SCODE User Manual

Table of Contents

1. Introduction

SCODE is an application designed to implement a supervised training model for protein complex identification in a weighted PPI network. This algorithm was developed by Yanjun Qi, Fernanda Balem, Christos Faloutsos, Judith Klein-Seetharaman, and Ziv Bar-Joseph
and published in the following paper:

Y. Qi, F. Balem, C. Faloutsos, J. Klein-Seetharaman, Z. Bar-Joseph, (2008). Protein Complex Identification by Supervised Graph Clustering , Bioinformatics 2008, 24(13), i250-i268 (The 16th Annual International Conference Intelligent Systems for Molecular Biology (ISMB), July 2008, (Impact Factor 4.328) (acceptance rate of ISMB08: 17% = 49/292)

2. Installation

SCODE is designed and tested for use with Cytoscape version 3.2. Cytoscape is a platform for visualizing biological graphical networks which enables applications and tools for network analysis, annotation, and a host of other features.You must have Cytoscape installed in order to run the application.

Cytoscape requires Java (versions 7 and 8 are compatible with Cytoscape 3.2), which can be downloaded and installed through the following link:

https://www.java.com/en/download/help/download_options.xml

Cytoscape is compatible with Windows XP and newer, Mac OS X 10.7 (Lion), and various distributions of Linux. It can be downloaded from the following link:

http://www.cytoscape.org/download.php

After downloading, run the executable file and follow the installation instructions.

Once you have installed Cytoscape, SCODE may be installed in one of two ways:

1. Use Cytoscape's installation manager.
    a. Open Cytoscape. In the top menu bar, Navigate to **Apps > App Manager**. In the window that appears, Cytoscape will load an alphabetized list of apps available for installation. Select 'SCODE' from this list and click '**Install**'.

2. Install manually using jar file.
    a. Download the SCODE jar file from http://apps.cytoscape.org/apps/scode. Then, locate the folder in which Cytoscape was installed on your computer (/Cytoscape). Navigate to **/apps/installed** and move the jar file into this folder.

3. Startup

After launching Cytoscape, you will be prompted with a window asking you if you would like to start a new session or launch a saved session. A session file (.cys) contains saved work from a network that has previously been modified in Cytoscape. It packages together all of the settings, files, data, and visualizations so that you may continue your work at another time or from a different machine.

You will need to provide a network for analysis with SCODE. From this window, you may choose to either construct a network from scratch, load a network from a file, or you may use Cytoscape's database search button in order to load a network from a database.

Once you have begun a session and loaded a network, you can launch the SCODE app by clicking on '**Apps**' in the top menu bar, followed by **SCODE > Open SCODE.**

