

## The role of data mining for biomarker discovery and diagnostics in metabolic disorders

Baumgartner C<sup>1</sup>, Pfeifer B<sup>1</sup>, Tilg B<sup>1</sup>, Weinberger K<sup>2</sup>, Ramsay S<sup>2</sup>, Graber A<sup>2</sup>

<sup>1</sup>Research Group for Clinical Bioinformatics, Institute of Biomedical Engineering, University for Health Sciences, Medical Informatics and Technology, Hall in Tyrol, Austria

<sup>2</sup>BIOCRATES life sciences GmbH, Innsbruck, Austria  
christian.baumgartner@umit.at

**Introduction** Recently, due to significant advances in high-throughput technologies, a wider set of the human metabolome - a thus far largely unexplored source of bioinformation - is now accessible [1]. Metabolite profiling technologies comprise a range of advanced analytical and data processing tools, with the objective of utilizing potential markers as a result of comparison of small molecule components of biological systems. Tandem mass spectrometry (MS/MS), for example, detects hundreds of metabolites simultaneously from micro liter quantities of biological samples, such as whole blood, serum, plasma, urine or other body fluids from minute amounts, with high precision and sensitivity [2,3]. Relative quantification is achieved by reference to a wide range of appropriate internal standards. Quality assured data, generated by modern LIMS-controlled technology platforms comprising automated sample preparation, mass spectrometer based analytics and technical validation [4], is rapidly becoming too voluminous to catalogue and interpret by hand, so that cutting-edge data mining tools are needed to identify novel and highly relevant information on pre-processed sample data [5-7]. However, the identification of biologically meaningful biomarkers is challenged by the deficiency of apriori knowledge related to the biomolecular nature of the disease as well as the biological variability of data. Advanced data mining and bioinformatics techniques are applied to increasingly comprehensive and complex MS data sets, with the objective to identify and verify robust and generalizable markers that are biochemically interpretable and biologically relevant in the context of the disease. Ultimately, validated and qualified predictive models can be used for disease screening and therapeutic monitoring [7,8].

**Methods and Results** Additionally to descriptive and test statistics, data mining techniques primarily include for the analysis of mass-spectrometric data feature subset selection methods (PCA, filters, wrappers), classification methods such as logistic regression analysis, support vector machines or neural networks, genetic programming, and cluster analysis. Instances of metabolic data derived from biological samples are represented as a numerical vector in a multi-dimensional space. Here, dimensions or features reflect a vector of analytes with calculated concentrations that relate to pre-defined and pre-annotated metabolites. As principal data mining tasks in biomarker discovery are "supervised", data vectors are defined by a set of tuples  $T_{DB} = \{(c_j, m) \mid c_j \in C, m \in M\}$ , where  $c_j$  is the class label of the collection  $C$  of pre-classified cohorts (diseased, various stages of disease, treated, normal), and  $M = \{m \mid m_1, \dots, m_n\}$  is the given set of metabolite concentrations. Success of data mining is affected by factors such as noise, redundancy, relevance or reliability inherent in the experimental data. Feature selection, an important data mining task for biomarker discovery, focuses on the process of identifying and removing as much of irrelevant or redundant information as possible and is used as pre-processing step before classification and biochemical interpretation. BMI - the biomarker identifier, an algorithm recently described by [8] makes use of a two-step data processing procedure to discern the discriminatory attributes between two classes of interest (e.g. diseased vs. normal). In seven inborn errors of metabolism, i.e. phenylketonuria (PKU), glutaric acidemia, Type I (GA-I), 3-methylcrotonylglycinemia deficiency (3-MCCD), methylmalonic acidemia (MMA), propionic acidemia (PA), medium-chain acyl CoA dehydrogenase deficiency (MCADD), and 3-OH long-chain acyl CoA dehydrogenase deficiency (LCHADD), BMI identified all key metabolites and prioritized them according to the current biochemical knowledge on disease metabolism, so that the algorithm is suitable for identifying and prioritizing selected metabolites in single pathway blockades disorders (Fig. 1).

