Machine learning approaches for deciphering complex pathomechanisms in cancer

Roland Eils
Division "Intelligent Bioinformatics Systems"
German Cancer Research Center, Heidelberg

Abstract: Recent years have seen a dramatic increase in the amount of genetic information stored in electronic format. It has been estimated that the amount of information in genomics and proteomics doubles every 20 months
and the size and number of databases are increasing even faster. It is widely accepted that a sophisticated exploration of such data is crucial in a variety of fields such as disease genetics and pharmacogenomics. While both corporate and institutional efforts have concentrated on the integration of heterogeneous data in genomics and proteomics, a systematic data exploration is still at its beginning. Although data mining has celebrated many successes in business operations applications as retail and marketing (see e.g. [1]),
its application to scientific and engineering data is not straightforward.

Data sets in life sciences are often significantly larger in volume, structurally more complex then traditional business data, and often rapidly changing in time.

In contrast to business environments, the body of existing background knowledge in life sciences is extensive.

I will report on our recent efforts [2,3] to adapt data mining technology in particular from the field of machine learning for effective knowledge discovery in tumour genetics. To exemplify the power of this approach, we
will describe how complex concepts such as survival or therapy response [4,5] can be learnt from heterogeneous clinical or molecular data.

References:

1. Ragg T., W. Menzel, W. Baum, M. Wigbers (2002) Bayesian learning for sales rate prediction for thousands of retailers. Neurocomputing 43, 127-144.
2. Dubitzky et al. (2000) A Comparison of Symbolic and Subsymbolic Machine Learning Approaches to Molecular Classification of Cancer and Gene Identification. In: Proceedings of "Critical Assessment of Techniques for Microarray Data Analysis (CAMDA'00 Conference)", Duke University, Durham, NC, USA, 2000.
3. Granzow et al. (2001) Tumor Identification by Gene Expression Profiles: A Comparison of Five Clustering Methods. ACM SIGBIO Newslett., 21(1): 16-22
4. Berrar et al. (2001b) New Insights in Clinical Impact of Molecular Genetic Data by Knowledge-driven Data Mining. In: Proceedings of the Second International Conference on Systems Biology. Omnipress, Wisconsin, USA. pp. 275-281.
5. Berrar et al. (2001c) Analysis of Gene Expression and Drug Activity Data by Knowledge-based Association Mining. In Proceedings of "Critical Assessment of Techniques for Microarray Data Analysis (CAMDA'01 Conference)", Duke University, Durham, NC, USA, 2001.