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 5

Due: May 25, 2018 13:15

Submit your solutions on paper, hand-written or printed at the *beginning* of the lecture. Alternatively, you can send an email with a single PDF attachment to markus-hollander@web.de. Your submission should include code listings for programming exercises. Additionally, hand in a .zip file with your source code via email.

Motifs and Annotations in PPI–Networks

Exercise 5.1: Cliques and Network Evolution (45 points)

Biological protein–protein–interaction (PPI) networks can change over time. In this exercise you examine how the frequencies of network motifs, in this case cliques, are affected by this evolution. Cliques are sub–networks that are fully connected, meaning every node in a clique has an edge to every other node in the same clique. Since finding all maximal cliques in a network can result in very long runtimes for larger networks, this exercise only considers cliques of size 3, 4 and 5.

(a) **Reading network files:** Implement a function or class that reads files into an undirected network. The files are tab–separated and contain two columns representing the identifiers of two interacting nodes. You will need this for Exercise 5.2 (a) as well.

(b) **Finding cliques:** Implement a function that computes the number of cliques of sizes 3, 4 and 5 in a network. Do not count the cliques of smaller size that are contained in a larger clique. For example, cliques of size 4 contain 4 cliques of size 3 that do not count towards the number of cliques of size 3.

(c) **Evolving networks:** Implement a function that takes a parameter t representing the number of time steps, as well as a network. For each time step, randomly insert or delete one edge in the network. This way the the number of edges remains about constant as the network evolves.

(d) **Cliques in evolving networks:** Read in the rat network and report the number of cliques of size 3, 4 and 5 at the beginning and after letting it evolve for 100 and 1000 time steps. Also plot the number of cliques of size 3, 4 and 5 at the beginning and after each time step as a function of time with $t = 100$. Comment on your results.

(e) **Randomising networks:** Implement a class or function that takes a network with m edges and returns a randomised version of that network. For $2m$ iterations, randomly select two edges $e_1 = (n_1, n_2)$ and $e_2 = (n_3, n_4)$ from the network and rewire them such that the start and end nodes are swapped, resulting in $e_{1'} = (n_1, n_4)$ and $e_{2'} = (n_3, n_2)$. Create a new network instead of overwriting the input network.

What is the goal of randomising networks this way?

(f) **Examining motif enrichment:** Implement a class or function that takes a parameter n and a network and computes if cliques of size 3, 4 and 5 are significantly ($p < 0.05$) enriched in that network:

(1) For each clique size i, compute the number of cliques $c_o(i)$ in the original network.

(2) Use part (f) to obtain n randomised versions of the original network and repeat step (1) to obtain $c_j(i)$ for each randomised network j.

(3) For each clique size i, compute the number of randomised networks $n_r(i)$ in which the number of cliques is at least as high as in the original network: $c_j(i) \geq c_o(i)$.

(4) For each clique size i, compute $p_i = \frac{n_r(i)}{n}$.

Use this function on the rat PPI–network with parameter $n = 100$ and report if certain clique sizes are significantly enriched in the network or not. Do not forget to mention the p-values.

Exercise 5.2: Annotations in Protein–Protein–Interaction Networks (55 points)

In Exercise Sheet 2 you were introduced to BioGRID, which offers information on protein–protein–interactions (PPI) in several organisms. In this exercise you are going to add protein function annotations from the Gene Ontology (GO) to PPI–networks from BioGRID.
On that basis you are going to investigate if interacting proteins are functionally more similar to each other than non–interacting proteins and if certain combinations of annotations are more frequent than expected.

(a) **Adding annotations to PPI–networks:** Unfortunately, the protein–protein–interaction information from BioGRID and the GO annotations do not come form the same source. As a result, the protein identifiers might not match. Fortunately, the GO annotation files contain accession numbers for the protein database UniProtKB.

Implement a class or function that processes the information of a PPI–network–, a UniProtKB– and a GO annotation file as follows:

(1) Constructs a network from the network file, see Exercise 5.1 (a). PPI–networks of several organisms have already been extracted from BioGRID and are provided in the supplement.

(2) The UniProtKB file is tab–separated and contains the UniProtKB identifier of a protein in the column "Entry" and additional names in the column "Gene names". The latter column can contain several alternative names that are separated by whitespace. Construct a mapping of UniProtKB identifiers and alternative names from that file.