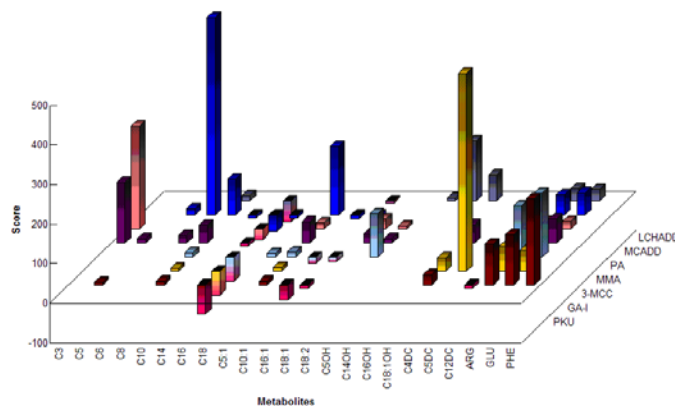


Fig. 1: Visualization of the abnormal metabolite profiles of seven inborn errors of metabolism identified by BMI. Cx are acylcarnitines, ARG = arginine, GLU = glutamate and PHE = phenylalanine.

For the clinical routine, the predictive performance and generalization power of candidate biomarkers is utilized to build classification models for disease screening. Typically high sensitivity and specificity is required to rule out other diseases. Additionally, the models have to consider and adjust for the real incidence rate of a disease to calculate false-positive rates, assuming that the prevalence of the disease was artificially controlled in the study. Particularly for screening applications, model-based classifiers such as logistic regression analysis or classification trees are rather used than instance- or kernel based methods like k-NN, support vector machines or neural networks. The use of explicit rules described by the models' target decision function is more practical for the daily screening routine and showed highest acceptance rates by the clinical personal.

Supervised and unsupervised data mining techniques are applied to reveal statistically significant, putative biomarkers. A further fundamental step of data mining is to verify putative marker candidates in a biochemical context by mining most likely pathways. Therefore, for querying appropriate knowledge bases powerful search and retrieval tools are needed that map and link discovered marker candidates against a variety of sources such as online repositories or internal databases. Metabolic explorer tools are needed to visualize directly affected biochemical pathways, map experimental metabolite concentrations on these graphs, and reconstruct a spectrum of theoretically possible pathways between relevant metabolites. Metabolic information is primarily extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and stored in a relational database. Dynamic visualization tools, and compound, structure and route finding algorithms were designed and implemented for the interactive mining of biochemical pathways and related entities, such as reactions, enzymes and metabolites, facilitating a direct functional annotation of experimental results. Generally, three steps need to be processed: *Searching* for entity mining and annotated pathway association, *browsing* for mapping of experimental metabolite concentrations, and *routing* for pathway reconstruction.

**Conclusion** The discovery, biochemical and biological interpretation, statistical verification, independent validation of biomarker candidates typically requires the interdisciplinary expertise and team work of biostatisticians, clinicians, biologists, analytical and biochemists and bioinformaticians, and involves the proficient planning, execution and control of all study steps, ranging from the experimental or clinical trial design to the discovery and validation of putative markers, respectively [9]. In particular, targeted metabolite concentration profiling in combination with appropriate data mining strategies have the potential to revolutionize diagnostics and drug development.

## References

- [1] Beecher C. The human metabolome. In Harrigan GG, Goodacre R (Ed). *Metabolic profiling: Its role in biomarker discovery and gene function analysis*. Kluwer Academic Publishers, Boston/Dordrecht/London; 2003, pp. 311-319.
- [2] Roschinger W, Olgemoller B, Fingerhut R, Liebl B, Roscher AA. Advances in analytical mass spectrometry to improve screening for inherited metabolic diseases. *Eur J Pediatr* 2003;162 (Suppl 1):S67-76.
- [3] Kaltashov IA, Eyles SJ. *Mass spectrometry in biophysics: Conformation and dynamics of biomolecules*. Wiley, New York; 2005.
- [4] Weinberger KM, Ramsay S, Graber A. Towards the biochemical fingerprint. *Biosystems Solutions*, 2005;12:36-37.
- [5] Huyn N. Data analysis and mining in the life sciences. *ACM SIGMOD Record*, 2001;30:76-85.
- [6] Fiehn O, Spranger J. Use of metabolomics to discover metabolic patterns associated with human diseases. In Harrigan GG, Goodacre R (Ed). *Metabolic profiling: Its role in biomarker discovery and gene function analysis*. Kluwer Academic Publishers, Boston/Dordrecht/London; 2003, pp. 199-215.
- [7] Baumgartner C, Böhm C, Baumgartner D, Marini G, Weinberger K, Olgemöller B, Liebl B, Roscher AA. Supervised machine learning techniques for the classification of metabolic disorders in newborns. *Bioinformatics* 2004;20:2985-2996.
- [8] Baumgartner C, Baumgartner D. Biomarker discovery, disease classification, and similarity query processing on high-throughput MS/MS data of inborn errors of metabolism. *J Biomol Screen* 2006;11:90-99.
- [9] Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* 2004;4:309-314.