SCODE User Manual

Table of Contents

1. Introduction
2. Installation
3. Startup
4. Network Analysis
    a. Search for Complexes
    b. Scoring Complexes
    c. Training a Bayesian Template
    d. Syntax for Bayesian Template Features
    e. Creating a Custom Bayesian Template
        i. Load from a .sif File
        ii. Create using Cytoscape
    f. Browse Results
    g. Evaluate Results

1. Introduction

SCODE is an application designed to implement a supervised training model for protein complex identification in a weighted protein-protein interaction (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:

http://www.cs.cmu.edu/~qyj/paper_CMU/ISMB08_qyj.pdf

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 (Supervised Complex Detection) 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 SCODE 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 Cytoscape, 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. 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.

PPI graphs must be represented as tab-delimited tables to be loaded from a file. At a minimum, the app will require two columns (one for each of the proteins in a pairwise interaction). A weight column indicating the strength of the pairwise interactions is suggested, but not required. Other columns may be included in the input file but will not be used.

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. Network Analysis

After launching SCODE, you will be prompted with a control panel containing many options for conducting your analysis. Results from an analysis can be saved by selecting the 'Save Results to File' button after running an operation.

Searching for Complexes

Searches can be conducted on one network at once that can be selected by the user under the "Protein Graph" field. Only graphs already in the session can be selected.

SCODE offers three search variants: improved simulated annealing (ISA), greedy improved simulated annealing (Greedy ISA), and sorted-neighbor improved simulated annealing (Sorted-Neighbor ISA).

1. **ISA**: This is the fastest option, but it tends to yield the lowest scoring complexes of all of the search variants. A random, neighboring node of the candidate complex is chosen and evaluated to determine if it should be added to the complex.
2. **Greedy ISA**: This option tests all of the neighboring nodes for expansion and selects the best one. The Greedy ISA search is relatively slow, but tends to result in more, larger, and higher-scoring candidate complexes.
3. **Sorted-Neighbor ISA**: This search variant is slower than ISA, and may perform more slowly than Greedy ISA depending on the density of the graph. Like Greedy ISA, it tends to produce higher scoring complexes. At each round, all neighboring nodes are sorted according to degree, and the highest degree neighboring nodes (some number specified as an input parameter) are scored. Only the best node is kept in the complex.

Once you have selected a search variant, you may choose to further refine your search. The following parameters may be customized for the search process:

a. **Search Limit**: The input here sets the maximum number of iterations that the simulated annealing search will take.
b. **Initial Temperature**: This sets the starting temperature of the annealing search. The higher the temperature, the longer the search will take.
c. **Temperature Scaling Factor**: The rate at which the temperature changes at each iteration of the search can be set here. The higher the initial temperature, the more likely that the search will continue.
d. **Overlap Limit**: The search will stop for candidate complexes that overlap another candidate be this specified ratio
e. **Minimum Complex Size**: Candidate complexes with fewer than this specified number of nodes will be discarded at the end of the search.

f. **Use Seeds From File:** A file of seed nodes to parse for searching can be selected here.
g. **Number of Random Seeds**: The number of seed nodes on which the search is performed. Seed nodes are selected by greatest degree.
h. **Number of Results to Display**: The maximum number of complexes to return from the search (ranked according to highest score).

Scoring Complexes

At each stage of the search, complexes are scored to determine whether to keep or discard expanded nodes. Complexes may be scored either with or without supervised learning, according to one of the three options below:

1. Score with edge information (no learning)
2. Provide a trained Bayesian model
3. Train a Bayesian template

Without supervised learning (option 1), complexes will be scored by averaging their normalized edge weights. If an edge weight column is not provided in the protein graph, then a default value of 1.0 will be applied for edges (implicitly, a weight of 0 is applied to unconnected nodes).

The program also allows for supervised learning with a Bayesian model. The latter option requires training to be done on the bayesian network before the operation can be executed. A training file of positive protein complexes must be provided (see 'Training a Bayesian Template' for more information).

The following additional scoring parameters may also be set:

a. **Minimum Complex Score**: Candidate complexes below this threshold will be discarded at the end of the search.
b. **Cluster Probability Prior**: Prior probability that a group of proteins forms a cluster (P(C), where C denotes cluster, or 1 - P(NC), where NC denotes not-a-cluster).

Training a Bayesian Template

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 two options for training a Bayesian template:

1. Train the built-in Bayesian template
2. Train a custom template (See the section below on creating a custom template)

These options require that an additional file of positive training data is provided. 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.

Several advanced parameters may also be set if you are training a model:

a. **Generate Negative Examples**: The value provided here specifies the number of negative examples that will be randomly generated.
b. **Ignore Missing Nodes**: Selecting this option will disregard proteins in a positive training example that are not found in a network.

Syntax for Bayesian Template 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 the number of nodes in a complex, with 3 possible bins for feature values.

Discretization/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)

Creating a Custom Bayesian Template

Your template 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.

The following are two strategies for creating a Bayesian template in Cytoscape: loading the network from a .sif file, or creating the network manually within Cytoscape.

Load from a .sif File

Columns of the .sif file must be tab-separated and adhere to the following format:
Column 1: Source node (tail of the directed edge)
Column 2: Interaction type (will not be used, but required for importing the network. "Interaction" will suffice.
Column 3: Target node (head of the directed edge)

To import the network, select **File > Import > Network > File…**
After selecting the file, click **OK** with the default settings in the Import Network popup window.

Create using Cytoscape

Start by creating an empty Cytoscape network via **File > New > Network > Empty Network**.

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** and clicking on the target node.

Browse Results
After running an operation, any complexes found will be added to the "Network" tab of Cytoscape. These complexes will have a score attached to their name, and a visual of the complex can be generated on the right side of the program by right clicking the complex and selecting "Create View."

In the table panel at the bottom of the program, a node table of the proteins within the complex and an edge table depicting the relationships of these proteins within the complex can be viewed as well.

Evaluate Results
SCODE provides an option for the user to have operations evaluated in order to measure the application's performance. After conducting a search on a network, a block containing options on how to evaluate the results appears. There are two parameters in this block:

- **p**: This value specifies the required level of overlap among the proteins in the discovered complex vs. a known complex in order to say that the discovered complex "recovered" it.
- **File of Testing Complexes**: Input a file containing known complexes from the network to compare with the complexes discovered in the results of the search.

Below this block is an option to 'Evaluate Results,' which will calculate and return the recall and precision of the search based on the file inputted for comparison.