(3) The GO annotation file is also tab–separated, apart from the initial header. The relevant columns are

- **Column 0:** Name of the protein or gene database. Skip all entries that are not from UniProtKB.
- **Column 1:** Accession number of the gene or protein in the database.
- **Column 2:** Exactly one alternative name for the gene or protein.
- **Column 4:** GO identifier of the annotation.
- **Column 8:** Indicator whether the annotation belongs to the cellular component (C), molecular function (M) or biological process (P) ontology. Skip all entries that do not belong to the biological process ontology.

Use the mapping constructed in (2) to find the protein(s) in the network that correspond to the protein identifiers/names in each valid annotation entry. Associate the GO annotation ID of the entry with the found protein(s) in the network.

(b) **Generating an overview:** Implement a function that computes the following information for a given annotated PPI–network:

- total number of proteins and interactions in the network
- total number of unique annotations in the network
- total number and percentage of proteins without any annotation

- smallest, average and highest number of annotations per protein
- smallest, average and highest number of associated proteins per annotation

In a table, report your findings for chicken, pig, and human.

(c) **Examining the most/least common annotations:** Implement a function that returns the n most common and n least common GO identifiers in a given annotated network. If there are several GO identifiers that are associated with the same number of proteins, choose the ones with the lower lexicographical order first.

Use the GO identifiers to look up the 5 most common and 5 least common annotations in humans on QuickGO and list your findings, including how often they occur in the network. Explain why those annotations might be the most or least common.

(d) **Investigating annotation enrichment:** The hypergeometric distribution can be used to find out if a given annotation is significantly overrepresented in interacting compared to non–interacting protein pairs. Let

- N be the number of protein pairs, regardless of whether they interact or not
- n be the number of interacting protein pairs
- K_A be the number of protein pairs where both proteins have annotation A
- k_A be the number of interacting protein pairs where both proteins have annotation A

The probability of at least k_A randomly selected interacting protein pairs where both proteins have annotation A is

$$p_A = P(X \geq k_A) = \sum_{i=k_A}^{\min(K,n)} P(X = i) = \sum_{i=k_A}^{\min(K,n)} \frac{\binom{K_A}{i}\binom{N-K_A}{n-i}}{\binom{N}{n}}.$$

Implement a function that computes p_A for every annotation A in a given annotated network and then reports:

- The number and percentage of annotations A with $p_A < 0.05$, $p_A > 0.5$, $p_A > 0.95$
- The n annotations with the smallest p_A and the n annotations with the highest p_A. If there are several annotations with the same p_A, choose the ones that are associated with more proteins first.

Apply this function to the annotated PPI–network of chicken with $n = 5$ and look up the GO annotations corresponding to the reported GO identifiers. List your findings, including the GO identifiers, p–values and annotations.

Comment on your results and also answer the following questions: Are interacting proteins functionally more similar than non–interacting proteins? Was this to be expected? Why (not)?

Hint: Try to avoid re–computing parts of the sum over and over.

(e) **Investigating annotation combinations:** Implement a function that computes if certain annotation combinations occur more frequently than expected. The function should take the combination size k and the number of random distributions r. Additionally, let n be the number of proteins in the network and n_A the number of proteins with annotation A.

(1) Compute the probability of each annotation in the annotated network as $P(A) = \frac{n_A}{n}$.

(2) Generate a list of all annotation combinations of size k that occur in the annotated network. For each combination $c = (A_1, .., A_k)$

   i. compute how often c occurs in the network as n_c

ii. compute the probability expected if the annotations were independent as

$$P_e(c) = P\left(\bigcap_{i=1}^{k} A_i\right) = \prod_{i=1}^{k} P(A_i)$$

iii. generate r random samples of size n using probability $P_e(c)$ and compute the number of random samples n_r in which c occurs at least as often as in the actual network: $n_c(\text{sample}) \geq n_c(\text{actual})$

iv. compute $p_c = \frac{n_r}{r}$

(3) Report:

- The number and percentage of combinations c with $p_c < 0.001$, $p_c < 0.05$, $p_c > 0.5$
- The m combinations with the smallest p_c and the m annotations with the highest p_c. If there are several combinations with the same p_c, choose the ones that occur more frequently in the network first.

Apply this function to the annotated PPI–network of chickens with parameters $k = 2$, $r = 100$ and $m = 3$. Look up the GO annotations corresponding to the reported GO identifiers. List your findings, including the GO identifiers, p–values and annotations. Briefly comment on your results.