MASTER'S THESIS REVIEW

Author	Mukhiddin Yusupov
Title:	Utilization of Methylation Data in Phenotype Molecular Models
Opponent:	Petr Pošík, Ph.D., Dept. of Cybernetics, petr.posik@fel.cvut.cz

The presented thesis deals with an important issue in bioinformatics: how to combine the gene expression (GE) data with other relevant sources of information (here, methylation data) to improve the performance of machine learning models used to detect or predict various types of diseases. Put simply, the thesis shows/confirms that the naive approach of a mere concatenation of both data sets does not bring anything useful, and that a smarter way of combining the individual features based on background knowledge can improve the classification performance. From this point of view, the results obtained by the student are promising and potentially interesting.

A different part of the student work is, however, the written thesis. It has an acceptable length of 45 pages. However, its quality is mediocre at the best. The whole thesis makes an impression that it was written in a hurry. The text contains many grammatical mistakes and typos, which decrease the readability of the text. Although the application area is quite complex, I must say that neither the relevant biological background, nor the state of the art are explained/described clearly and comprehensibly. Long parts of the thesis are written using unstructured text only, without any accompanying figures, schemata, or any other aid that would help the reader understand. Some terms were used without definition (site, region, area - is it the same as region?, etc.). The whole text is rather vague, without precise definitions or descriptions of the applied analytical methods.

The description of the results in Sec. 5.2 also contains many deficiencies. E.g. in the description of methylation data preparation, hierarchical clustering is used, but I miss any specification of the distance metric used, how the similarity between 2 clusters was computed, what data were actually clustered (how many features they had), or whether the data were preprocessed somehow before clustering. In Fig. 5.2, the results of principal components analysis (PCA) and multidimensional scaling (MDS) are presented and the student comments on that: "The PCA and MDS plots tell us about the same amount of information." Well, of course: in the setting used by the student, they shall give us exactly the same results, as can be easilly seen from Fig. 5.2, where the right plot is just the left plot rotated by 180 degrees. In figures 5.5, 5.6 and 5.7, there is also an inconsistence in the plots (see the question below). Generally, the discussion of the results is very weak; it basically only says what can be seen in the graphs and tables, but I often miss any interpretation of the results.

Despite all my complaints, I must state that the thesis fulfilled all the requirements listed in the diploma thesis assignment. The only exception is maybe the item nr. 7: the student should have implemented the methods as a functional plugin of the miXGENE system. This task is addressed only by the very last sentence of the thesis, where the student states that the method is implemented in such a way that allows for implementation in miXGENE, but the actual plugin is missing.

My final evaluation of the master's thesis is

D - satisfactory.

Questions for the student:

1. In figures 5.5, 5.6 and 5.7, we can find box plots and density plots. Both have on one of their axes a variable called "log2 CpG site intensity". However, in the box plots, this variable has values between 10 and 30, while in the density plots it has values between 0 and 40 thousands. Could you explain this this?

In Prague, May 28, 2015

Petr Pošík, Ph.D., opponent