Bioinformatics III

Prof. Dr. Volkhard Helms
Maryam Nazarieh, Duy Nguyen, Thorsten Will
Winter Semester 2016/2017

Saarland University
Chair for Computational Biology

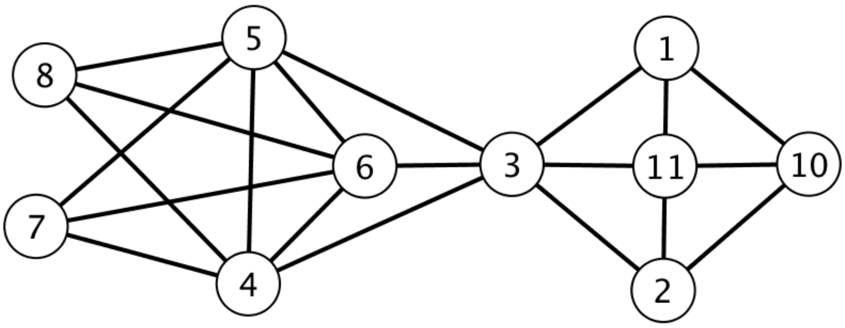
Exercise Sheet 4

Due: November 25, 2016 13:15

Submit your solutions on paper, hand-written or printed at the *beginning* of the lecture or in building E2 1, Room 3.09.

Exercise 4.1: Graph Modular Decomposition (50pts)

A module of a graph $G = (V, E)$ is a set $X \in V$ of vertices where all vertices in X have the same neighbors in $V \backslash X$.



- Every single vertex of V satisfies the definition of a module is a trivial module. List the trivial modules in the graph given above.
- A module is called series if all included nodes are direct neighbors of each other. List all nodes which make series modules in the given graph and specify the series modules.
- A module is called parallel if all included nodes are non-neighbors. List all nodes which make a parallel module and specify the parallel module.
- A module is called prime if it does not fulfill the conditions either of a series or of a parallel module. List all nodes which make a prime module and specify the prime module.
- Draw a modular decomposition tree of this graph formed by prime, series and parallel modules.

Exercise 4.2: Gene Expression Prediction (50pts)

Decision tree is a classification method in which each internal node represents a test on a feature, each branch represents the outcome of the test and each leaf represents a class.
In this problem, we have a set of genes in the test data which are needed to be categorized into two groups of expressed and not expressed genes.

- Use the training data given in the table below to train the decision tree. The binary valued features (0,1) are DNA-methylation and histone modifications (H3K27me3, H3K27ac).

Gene	DNA methylation	H3K27ac	H3K27me3	class
gene1	1	1	1	not-expressed
gene2	0	0	0	not expressed
gene3	0	1	0	expressed
gene 4	0	1	1	not expressed

- Determine for each feature, the best split by minimizing $\frac{N_L}{N}I(N_L) + \frac{N_R}{N}I(N_R)$, where $I(N)$ stands for node impurity of node N, $I(N) = 1 - max\ p_N(k)$ where $p_N(k)$ is the fraction of training points at node N of class k and $k = 1, ..., K$.
- Grow the tree until each leaf is maximal pure.
- Describe the path(s) from the root to the leaf which ends to gene expression.
- Label the class of the test data given below using the trained decision tree.

Gene	DNA methylation	H3K27ac	H3K27me3
gene5	1	0	0
gene 6	0	0	1

Good Luck!