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 worked on. 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. Analysis

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.

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 options to train a new network, use a trained model or use a custom Bayesian network.

a. **Train a new network**: This requires a list of complexes that will be used to train the model. You must choose a column in your training set representing the edge weights. You may also generate a customized number of negative training examples. Positive training data must then be loaded from a file. You may choose to save the resulting Bayesian network to a file.
    i. **Creating a Custom Bayesian Network**: To create a custom Bayesian network, start with an empty Cytoscape Network. Your graph must contain a node labeled "Root", which represents classification of candidate complexes (cluster/non-cluster). All nodes must be connected by directed edges. Those edges must not form cycles.

A node's name will determine the feature that it represents. For example 'Count: Node (3)' represents the 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.

The general syntax for features is [] : {args}. Statistics are used to transform the values returned by a feature, which are generally calculated per node. *Note that this syntax is case insensitive. All features/statistics may be entered as is, unless specific examples are given.*

b. **Use Trained Network**: If 'Train a new network' is left unchecked, you may select a trained network from the drop-down menu.

5. Sample Session

A demo is available for download at
https://drive.google.com/file/d/0B0zcluwHLquUZkpoT3RtdW9vRHM/view?usp=sharing

This demo contains two files: DEMO.cys (a session file containing a pre-created Bayesian Model) and training-tap06.txt (a text file containing the positive training data that will be used to train the network).

First, open Cytoscape. In the welcome window, click the button to '**Open Session File**' and select DEMO.cys.
Once the session file has been read, you'll notice two networks in the 'Control Panel' on the left of the screen

- Default Bayesian Model: the Bayesian model that will be used for training
- mapped gt1.txt contains the protein network that will be trained

In the top menu bar, select **Apps > SCODE > Open SCODE**.
The window that appears is where you will set the search and training parameters.
For the fastest but least-quality training, use ISA. For slower but better-quality complex identification, use M-ISA. For the slowest but best quality complexes, use Greedy ISA.

You may adjust the remaining search parameters as is fitting. For more information about the search parameters, see Section 4a.

Under '**Train Model**', select the checkbox for '**Train New Model'** and **'Use Custom Bayesian Network'**.

- Select the network titled ‘Default Bayesian Network’ to use as your custom Bayesian network
- The edge weight column is labeled ‘weight’

Select the file ‘training-tap06.txt’ to load positive training data.