Disease Classification Issues in Cancer Genomics

Harri Lähdesmäki[1,], Nathan Price[1,*], Seungchan Kim[4], Cristian Mircean[6], Tohru Kobayashi[5], Greg Fuller[2], Jeff Trent[4], Ioan Tabus[6], Edward R. Dougherty[3], Wei Zhang[2] and Ilya Shmulevich[1]*

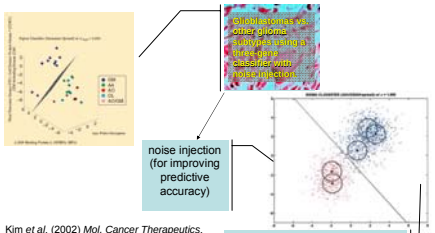
[1]Institute for Systems Biology, Seattle, WA

A collaboration between ISB and MDACC

classification in cancer genomics

- collect measurements (e.g. gene expression)
- determine which of those (e.g. genes) can be used as features
- apply a classification rule to construct a classifier
- apply an estimation rule to estimate the classifier error

multivariate feature selection



Kim *et al.* (2002) *Mol. Cancer Therapeutics*, 1(13):1229–1236.

Kobayashi *et al.* (2003) *Cancer Research* 63: 60-66.

an important issue: *sample heterogeneity*

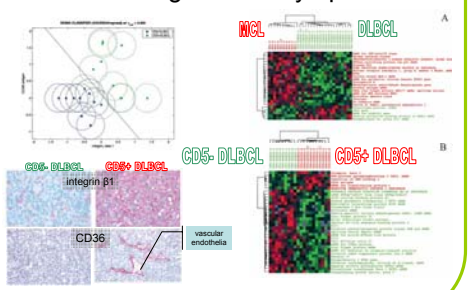
- tissue samples used in cancer studies are usually contaminated with the surrounding or infiltrating cell types.
- this sample heterogeneity hinders further statistical analysis, significantly so if different samples contain different proportions of these cell types.
- thus, sample heterogeneity can result in the identification of differentially expressed genes that may be unrelated to the biological question being studied.

in silico microdissection

- the computational method provides estimates of the expression values of the pure (non-heterogeneous) cell samples.
- the inversion of the sample heterogeneity can be facilitated by providing accurate estimates of the mixing percentages of different cell types in each measurement.
- for those cases where no such information is available, an optimization-based method for joint estimation of the mixing percentages and the expression values of the pure cell samples is used.

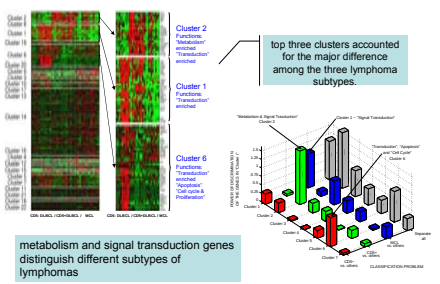
Lähdesmäki *et al.* (2005) *BMC Bioinformatics*. 6(1):54.

CD5+ and CD5- Diffuse Large B-Cell Lymphomas



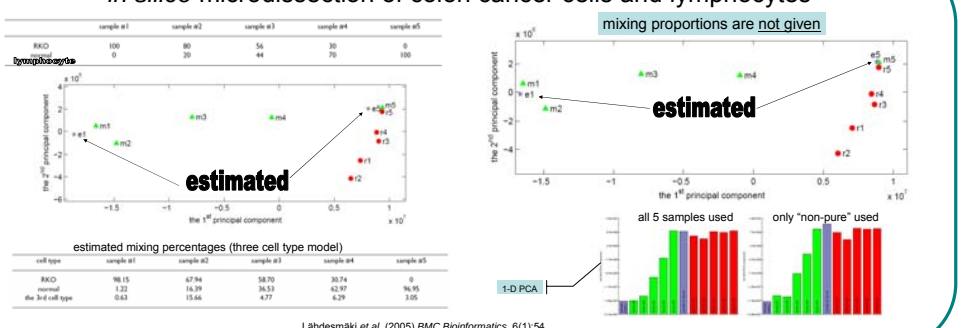
Kobayashi *et al.* (2003) *Cancer Research* 63:60-66.

gene shaving (22 clusters)



Mircean *et al.* (2004) *Int J Oncol*. 24(3):497-504.

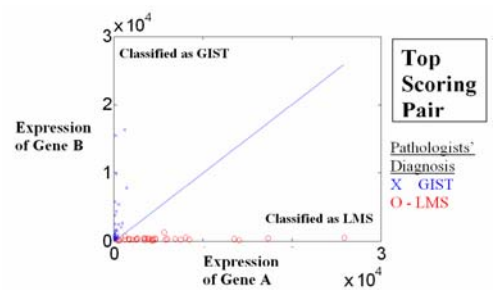
in silico microdissection of colon cancer cells and lymphocytes



Lähdesmäki *et al.* (2005) *BMC Bioinformatics*. 6(1):54.

Top Scoring Pair (TSP) Classification of Tumors

- Two similar types of tumors, Gastrointestinal Stromal Tumors (GIST) and Leiomyo-sarcomas (LMS) were until fairly recently thought to be the same type of cancer.
- We used a classification approach known as Top Scoring Pair (TSP) (Tan *et al*, Bioinformatics, 2005. 21(20), 3896-3904). It performs perfectly on the data, and its accuracy on future data sets was estimated using leave-one-out cross validation. (LOOCV) at 99%.
- A Simple Classification Rule: If Expression of Gene A > Expression of Gene B, then the tumor is classified as an LMS, else it is classified as a GIST.



Estimated Accuracy of this 2-gene classifier (LOOCV) was 99%.

acknowledgments

- Wei Zhang, Greg Fuller – [2]University of Texas MD Anderson Cancer Center, Houston, TX, USA
- Edward R. Dougherty – [3]Texas A&M University, College Station, TX, USA
- Seungchan Kim, Jeff Trent – [4]Translational Genomics Research Institute, Phoenix, AZ, USA
- Tohru Kobayashi – [5]Mie University School of Medicine, Tsu, Japan
- Cristian Mircean, Ioan Tabus – [6]Institute for Signal Processing, Tampere University of Technology, Finland