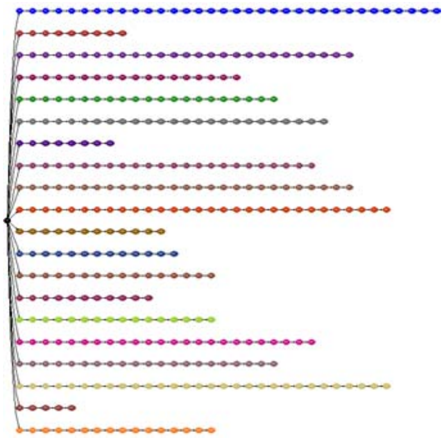
The non-embedded result contains the result files with the star network topological indices and several plots of the graph:



```
PDB        Chain      Seq        Sh0        Sh1        Sh2        Sh3        Sh4        Sh5
Tr0        Tr1        Tr2        Tr3        Tr4        Tr5        H          W          S6
S          J          X0         X1(R)      X2         X3         X4         X5
7ODC       A
SSFTKDEFDCHILDEGFTAKDILDQKINDKDAFYVADLGDILKKHLRWLKALPRVTPFYAVKCNDSRA
IVSTLAAIGTGFDCASKTEIQLVQGLGVPAERVIYANPCKQVSQIKYAASNGVQMMTFDSEIELMKVA
RAHPKAKLVLRIATKFGATLKTSRLLLERAKELNIDVIGVSFHVGSGCTDPDTFVQAVSDARCVFDMA
TEVGFSMHLLDIGGGFPGSEDTKLKFEEITSVINPALDKYFPSDSGVRIIAEPGRYYVASAFTLAVNI
IAKKTVWEQTFMYYVNDGVYGSFNCILYDHAHVKALLQKRPKPDEKYYSSSIWGPTCDGLDRIVERCN
LPEMHVGDWMLFENMGAYTVAAASTFNGFQRPNIYYVMSRPMWQLMK 5.91683715007
6.80911581588    6.41546759481    6.87788629113    6.55167891758
6.90893894584    388       0.0    194.5     0.0    146.125 0.0
1569.1127847     9926318.0        13322.3405011    39747914.0
7680365149.45    279.731795493    190.804413284    127.848026702
85.4022066419    56.8529455389    37.6656998544
```

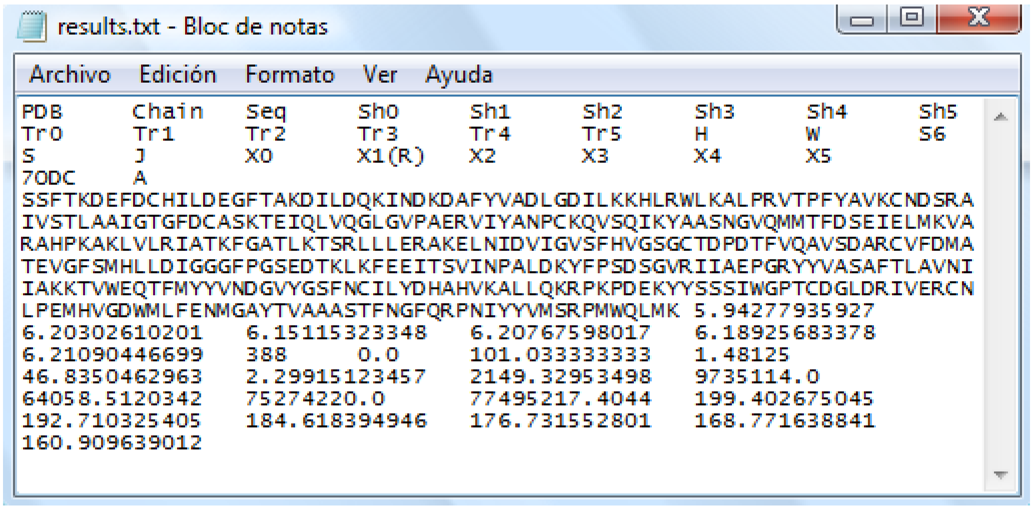


The non-embedded 7ODCA graph made with *twopi* (left) and *neato* (right)

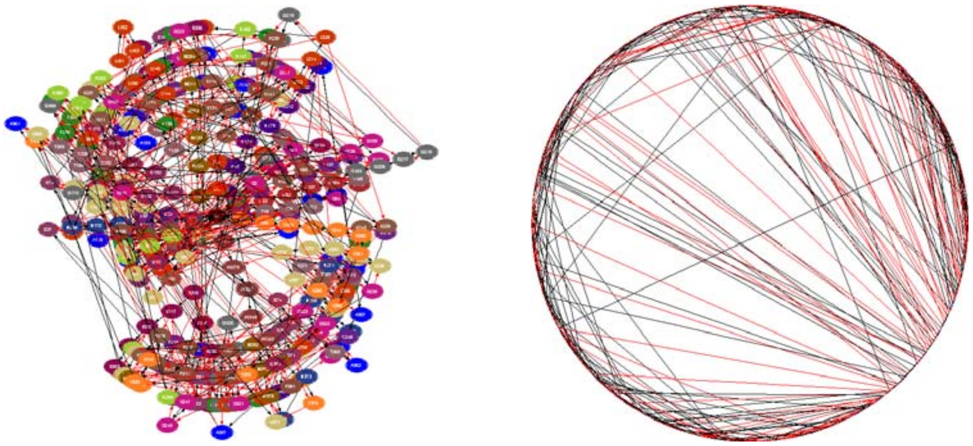


The non-embedded 7ODCA graph made with *dot*

In the case of the embedded graphs, the result files with the star network topological indices and several plots of the graph are different:



```
PDB       Chain     Seq       Sh0       Sh1       Sh2       Sh3       Sh4       Sh5
Tr0       Tr1       Tr2       Tr3       Tr4       Tr5       H         W         S6
S         J         X0        X1(R)     X2        X3        X4        X5
7ODC      A
SSFTKDEFDCHILDEGFTAKDILDQKINDKDAFYVADLGDILKKHLRWLKALPRVTPFYAVKCNDSRA
IVSTLAAIGTGFDCASKTEIQLVQGLGVPAERVIYANPCKQVSQIKYAASNGVQMMTFDSEIELMKVA
RAHPKAKLVLRIATKFGATLKTSRLLLERAKELNIDVIGVSFHVGSGCTDPDTFVQAVSDARCVFDMA
TEVGFSMHLLDIGGGFPGSEDTKLKFEEITSVINPALDKYFPSDSGVRIIAEPGRYYVASAFTLAVNI
IAKKTVWEQTFMYYVNDGVYGSFNCILYDHAHVKALLQKRPKPDEKYYSSSIWGPTCDGLDRIVERCN
LPEMHVGDWMLFENMGAYTVAAASTFNGFQRPNIYYVMSRPMWQLMK 5.94277935927
6.20302610201     6.15115323348     6.20767598017     6.18925683378
6.21090446699     388       0.0     101.033333333     1.48125
46.8350462963     2.29915123457     2149.32953498     9735114.0
64058.5120342     75274220.0        77495217.4044     199.402675045
192.710325405     184.618394946     176.731552801     168.771638841
160.909639012
```



The embedded 7ODCA graph made with *twopi* (left) and *circo* (right)

S2SNet contact

For any question, error or suggestion, please send an email to muntisa@gmail.com. Thank you for using our software!