Bioinformatics III

Prof. Dr. Volkhard Helms   Saarland University
Mohamed Hamed, Ruslan Akulenko, Christian Spaniol   Chair for Computational Biology
Winter Semester 2013/2014

Exercise Sheet 9
Due: Jan 29. 2014 10:15

Bayesian approaches for predicting network interactions

Submit your solutions on paper, hand-written or printed at the beginning of the lecture or in building E2 1, Room 3.03. Alternatively you may send an email with a single PDF attachment. If possible, please include source code listings. Additionally hand in all source code via mail to mhamed@ bioinformatik.uni-saarland.de.

Exercise 9.1: Terminologies of Bayes thinking   (30 points)

Discuss shortly (IN YOUR OWN WORDS) the following terms:

1. The difference between the frequency statistics and Bayesian statistics
2. Naïve Bayes classifier and Bayes optimal classifier
3. Bayesian model averaging
4. Max posteriori hypothesis
5. The advantages of Bayesian approaches over other predicting algorithms

Exercise 9.2: Predicting network Interactions using Bayes classifier (70 points)

Apply Naïve Bayes classifier (as discussed in V4) to predict the unknown gene interactions from the given expression data based on the following features between all gene pairs

1-Pearson correlation measure
2-Spearman measure
3-Functional similarity

Steps:

1. Calculate the Pearson and Spearman correlations from the expression data for each gene pair.
2. Merge these measure correspondingly with the given functional similarity scores for all gene pairs
3. Use the labels in the training set file to calculate the required probabilities
4. For the prior knowledge, assume the following:
   - Pearson and Spearman correlation of 500000 pairs (out of 4 Million pairs) is found to be significantly associated with their gene interactions.
   - For Functional similarity: No evidence

Please submit a _**soft**_ copy of your code and the prediction table of the unknown interactions in the following format:

Gene1	Gene2	Interaction
G1	G2	1
G1	G3	0

Where "1" means that there is gene interaction between G1 and G2 and "0" means no interaction.
The unknown interactions are all gene pairs except the ones labeled in the training set.

*Note: Don't multiply probabilities, but rather use the log space.

Good Luck!