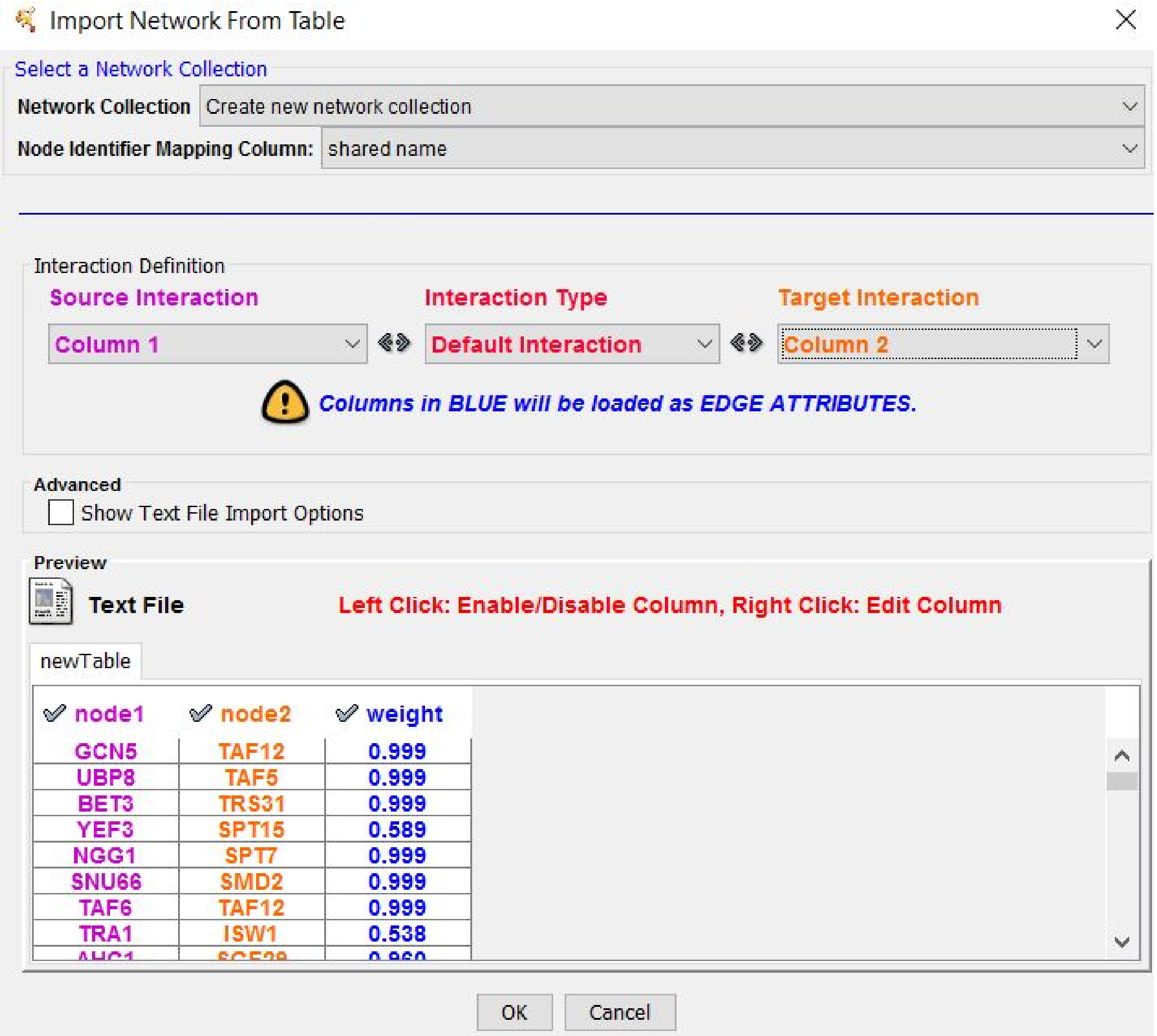
4. Importing a PPI graph to Cytoscape from a file

PPI graphs must be represented as tab-delimited tables to be loaded from a file. At a minimum, the app will require three columns:

- Column 1: Source protein
- Column 2: Target protein
- Column 3: Weight of interaction between the two proteins

Other columns may be included in the input file but will not be used.

To load the PPI graph into Cytoscape, navigate to File > Import > Network > File from the Cytoscape menu bar, and select the graph file. The following window should appear:



Import Network From Table

Select a Network Collection

Network Collection: Create new network collection

Node Identifier Mapping Column: shared name

Interaction Definition

Source Interaction: Column 1

Interaction Type: Default Interaction

Target Interaction: Column 2

Columns in BLUE will be loaded as EDGE ATTRIBUTES.

Advanced

☐ Show Text File Import Options

Preview

Text File

Left Click: Enable/Disable Column, Right Click: Edit Column

newTable

✓ node1	✓ node2	✓ weight
GCN5	TAF12	0.999
UBP8	TAF5	0.999
BET3	TRS31	0.999
YEF3	SPT15	0.589
NGG1	SPT7	0.999
SNU66	SMD2	0.999
TAF6	TAF12	0.999
TRA1	ISW1	0.538

The ‘Source Interaction’ is the first column of proteins in your input file, and ‘Target Interaction’ represents the second set (the each row in your graph file is a directed edge between two nodes). ‘Interaction Type’ may be left as ‘Default’ if your data does not already have a column denoting this information (by default it will be pp, or protein-protein).

Under the preview, the ‘Source Interaction’ and ‘Target Interaction’ columns should now be marked with a check. In addition, click on the weight column in order to load the weights as edge attribute information. It will become highlighted in blue after selection.

Click ‘OK’ to continue.

5. Positive training data

Positive training data is a tab/space delimited file in which each row represents a positive complex exemplar. The columns denote the following:

- Column 1: A numerical identifier for the complex
- Column 2: A name/textual identifier for the complex
- Column 3: A space-separated list of the proteins that are members of the complex. The protein names must match their identifiers in the input protein graph.

6. Syntax for Bayesian model features

The general syntax for a feature (case insensitive) is

Statistic : Feature {args} (bins)

Consider an example: Count: Node (3)
In this example, Count is a statistic applied to the Nodes in the complex, and the feature is discretized into 3 bins. This gives number of nodes in a complex, with 3 possible bins for feature values.

Descretization/binning is based on the range of a feature's training values. So if the model is trained on complexes composed of 3-11 nodes, bin 1 would account for complexes of 3-5 nodes, bin 2 for complexes of 6-8 nodes, and bin 3 for complexes of 9-11 nodes. Statistics are used to transform the values returned by the feature, which are generally calculated per node.

