Analysis and visualization tools for PCHi-C interaction networks

Short description

Software consists of two parts:

- Analysis tools for PCHi-C interaction networks,
- Visualizer of PCHi-C interaction networks.

Analysis tools for PCHi-C interaction networks

Purpose

Software module is designed to provide topological structure analysis of Hi-C interaction networks - obtaining information about all components in the given graph having significance score in the given interval. Results obtained using this software are described in the paper "Topological structure analysis of Hi-C interaction networks". This software serves as a supplementary material for the paper and allows to reconstruct completed experiments using the same data source or provide similar experiments using other data sets. Software package contains source code in C++ (*FindNetworkComponents.cpp*) together with the corresponding data files.

Requirements

To compile source code C++ compiler supporting C++11 standard is required.

Compilation

If g++ compiler is used, we suggest the following command line options (for 64 and 32 bit versions respectively):

```
g++ -m64 -O3 -std=c++11 FindNetworkComponents.cpp -o
<name_of_executable_file>
g++ -m32 -O3 -std=c++11 FindNetworkComponents.cpp -o
<name_of_executable_file>
```

Usage

To run executable, values of the following parameters should be specified:

1) `input_file` - name of the input data file,
2) `output_file` - name of the output text file,
3) `nmax` - the maximal number of vertices in the component,
4) `nmin` - the minimal number of vertices in the component,
5) `a` - the lowest proportion of edges which should be preserved after transformations to define changes as small,
6) `b` - the highest proportion of edges which should be preserved after transformations to define changes as important.

The current version of the program includes some hardcoded parameters (like names and coding of cell types) specific for processing a particular PCHi-C dataset, but the program code can be easily modified to work with other datasets of similar type.

Input data file should contain comma-separated values and contain caption row, like:

```
chr,bait,oe,Mon,Mac0,Mac1,Mac2,Neu,MK,EP,Ery,FoeT,nCD4,tCD4,aCD4,naCD4,nCD8,tCD8,nB,tB
1,840031,896532,0,0,1,1,0,0,1,0,0,0,1,0,0,0,0,0,0
1,840031,915517,1,0,1,1,0,0,1,0,0,0,0,0,0,1,1,0,0
1,849393,896532,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0
1,849393,1209656,0,0,1,1,0,0,1,0,0,0,0,0,0,0,0,0,0
1,862350,840031,0,0,1,1,1,1,1,0,0,0,0,0,0,0,1,0,0
1,862350,1153602,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0
```

In each row there should be exactly 20 values in the following order: number of chromosome (chr, values from 1 to 23), identifier of promoter (bait), identifier of interaction region (other end, oe), 17 values denoting whether this interaction takes place for a particular cell type.

Example of `FindNetworkComponents64.exe` call:

```
FindNetworkComponents64.exe Data/chr1.csv chr1_result_100.txt 100
10 0.75 0.25
```

Output file format

Output file is text file containing description of all found components. Description of each component contains exactly five rows numbered from 1 to 5. Rows according to numbering contains information about:

"1" - number of vertices, edges and significance score,
"2" - cell types used for filtering (if any),
"3" - cell types preserving proportion of edges above `a`,
"4" - cell types leading to the proportion of edges below `b`,
"5" - supplementary information - identifiers of two vertices inside the component.

Example of the output file (first 15 rows):

```
1 10 9 56
2 MK
3 tB nB tCD8 nCD8 naCD4 aCD4 tCD4 nCD4
4 FoeT Ery EP Mac2 Mac1 Mac0 Mon
5 65364341 65522955
1 32 47 54
2 Ery Mon
3 EP MK Neu Mac2 Mac1 Mac0
4 tB nB tCD8 nCD8 naCD4 aCD4 tCD4 nCD4 FoeT
5 44797652 45135428
1 33 62 54
2 Mac2
3 Ery EP MK Mac1 Mac0 Mon
4 tB nB tCD8 nCD8 naCD4 aCD4 tCD4 nCD4 FoeT
5 44797652 45135428
```

Together with the specified output file, also file containing comma-separated values with name encoding parameter values will be produced.

Visualizer of PCHi-C interaction networks

Purpose

Visualizer `tissueComp` is designed to visualize connections between various gene parts for different tissue types for a particular chromosome as a graph diagram using data from the following files:

- file in JSON format containing the description of complete PCHi-C interaction network.
- 23 files of comma separated values containing description of identified components for each of 23 chromosomes.

Setup

Software module `tissueComp.html` together with file `segmTissueBitsGraph17-log-mark.json` and data files `ic-<number>-17-17-100-10-0.750000-0.250000.csv` (where `number=1..23`) should be copied in the directory named `Viz` and at the same level there should be directory `Utils` containing the following `JavaScript` and `CascadedStyleSheet` files:

- `bootstrap.min.css`
- `bootstrap.min.js`
- `d3.min.js`
- `jquery-3.1.1.min.js`
- `underscore-min.js`
- `vis-network.min.css`
- `vis-network.min.js`

File `tissueComp.html` should be executed. The data files may be replaced by another data files in the corresponding formats.

