Abstract

Integrative analysis methods have become essential tools for the extraction of a small set of features which is assumed to be the driver of the measured molecular-biological processes.
Objectives: In the presented master's thesis, three integrative analysis methods based on different mathematical concepts are compared: sparse Canonical Correlation Analysis (sCCA), Non-Negative Matrix Factorization (NMF) and Microarray Logic Analyzer (MALA). They are applied on synthetic data as well as on biological breast cancer data derived on three different levels: the DNA level, the transcript level, and the protein level.
Methods: The resulting sets of selected features are compared with each other directly as well as on a more general level, the associated Gene Ontology (GO) terms. Additionally, the sets of selected features in the biological datasets are compared to genes known to be involved in cancer development.
Results: The observed overlap on the feature level is modest in both the synthetic and the biological datasets. Considering the associated GO terms, the overlap increases in at least one GO category for both datasets. The features selected from the biological dataset by each of the three methods cover about 10% of the features involved in pathways in cancer according to the KEGG database.
Conclusion: The results of integrative analysis of biological data can hardly be validated, however, they can be compared to the results of other integrative analysis methods. The feature sets resulting from the methods under comparison are not congruent. A better agreement between the results can be observed on a higher functional level, the GO term level.