The list of available features is:

- Cluster Coefficient
- Degree
- Degree Correlation
- Density
- Density at Cutoff N (e.g. Density at cutoff 1.2)
- Edge Table Feature (e.g. edge{ColumnName} -- The column must contain numeric values)
- Edge Table Correlation Feature (e.g. edge{ColumnName1, ColumnName2, …} -- The columns must contain numeric values)
- Node
- Node Table Feature (e.g. node{ColumnName} -- the column must contain numeric values)
- Node Table Correlation FEature (e.g. node{ColumnName1, ColumnName2, …} -- the columns must contain numeric values)
- Singular Value
- Topological Coefficient

The list of available statistics is:

- Mean
- Median

- Max
- Variance
- Count
- Ordinals (e.g. 1st, 2nd, 3rd)

7. Creating a custom Bayesian network

Start by creating an empty Cytoscape network via **File > New > Network > Empty Network**. Your graph must contain a node labeled “Root”, which represents the classification of candidate complexes (whether a complex or not a complex). All nodes must be connected by directed edges, and the edges must not form cycles. A node's name will determine the feature that it represents.

To add a node to the network, right click on the white region of the network window view and select **Add > Node**.
To rename the node to a feature, right click on the node and select **Edit > Rename Node**.

Some sample features might include:

Count : Node (3)	The number of nodes, divided among 3 bins
Max : Degree (4)	The maximum degree of a node in the complex, divided among 4 bins

After creating a pair of features, create a directed edge between them by right clicking the source node, then selecting **Add > Edge**.

8. Saving and loading networks from session files

Session files save all of your network collections into a .cys file. This will allow you to save input PPI graphs to be loaded and analyzed later, to save newly created custom bayesian networks, newly trained bayesian networks, or any other networks that are produced.

To do so, click File > Save As, and your networks will be saved into a .cys file in the directory of your choosing.

To load a previously saved session file, click File > Open, and select the .cys file from your computer.

9. Analysis

Bayesian Model Training

Bayesian networks are probabilistic graphical models that make it easy to define relationships between supposed features of complexes. Each node in a Bayesian network represents a feature (e.g. Number of nodes in a complex). Each edge between nodes represents a dependency between features or conditioning of one feature by another (e.g. the complex's density given the number of nodes in the complex). The values of each feature are discretized before training or scoring a candidate complex.

SCODE offers three options for providing a Bayesian network:

1. Train the built-in bayesian model
2. Provide a trained model
3. Train a custom model

Options 1 and 3 require that an additional file of positive training data is provided. The format of this file is explained in Section 5.

A trained model may be loaded into Cytoscape from a session file and used instead of training a new model. You may also elect to train a network that you have created yourself within Cytoscape (see Section 6).

Several advanced parameters may also be set if you are training a model using option 1 or 3:

- Cluster Probability Prior
- Generate Negative Examples
- Ignore Missing Nodes

Search Parameters

After launching SCODE, you will be prompted with a window allowing you to set the parameters for your analysis. SCODE offers three search variants: improved simulated annealing (ISA), modified improved simulated annealing (M-ISA), and greedy improved simulated annealing (Greedy ISA).

ISA: This is the fastest option and will perform the worst. Each round, a candidate is expanded (or not) using a single, random neighboring node

M-ISA: This a slower option that will perform better than ISA. Each round, a candidate is expanded by testing the M highest degree neighboring nodes. The best of these M nodes is used for expansion.

Greedy ISA: This option, the slowest, tests all the neighboring nodes for expansion and selects the best one. This will result in more, larger, higher-scoring candidate complexes.

Once you have selected a search variant, you may choose to use some customizable number of selected nodes as seeds. Selecting seeds allows for reproducibility of the analysis, by allowing certain nodes in the network to be used as starting points.

The following parameters may be customized for the search process:

a. **Number of seeds**: The number of seed nodes on which the search is performed. Seed nodes are selected by greatest degree.
b. **Search Limit**: The maximum number of iterations that the simulated annealing search will take.
c. **Initial Temperature**: Starting temperature of the annealing search. The higher the temperature, the longer the search will take.
d. **Temperature Scaling Factor**: The rate at which the temperature changes at each iteration of the search. The higher the initial temperature, the more likely that the search will continue.
e. **Overlap Limit**: The search will stop for candidate complexes that overlap another candidate be this specified ratio
f. **Minimum Complex Score**: Candidate complexes below this threshold will be discarded at the end of the search.
g. **Minimum Complex Size**: Candidate complexes with fewer than this specified number of nodes will be discarded at the end of the search.