Biochemometrics for the analysis of several data sets

Achim Kohler[1,2,3] and Harald Martens[1,2,4,5]

[1]Centre for Biospectroscopy and Data Modelling, Matforsk, Ås, Norway
[2]CIGENE – Center for Integrative Genetics, University of Life Sciences, 1432 Ås, Norway
[3]Department of Mathematical Sciences and Technology (IMT), Norwegian University of Life Sciences, Ås, Norway
[4]Department of Chemistry, Biology and Food Science, Norwegian University of Life Sciences, N-1432 Ås, Norway
[5]Faculty of Life Sciences, University of Copenhagen, Denmark

Modern instrumentation makes it possible to study biological systems from genotype to phenotype, in terms of high-dimensional and very informative data at different stages up and down along the causal chain: DNA, mRNA, proteome and metabolome can be measured by highly advanced instrumentation. There is a need to analyze this mass of data, to relate the different data sets to each other in light of background knowledge and other available data. The use of bio-spectroscopy in systems biology for low-cost high-resolution phenotype screening has a very high-potential, due to its combination of simplicity and high information content. It is therefore an excellent tool for screening a high number of samples in order to select subsets of particularly interesting samples for more detailed, expensive and time-consuming measurements.

Pre-processing is an important first step of the data analysis in order to remove non-relevant information and reduce linearity problems and noise. By using *a priori* knowledge in model-based pre-processing, physical information, e.g. light scattering, can be separated from biochemical information in spectra.

Integrating biospectroscopic data with other types of data requires data analytical tools for extracting common underlying patterns of mutual information, and for visualizing and interpreting these results. Different multiblock methods can be used to relate different data sets in order to (a) extract common underlying co-variation structures in e.g. FT-IR and DNA, mRNA or proteomic data, and (b) to interpret the obtained results, e.g. patterns of variation in FT-IR bands, to known gene ontologies and -functions or to environmental factors.