BMC Bioinformatics

Oral presentation

Semi-supervised class discovery using quantitative phenotypes – CVD as a case study

Israel Steinfeld[1], Roy Navon*[1], Diego Ardigò[2], Ivana Zavaroni[2] and Zohar Yakhini[1]

Address: [1]Agilent Laboratories, Tel Aviv, Israel and [2]Departments of Internal Medicine and Biomedical Sciences, University of Parma, Italy

Email: Roy Navon* - roynavon@post.tau.ac.il

* Corresponding author

from Third International Society for Computational Biology (ISCB) Student Council Symposium at the Fifteenth Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)
Vienna, Austria. 21 July 2007

Published: 20 November 2007

BMC Bioinformatics 2007, **8**(Suppl 8):S6 doi:10.1186/1471-2105-8-S8-S6

This abstract is available from: http://www.biomedcentral.com/1471-2105/8/S8/S6

Background

Genomic studies typically focus on comparing disease to healthy population. In our work, various parameters, including peripheral blood mononuclear (PBM) cells expression profiling, were stratified solely from healthy subjects. To analyze the data we developed a semi-supervised class discovery method, constraining the search space to patterns that respect an order induced by the rich quantitative annotations. We show that our method is robust enough to detect known clinical parameters with accordance to expected values. We also use our method to elucidate cardiovascular disease (CVD) putative risk factors.

Methods

One of the basic tasks in gene expression data analysis is finding differentially expressed genes between 2 classes (such as tumor vs. normal). Among the various methods for measuring differential expression (e.g. Student t-test), we focus on TnoM [1] which is a non-parametric statistical score that affords an exact p-value.

When many partitions of the sample set are possible, one would like to assess the statistical significance of any partition considered, and to compare between partitions. In overabundance [2] analysis the exact p-value of the TNoM score is used to estimate the expected number of differentially expressed genes. By comparing to the actually observed number we can calculate the overabundance of differentially expressed genes. This quantity can be used as a figure of merit: higher overabundance indicating a more profound change in the cell state.

Typical class discovery in gene expression data searches over all possible partitions of the set of samples and uses heuristic methods to do so [2,3]. Given any quantitative phenotype, we can constrain the search space to patterns that respect the order it induces on the set of samples. This approach reduces the search space from $O(3^n)$ to $O(n^2)$ making the search feasible (Figure 1).

Results

We applied our method to PBMC gene expression profiling data, collected from 49 healthy subjects. Clinical, laboratory measurement and CVD prognostic indicators were also collected, adding more then 160 phenotypic parameters for each subject. One of the interesting phenotypic parameters is Carotid Intima-Media Thickness (IMT) [4], a CVD prognostic indicator. Using semi-supervised class discovery with the IMT values we received IMT threshold levels that are in agreement with the known prognosis values (Figure 1). The differentially expressed genes in this partition were enriched with GO terms related to vesicle-mediated transport ($p < 10^{-8}$) and glycolysis ($p < 10^{-6}$), giving mechanistic insights to the difference between the two cell states.

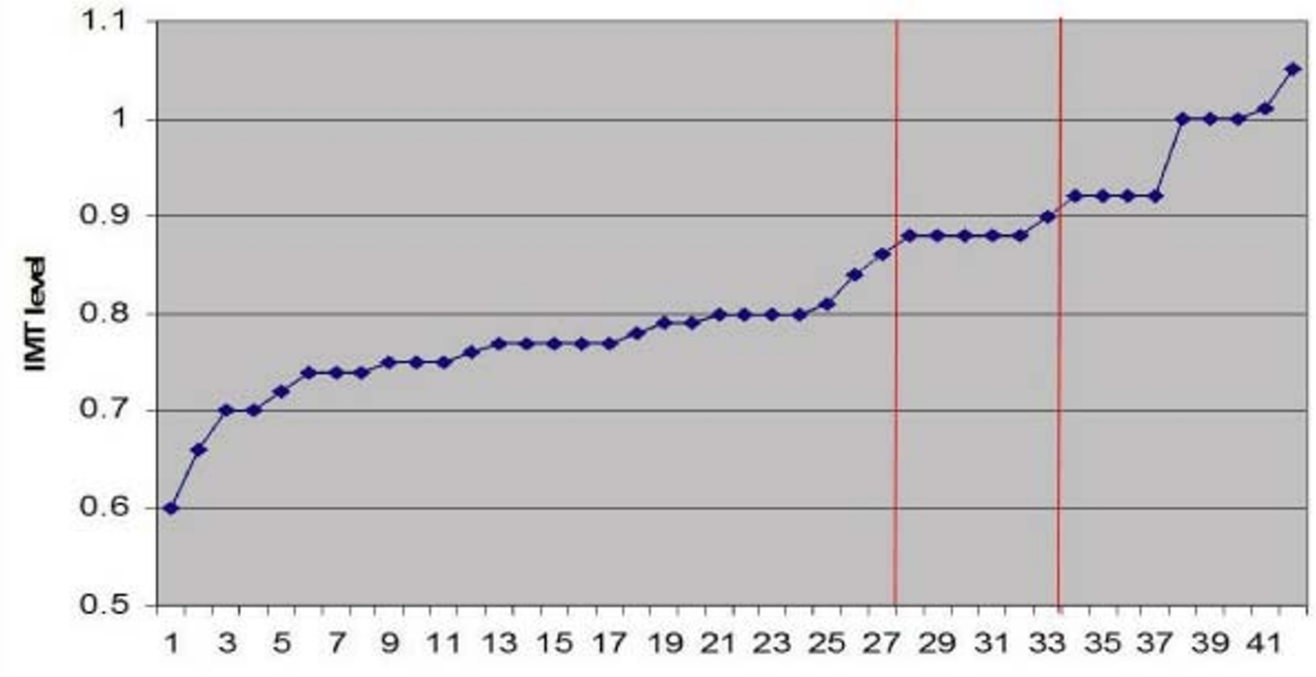

Figure 1
IMT levels available for 42 subjects are presented. All threshold pairs of IMT levels were tested, each representing a partition of the samples to high IMT levels, low IMT levels and samples not used. Marked in red is the threshold pair with the highest overabundance of genes, giving rise to the partition of 27 samples with IMT values of 0.6–0.86 vs. 9 samples with IMT values of 0.92–1.05.

Acknowledgements

This study is supported by a EU FP6 grant in the framework of the Multi Knowledge Project. We thank Eran Eden, Doron Lipson and Benny Chor for useful discussions.

References

1. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z: **Tissue classification with gene expression profiles.** *J Comput Biol* 2000, **7(3–4):**559-83.
2. Ben-Dor A, Friedman N, Yakhini Z: **Overabundance Analysis and Class Discovery in Gene Expression Data.** *Agilent Technical Report* 2002. AGL-2002–4
3. von Heydebreck A, Huber W, Poustka A, Vingron M: **Identifying splits with clear separation: A new class discovery method for gene expression data.** *Bioinformatics* 2001, **17:**S107-S114.
4. Lorenz MW, Markus HS, Bots ML, Rosvall M, Sitzer M: **Prediction of clinical cardiovascular events with carotid intima-media thickness: a systematic review and meta-analysis.** *Circulation* **115(4):**459-67. 2007, Jan 30; Epub 2007 Jan 22