## Molecular Biomarkers in Clinical Research

## Assignment for lecture on “Bioinformatics and Biomarker Discovery”, 8/9/10

The objective of this assignment is to familiarize you with one of the commonly used tools in data analysis, viz. WEKA. The WEKA machine learning package is used in this assignment. Please download and install it from <http://www.cs.waikato.ac.nz/ml/weka/>.

There are six datasets for this assignment. These datasets are available at the url below. Please download them: <http://www.comp.nus.edu.sg/~wongls/courses/mci10/datasets.zip>.

Q1/ Please describe an approach that you would use to identify attributes in a given training data sets that are important for making accurate prediction in future testing data. [5 marks]

Q2/ For each of the six datasets, please use the approach that you have described in Q1 to identify the three most important attributes for making correct classification of the data. [5 marks]

Q3/ For each of the six datasets, please use the three attributes that you have identified in Q2 as the basis for training a classifier (say the C4.5 decision tree classifier) for the corresponding training dataset. Test the performance of the classifier on the corresponding test dataset. [5 marks]

Q4/ Select one of the classifier that you have built in Q3. Explain how it makes prediction. E.g., if a C4.5 classifier was built, you can show its decision tree and explain one of the rules derived from it. [5 marks]

If you are keen to learn more, you can try this optional question:

Q5/ If the performance in a datasets is not good in Q3, suggest how you would improve it.