Analysis of large-scale molecular biological data using self-organizing maps

Dipl.-Inf. Henry Wirth

Modern high-throughput technologies such as microarrays, next generation sequencing and mass spectrometry provide huge amounts of data per measurement and challenge traditional analyses. New strategies of data processing, visualization and functional analysis are inevitable. This thesis presents an approach which applies a machine learning technique known as self organizing maps (SOMs). SOMs enable the parallel sample- and feature-centered view of molecular phenotypes combined with strong visualization and second-level analysis capabilities.

We developed a comprehensive analysis and visualization pipeline based on SOMs. The unsupervised SOM mapping projects the initially high number of features, such as gene expression profiles, to meta-feature clusters of similar and hence potentially co-regulated single features. This reduction of dimension is attained by the re-weighting of primary information and does not entail a loss of primary information in contrast to simple filtering approaches. The meta-data provided by the SOM algorithm is visualized in terms of intuitive mosaic portraits. Sample-specific and common properties shared between samples emerge as a handful of localized spots in the portraits collecting groups of co-regulated and co-expressed meta-features. This characteristic color patterns reflect the data landscape of each sample and promote immediate identification of (meta-)features of interest. It will be demonstrated that SOM portraits transform large and heterogeneous sets of molecular biological data into an atlas of sample-specific texture maps which can be directly compared in terms of similarities and dissimilarities. Spot-clusters of correlated meta-features can be extracted from the SOM portraits in a subsequent step of aggregation. This spot-clustering effectively enables reduction of the dimensionality of the data in two subsequent steps towards a handful of signature modules in an unsupervised fashion.

Furthermore we demonstrate that analysis techniques provide enhanced resolution if applied to the meta-features. The improved discrimination power of meta-features in downstream analyses such as hierarchical clustering, independent component analysis or pairwise correlation analysis is ascribed to essentially two facts: Firstly, the set of meta-features better represents the diversity of patterns and modes inherent in the data and secondly, it also possesses the better signal-to-noise characteristics as a comparable collection of single features.

Additionally to the pattern-driven feature selection in the SOM portraits, we apply statistical measures to detect significantly differential features between sample classes. Implementation of scoring measurements supplements the basal SOM algorithm. Further, two variants of functional enrichment analyses are introduced which link sample specific patterns of the meta-feature landscape with biological knowledge and support functional interpretation of the data based on the ‘guilt by association’ principle.

Finally, case studies selected from different ‘OMIC’ realms are presented in this thesis. In particular, molecular phenotype data derived from expression microarrays (mRNA, miRNA), sequencing (DNA methylation, histone modification patterns) or mass spectrometry (proteome), and also genotype data (SNP-microarrays) is analyzed. It is shown that the SOM analysis pipeline implies strong application

capabilities and covers a broad range of potential purposes ranging from time series and treatment-vs.-control experiments to discrimination of samples according to genotypic, phenotypic or taxonomic classifications.