Bioinformatics III

Prof. Dr. Volkhard Helms — Saarland University
Daria Gaidar, Markus Hollander, Duy Nguyen, Thorsten Will — Chair for Computational Biology
Summer Semester 2018

Exercise Sheet 2

Due: April 27, 2018 13:15

Submit your solutions on paper, hand-written or printed at the beginning of the lecture or in building E2.1, Room 3.02. Alternatively you may send an email with a single PDF attachment. Always include source code listings. Additionally hand in all source code via mail to thorsten.will@bioinformatik.uni-saarland.de.

2 Scale-free networks and real interaction networks

We continue to evolve the classes from the first assignment. The assignment of this week deals with scale-free networks, characterizing network structure and real data on protein-protein interaction networks.

Exercise 2.1: The scale-free network (45 pts)

First, (a) construct a scale-free network according to the Barabási-Albert model. Then (b) examine the degree distribution of such networks and determine some characteristics in comparison to random networks. Finally, in (c) try to fit the degree distribution to a theoretical distribution.

(a) Implement the algorithm given in the lecture to set up a scale-free network according to the Barabási-Albert model (see Lecture 2, slide 8). Start from the first three connected nodes and add each new node with a given number of links. Connect the new links with increasing preference to nodes that have higher degrees. This **ScaleFreeNetwork**-class should again use the abstract network class that you wrote in the first assignment.
To obtain a much faster implementation and full points, think of a method to map the probabilities to connect to nodes somehow instead of computing them from scratch in each iteration.

(b) Determine the degree distributions for scale-free networks of 10 000 and 100 000 nodes (each with two new links per iteration), respectively, and plot them with double logarithmic axes. A new pre-implemented method in **Tools.py** will help you with that. What are the differences?
Next, compare one of the distributions to the degree distribution of an equally sized random network (play around with the plot-scaling). What are the major differences?

(c) The degree distribution of a scale-free network follows a power law, which has the form

$$P(k) \sim k^{-\gamma}.$$

To simplify the exercise, we assume $P(k) = Ck^{-\gamma}$, with C being a fixed normalization constant to obtain a proper distribution. Try to fit this theoretical distribution to the degree distribution of a random network using the Kolmogorov-Smirnov distance. Follow this guideline:

- Implement **Tools.getScaleFreeDistributionHistogram(gamma, k)** which returns such a simple power law distribution (histogram[i] = math.**pow**(i, −gamma) and normalization afterwards).

- Implement the KS distance in **Tools.simpleKSdist(histogram_a, histogram_b)**: The KS distance of two distributions is the maximal distance between their respective **cumulative** distributions F_i:

$$D = \sup_x |F_1(x) - F_2(x)|$$

  Thus, first build cumulative distributions from the normalized histograms, then find the position where the distributions deviate the most and return this distance.

- Use the KS distance to determine a γ (between 1 and 3, 0.1 steps sufficient) that fits best to the degree distribution of a scale-free network with 10 000 nodes and two new links per iteration. Compare the empirical distribution of the network to the theoretical distribution with optimal γ in a double-log. plot. Comment on the quality of your fit, reason why it may fail and how it could be vastly improved.

Exercise 2.2: Classify real-world network examples (10 Points)

Characterize (with a short explanation) the following examples of networks into the following categories: random, scale-free, hierarchic, and clustered. Are they directed, undirected? Some of the examples might fit into more than one category. If so, explain your choice.

(a) file sharing services such as ...

- Rapidshare, Megaupload, etc.
- Bittorrent
- Dropbox or Google Drive

(b) social networks such as ...

- Twitter or Instagram
- Facebook or LinkedIn

(c) broadcasting networks like ...

- cable television
- satellite television

Exercise 2.3: Real interaction networks (45 pts)

BioGRID ("Biological General Repository for Interaction Datasets") is a protein interaction database which, in version 3.4.159 (March 2018), contains data of 1,548,143 raw protein and genetic interactions from major model organism species compiled from 64,826 publications. The supplement contains this release as a tab-separated file ("`BIOGRID-ALL-3.4.159.tab.txt`"). The format is documented in the beginning of the file, make yourself familiar with that.
In this exercise you implement the class **BioGRIDReader** which should help you to deal with such data.

(a) The class should read the file in its initialization and store the necessary data in a data structure that simplifies your later queries. For every organism found in the file (as **NCBI taxon identifiers**) one should be able to retrieve all interactions as pairs of **official gene symbols** easily.

(b) Implement **getMostAbundantTaxonIDs(n)** and use it to return the **five** organism with the most interactions annotated in BioGRID as well as their respective number of interactions. Argument why the order is not surprising.

(c) How big is the human interaction network and which are the **10** proteins with the highest degree? Take one of them as an example and briefly explain the biology behind the connectivity.

(d) Implement **GenericNetwork**, a network class that imports networks from files. Then implement **writeInteractionFile(taxon_id, filename)** to be able to create organism-specific network files that can be used by the GenericNetwork-class. Build a network for human (taxon 9606), determine and plot the corresponding degree distribution. Discuss if the distribution behaves more like a scale-free or a random network.

Have fun!