SUMMARY

Biomarkers are a useful tool to measure exposures and their biological effects and to evaluate system susceptibility. They have been widely applied in the nutrition area to evaluate dietary intake and its resulting biological consequences and the nutritional status of the study subjects. However, the existing classification schemes of biomarkers are ambiguous leading to uncertainty about their application. With the development of metabolomics technology, the biomarker research field is experiencing rapid changes and new biomarkers are continuously discovered. Challenges also emerge as metabolomics data is large scale, noisy and complex. Therefore, efficient data analysis tools are needed to extract the most relevant information from it.

This thesis aims to develop a classification scheme for the currently applied and newly discovered dietary and health biomarkers in the nutrition field, to optimize the data analysis strategy in untargeted metabolomics studies for the more precise discovery of biomarkers and to apply such strategy for the discovery of biomarkers related to onion consumption.

In paper I, an improved scheme for biomarker classification based on their intended use rather than the technology or outcomes is developed. Six subclasses are suggested: food compound intake biomarkers (FCIBs), biomarkers of food or food component intake (BFIs), dietary pattern biomarkers (DPBs), food compound status biomarkers (FCSBs), effect biomarkers, physiological or health state biomarkers. This scheme provides a well-defined ontology for the field and is beneficial for further increase in the application of biomarkers in the nutrition and health area.

Different bi- and trilinear PLS models for variable selection in untargeted metabolomic studies with time-series design were compared in Paper II. In total, five PLS models with different combinations of bilinear/trilinear X and group/time-response dummy Y were evaluated on simulated datasets with varying characteristics (number of subjects, number of variables, inter-individual variability, intra-individual variability and number of time points) and on two real datasets. Bilinear PLS model with group × time-response as dummy Y provided the highest recall (true positive rate) around 83-95 % with high precision and is independent of the influence of most characteristics of the datasets. Trilinear PLS models tend to select a small number of variables with high precision but relatively high false negative rate. They are also less affected by the noise compared to bilinear PLS models. In general, bilinear models tend to provide higher sensitivity by

increasing the number of selected variables while trilinear models tend to provide higher precision by sacrificing the number of selected variables.

A randomized, controlled crossover meal study to discover and identify BFIs for onion intake was conducted in Paper III. An untargeted UPLC-qTOF-MS metabolic profiling analysis was performed on the urine and blood samples and the profiles were successfully analysed by the modified PLS model with best variable selection performance in Paper II, multilevel PLSDA and nearest shrunken centroid (NSC) to select features associated with onion intake. Eight biomarkers were tentatively identified as biomarkers of food intake for onion and six of them originate from S-substituted cysteine derivatives such as isoalliin, propiin etc., which are considered the most specific for onion intake. Most of the biomarkers were completely excreted within 24 hours and no accumulation was observed during 2 weeks indicating their potential to monitor only recent intake of onions. All the biomarkers showed good performance for predicting onion intake with the area under the curve values (AUC) ranging from 0.81 to 1,

Bilinear PLS model with group × time-response as dummy Y showing best performance in Paper II was successfully applied in the onion intervention study in Paper III and outperformed other adopted methods. When limited sampling time points are available, NSC could also serve as a good choice in addition to PLS based methods.