

Algorithms in Bioinformatics 2, SoSe2007

Assignment sheet # 7

Christian Rausch

June 4, 2007

1 Resequencing by DNA Chips (2 points)

Research the web to find out about “resequencing using DNA chips”. Give an overview of the technique. For a fully sequenced reference genome, please discuss what kind of additional information one hopes to obtain by applying resequencing technology to the genome of other individuals of the same species? Can resequencing also be useful for investigating the genome of a closely related species?

2 Using the WEKA Package for Gene Expression Analysis (8 points)

WEKA [<http://www.cs.waikato.ac.nz/~ml/weka/>, tutorial: <http://weka.sourceforge.net/wekadoc/index.php/en%3APrimer>] is a comprehensive java toolbench for machine learning and data mining.

2.1 Studies on Cancer Tissues using Microarrays (2 points)

Please, read the two publications that are part of the supplementary data of this assignment sheet. The study by Alon et al. published 1999 in PNAS is on colon cancer, the study by Dhanasekaran et al. published 2001 in Nature is on prostate cancer. Briefly describe their experimental setting and findings. What kind of microarrays were used in the two studies?

2.2 Classification of Microarray Data with SVMs Using the WEKA Package (3 points)

Use at least one SVM implementation in the WEKA package (e.g. LibSVM) to classify the microarray data by Alon et al. and Dhanasekaran et al. that are part of the supplementary data file of this assignment sheet. Train linear SVMs and SVMs with a Gaussian RBF kernel on the training data. Vary the SVM and kernel parameters in a grid search to determine the optimum classification with the best Matthews correlation coefficient MCC on the test data. Recall that the exponential parameter of the RBF kernel is $\gamma_0 = \frac{1}{2\sigma_0^2}$ with $\sigma_0^2 = \text{mean}(\|x_i - x_j\|^2)$, the mean Euclidean distance between all training data points.

2.3 Feature Selection Using the WEKA Package (3 points)

Typically one applies feature selection prior to machine learning on microarray data with the goal to remove all genes (features) from the dataset that are not relevant for the classification problem. Decide for one feature selection algorithm implemented in the WEKA package and briefly state its basic concept. Reduce the features by 10%, 50% and 90% and repeat the classification of Section 2.2. Compare your results with the classification on the unreduced dataset.

Assignments due: **Monday, June 11, 10am**