KINC Specification

Knowledge Independent Network Construction

Joshua Burns
Stephen Ficklin

November 17, 2015

Contents

1 Introduction

This software is designed to take biological information and transform it into new types of data based off statistical analysis.

2 Abstract Classes

These interface classes describe the overall structure of the kinc program and how implemenation classes work with one another.

The structure of this code will allow for user implemented plugins in the future, since the interaction of the interface classes are clearly defined.

2.1 BioData

This is responsible to storing and representing a certain type of biological data. It is responsible for reading and writing to a file the data it holds, merging two or more files of its type together into one, modifying the data it holds, and providing results to queries about the data it holds.

These classes are also responsible for importing from human readable file formats to their binary file format and exporting back out to human readable formats.

2.2 BioAnalytic

This is responsible for taking in one or more BioData objects and through statistical analysis producing one or more new BioData objects.

3 BioData Classes

3.1 Expression

This is responsible for storing expression sample data for genes.

3.1.1 File Structure

1. Special Identifier that identifies this file as an expression file.
2. User defined name for expression data.
3. History of this data, where it came from.
4. Number of genes, then number of samples per gene.
5. List of all genes.
6. 2 Dimensional list of all gene samples, per gene.

3.2 Correlation

This is responsible for storing correlation data between genes.

3.2.1 File Structure

1. Special Identifier that identifies this file as a correlation file.
2. User defined name for correlation data.
3. History of this data, where it came from.
4. Number of genes, number of correlations per gene pair.
5. List of correlation methods used.
5. List of all genes.
6. Special lists of all gene correlations sorted by their correlation value. Each entry will have a correlation value, then a number representing how many gene correlations are this value, then the list of gene correlations. For each type of correlation used a list is made.
7. Special 2 Dimensional list of all gene correlations, special being a diagonal matrix instead of square.

3.3 Network

This is responsible for storing network data between genes.

3.3.1 File Structure

1. Special Identifier that identifies this file as a network file.
2. User defined name for network data.
3. History of this data, where it came from.
4. Number of genes.
5. List of all genes.
6. Special 2 Dimensional list of all gene edges, special being a diagonal matrix instead of square.

3.4 Annotation

This is responsible for storing additional information for genes.

3.4.1 File Structure

1. Special Identifier that identifies this file as an annotation file.
2. User defined name for annotation data.
4. Number of genes, number of annotations.
5. List of all genes.
6. List of all annotation types.
7. 2 Dimensional list of all gene annotations, per gene.

4 BioAnalytic Classes

4.1 Pearson

This takes an Expression BioData object and produces a Correlation BioData object. It uses the Pearson correlation statistical method for giving correlation values.

4.2 Spearman

This takes an Expression BioData object and produces a Correlation BioData object. It uses the Spearman correlation statistical method for giving correlation values.