Supported Browsers

`tissueComp.html` is designed and tested for Firefox version 67.0.4.

Main window

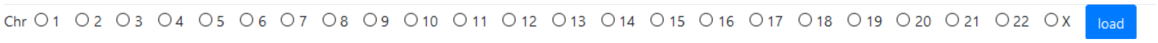

Chr ○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10 ○11 ○12 ○13 ○14 ○15 ○16 ○17 ○18 ○19 ○20 ○21 ○22 ○X
load

Fig. 1. “Choose chromosome” ribbon

In the Fig.1 first tool of the `tissueComp` is depicted: ribbon with radio buttons for chromosome selection. To start visualization, number of chromosome should be chosen (radio button from the ribbon should be checked) and button “load” should be pushed.

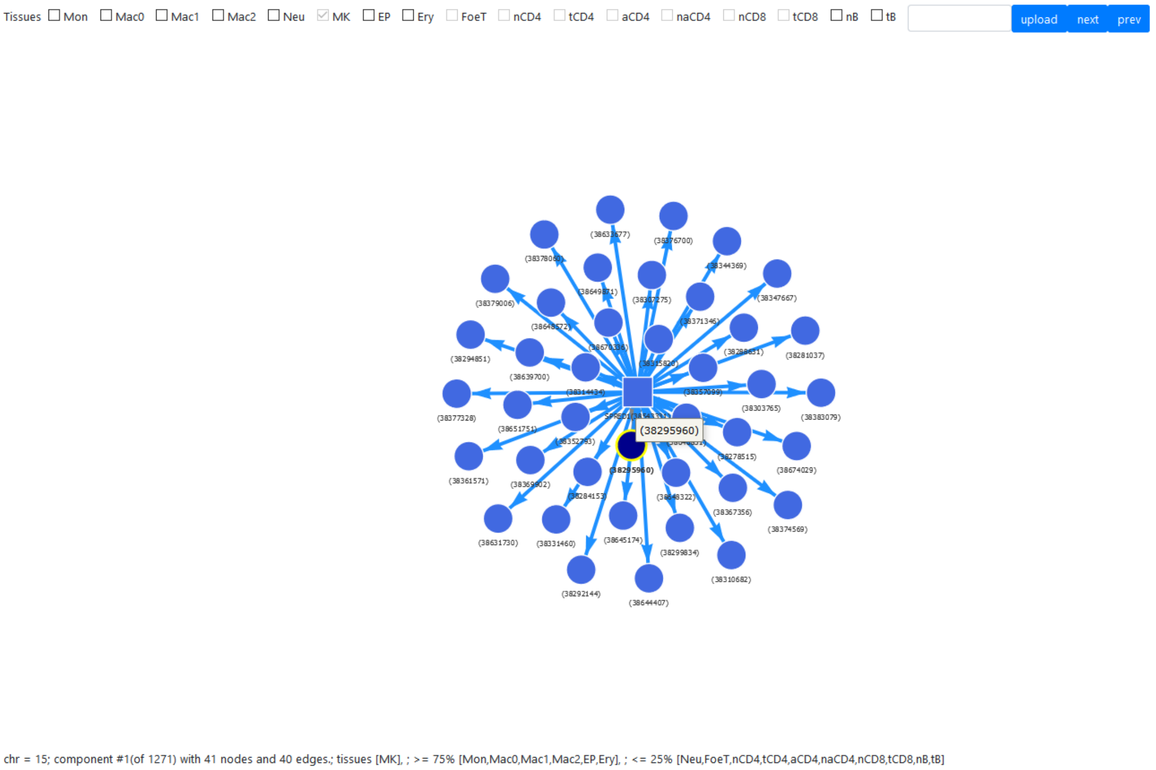
After that graph diagram will appear at the lower part of the screen and ribbon with checkboxes corresponding to the 17 cell types, gene name selection textbox, buttons “upload”, “next”, and “prev” on the top. Example of the obtained diagram is given in Fig. 2.

The diagram of the first component appears at the screen step-by-step in a floating mode - vertices are not in fixed positions and can slowly move around. Process stabilises in few minutes depending on the size of a diagram.

Each checkbox initially can be in one of three states:

- “Disabled/checked” - the corresponding cell type is used for filtering and is present at all edges shown in the diagram;
- “Disabled/unchecked” - the corresponding cell type is not present in the particular component;
- “Enabled/unchecked” - the corresponding cell type is present at least on the one edge in the component.

It is possible to check enabled checkboxes and after pressing “upload”, edges not containing cell type corresponding to the checked checkbox will be shown as dashed arrow in the diagram.



Tissues □ Mon □ Mac0 □ Mac1 □ Mac2 □ Neu ☑ MK □ EP □ Ery □ FoeT □ nCD4 □ tCD4 □ aCD4 □ naCD4 □ nCD8 □ tCD8 □ nB □ tB

chr = 15; component #1(of 1271) with 41 nodes and 40 edges.; tissues [MK], ; >= 75% [Mon,Mac0,Mac1,Mac2,EP,Ery], ; <= 25% [Neu,FoeT,nCD4,tCD4,aCD4,naCD4,nCD8,tCD8,nB,tB]

Fig. 2. Example of the base diagram (chromosome:15, base tissue: MK).

