# SNF Documentation

Naveen Venayak

April 29, 2015

**Abstract**

This documentation describes the SNF package in the Systemikon workbench. This package is designed to generate synthetic data representing four distinct biological interaction data sets, fuse these into a single graph using SNF, and finally cluster the resulting graph.

# 1 Introduction

Owing to the increasing number of biological databases containing diverse but related datatypes, simultaneous analysis of multiple data types can lead to improved prediction. Canonical methods for analyzing multiple data types was to cluster each type individually, and then compare results. However, this can often lead to conflicting conclusions. Alternative methods can be applied, such as concatenation of all data for analysis, but this tends to lead to a decreased signal to noise ration, which is a significant problem when analyzing biological data. Thus, the SNF method was proposed. Using message-passing theory, SNF aims to iteratively fuse the graphs. From here, multiple clustering methods can be used, including spectral clustering.

# 2 Code Description

## 2.1 Code overview

This code requires the SNF matlab code from (Wang et al., 2014). A functional R server is required to generate synthetic data; however, if synthetic data is not required this is not essential.

The key additional functions are:

- pWithinSearch
  This function generates synthetic and iteratively generates new data sets with a range of pWithin values, to generate multigraphs with reduced connectivity. Then, the SNFpermutation function is called to determine cluster quality. Finally, the result is plotted.

- SNFpermutation
  This function requires a graph input of the following form:

  - an m x n matrix with m nodes and the $m^{th}$ row defined as [geneA geneB iRefScore GEOScore GOScore pathWScore] for the $m^{th}$ edge
  - groundTruth array which indicates the true cluster identity of each gene, each index representing the respective gene number

### 2.2 Synthetic Data Generation

Synthetic data is generated using a common framework implemented by Prof. Boris Steipe and is detailed in the respective code file.

### 2.3 Similarity Network Fusion

### 2.4 Spectral Clustering

## 3 Example

## References

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–7.