Network Biology

Coding Assignment 2

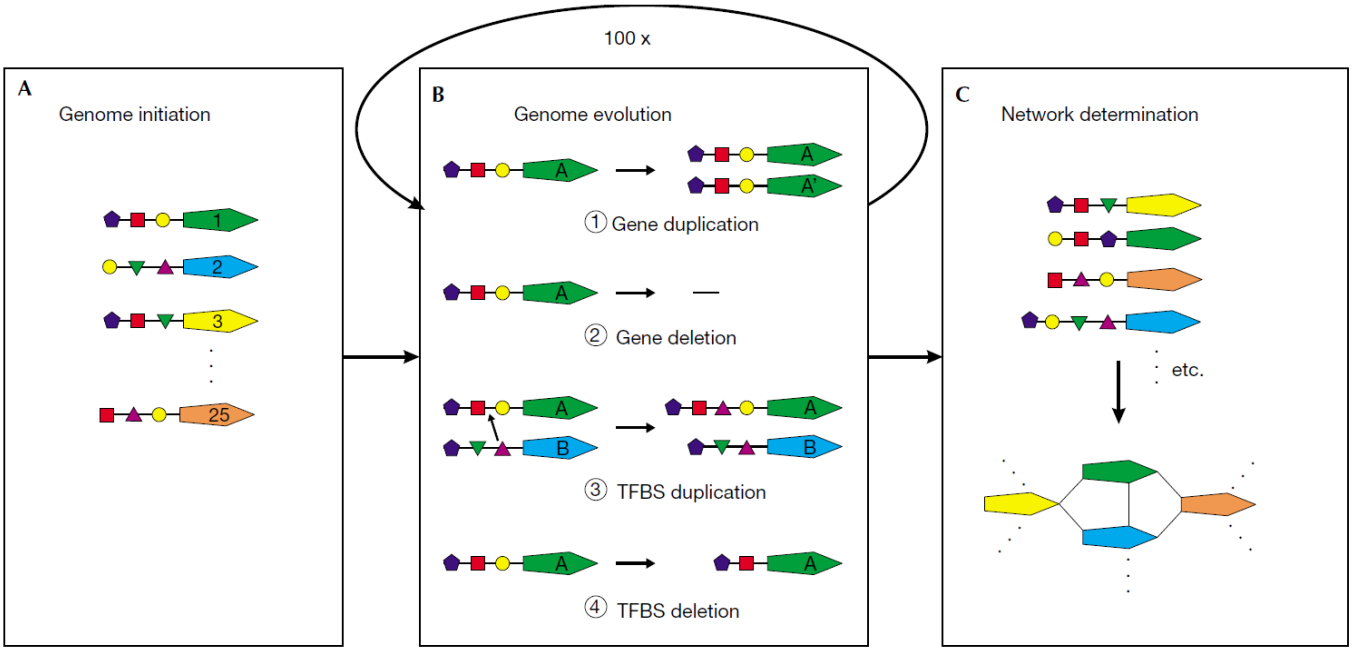
Note: You are responsible for the backup of data as well as results, which will be used for evaluation.

---

1. Implement the "evolutionary model of transcription regulation".
   i. Start with an initial pool of small number of genes (say, 25) and transcription factor binding sites (say, 10) such that each gene has a fixed number of TFBS (say, 3).
   ii. Implement genome evolution involving gene duplication, gene deletion, TFBS duplication, and TFBS deletion. Play with the probability values for each of these events.
   iii. After a certain number of steps of evolution (say, 100), construct the co-expression network with the assumption that 'two genes co-express if they share one or more TFBS'.

   For specific initial values of number of genes and TFBS, probabilities used in the evolution step, and total number of steps in evolution: **[10+10+5]**

   a. Report the **final co-expression network** (image exported from Cytoscape).
   b. Plot the **degree distribution** of the networks for following cases: Gene Duplication ≪ TFBS Duplication, TFBS Duplication ≫ Gene Duplication, and Gene Duplication ≈ TFBS Duplication.
   c. Provide your **conclusions** based on these results.



('The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model', V van Noort, B Snel & M Huynen, *EMBO Rep.* 2004; 5(3): 274–279.)

2. Create Residue Interaction Graph (RIG) and Long-Range Interaction Network (LIN) models for ***any five*** **of the following thirty single-domain two-state folding proteins**, using the strategy listed below: 1hrc, 1imq, 1ycc, 2abd, 2pdd, 1aps, 1cis, 1coa, 1fkb, 1hdn, 1pba, 1ubq, 1urn, 1vik, 2hqi, 2ptl, 2vik, 1aey, 1csp, 1mjc, 1nyf, 1pks, 1shf, 1shg, 1srl, 1ten, 1tit, 1wit, 2ait, 3mef. **[10]**
    - Download the PDB file.
    - Extract the atomic coordinates information (in any programming language, for extracting the coordinates flawlessly).
    - Further, extract the coordinates of Cα atoms of each amino acid.
    - Using the data of representative Cα atoms, **write a code** to create the RIG model using a cut-off of 7.
    - **Write a code** to compute the LIN model, using the threshold of 12 amino acids.

    a. Compute the characteristic path length (L) and clustering coefficient (C) of both RIG and LIN models for each of the above proteins. **Provide your results in a tabular form**.
    b. Write your observations about their topological properties and expected rate of folding.

3. Implement Bartoli's model of protein structure for the five proteins that you chose in response to Question 2. Compute the characteristic path length (L) and clustering coefficient (C) for the RIG model equivalent of Bartoli's models (for 100 instances).

    Following is the procedure for generating contact maps using Bartoli's model:
    (i) Assign 1s to the first two diagonals (up and down the main diagonal) of the adjacency matrix in order to define the backbone contacts.
    (ii) Randomly select a pair of residues *i* and *j* with a probability that decreases linearly with the distance separating these residues in the protein sequence.
    (iii) Assign 1s to the entries of the adjacency matrix corresponding to all nine residue pairs generated by the Cartesian product of $\{i-1, i, i+1\} \times \{j-1, j, j+1\}$.
    (iv) Iterate the last procedure until the number of links in the random graph is *close to* those of the real protein.

    a. **Compare the results** of the RIG and LIN of the original structure to that of their Bartoli counterparts, **and write your observations**. Can you modify step (ii) to create contact maps that are closer to those from real protein structures? Report your results. **[10]**
    b. Compute the plot depicting the 'number of amino acid contacts made' and 'Cartesian distance between them' for all five proteins. Could it be possible to create a model to generate a contact map of a protein, that accounts for this feature? **[5]**