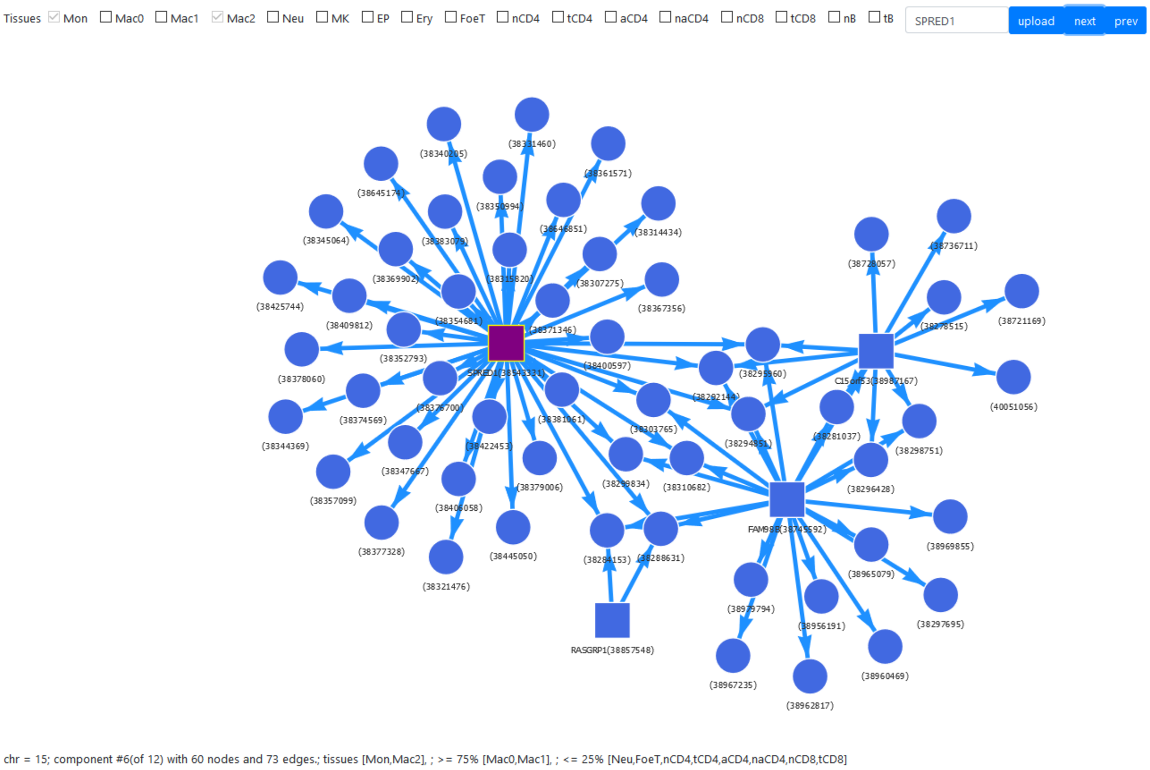
Each vertex and edge can be selected and moved. Also information about the corresponding gene or genome fragment (for vertices), or a represented cell types (for edges) is available as a tooltip after moving above the corresponding diagram element.

At the bottom of the screen the following information about the depicted component is shown as a text line: number of a chromosome, number of the component shown, total number of components, number of nodes and edges in the shown graph, list of the base cell types, list of cell types that preserve at least 75% of edges, list of cell types that preserve at most 25% of edges. Example of the information line corresponding to the Figure 2 is given in the Figure 3.

```
chr = 15; component #1(of 1271) with 41 nodes and 40 edges.; tissues [MK], ; >= 75%
[Mon,Mac0,Mac1,Mac2,EP,Ery], ; <= 25% [Neu,FoeT,nCD4,tCD4,aCD4,naCD4,nCD8,tCD8,nB,tB]
```

Fig. 3. Example of the information line

Further data analysis can be made by specifying name of particular gene. After name is chosen, visualisation should be renewed by pressing button “upload”. Vertex corresponding to the specified gene will be highlighted.



Tissues ☑ Mon ☐ Mac0 ☐ Mac1 ☑ Mac2 ☐ Neu ☐ MK ☐ EP ☐ Ery ☐ FoeT ☐ nCD4 ☐ tCD4 ☐ aCD4 ☐ naCD4 ☐ nCD8 ☐ tCD8 ☐ nB ☐ tB SPRED1

chr = 15; component #6(of 12) with 60 nodes and 73 edges.; tissues [Mon,Mac2], ; >= 75% [Mac0,Mac1], ; <= 25% [Neu,FoeT,nCD4,tCD4,aCD4,naCD4,nCD8,tCD8]

Fig. 3. Example of the updated diagram looking for the gene SPRED1 (component #6).

Navigation between diagrams of components can be done using buttons “next” and “prev”.