Bioinformatics Advance Access published June 4, 2004

Supervised machine learning techniques for the classification of metabolic disorders in newborns

Baumgartner C, PhD[1*], Böhm C, PhD[2], Baumgartner D, MD[3], Marini G, PhD[4], Weinberger K, PhD[4], Olgemöller B, MD, PhD[5], Liebl B, MD[6], Roscher AA, MD[7]

[1]Research Group for Biomedical Data Mining, University for Health Sciences, Medical Informatics and Technology, Innrain 98, A-6020 Innsbruck, Austria,

[2]Institute for Computer Science, University of Munich, Oettingenstrasse 67, D-80538 Munich, Germany,

[3]Department of Pediatrics, Medical University of Innsbruck, Anichstrasse 35, A-6020 Innsbruck, Austria,

[4]Biocrates Life Sciences Biotechnology GmbH, Innrain 66, A-6020 Innsbruck, Austria,

[5]Laboratory Becker, Olgemöller & Colleagues, Führichstrasse 70, D-81671 Munich, Germany,

[6]Public Health Newborn Screening Center of the State of Bavaria, Landesuntersuchungsamt Südbayern, D-85762 Oberschleissheim, Germany,

[7]Department of Biomedical Genetics and Molecular Biology, Dr. von Hauner Children's Hospital, University of Munich, Lindwurmstrasse 4, D-80337 Munich, Germany.

Running head:
Machine learning in metabolic disorders

To whom correspondence should be addressed*:
Dr. Christian Baumgartner
Research Group for Biomedical Data Mining
Institute for Information Systems
University for Health Sciences, Medical Informatics and Technology
Innrain 98
A-6020 Innsbruck
Austria
Phone: +43 (0)50 8648 3827
PC-Fax: +43 (0)50 8648 673827
Mailto: christian.baumgartner@umit.at

Abstract

Motivation: During the Bavarian newborn screening program all newborns have been tested for about 20 inherited metabolic disorders. Due to the amount and complexity of the generated experimental data machine learning techniques provide a promising approach to investigate novel patterns in high-dimensional metabolic data which form the source for constructing classification rules with high discriminatory power.
Results: Six machine learning techniques have been investigated for their classification accuracy focusing on two metabolic disorders, PKU and MCADD. Logistic regression analysis led to superior classification rules (sensitivity >96.8%, specificity >99.98%) compared to all investigated algorithms. Including novel constellations of metabolites into the models, the positive predictive value could be strongly increased (PKU 71.9% vs. 16.2%, MCADD 88.4% vs. 54.6% compared to the established diagnostic markers). Our results clearly proof that the mined data confirm the known and indicate some novel metabolic patterns which may contribute to a better understanding of newborn metabolism.

Availability:
WEKA machine learning package: www.cs.waikato.ac.nz/~ml/weka
Statistical software package ADE-4: http://pbil.univ-lyon1.fr/ADE-4

Contact:
christian.baumgartner@umit.at

Keywords:
Metabolomics; supervised machine learning; newborn screening; phenylketonuria, medium-chain acyl-CoA dehydrogenase deficiency.

Introduction

Background

Newborn screening programs for severe metabolic disorders, which hinder an infant's normal physical or mental development, are well-established (Liebl *et al.*, 2002, Liebl *et al.*, 2003). Otherwise not apparent at this early age, these metabolic disorders can be addressed by effective therapies. New and refined screening methodologies based on tandem mass spectrometry of metabolites have been developed for routine deployment (Millington *et al.*, 1984). The functional endpoint of metabolic cycles, which offer a precise snapshot of the current metabolic state, can be detected in a single analysis of a small blood sample that is collected during the first few days of life. Screening simultaneously for more than 20 inherited metabolic disorders by quantifying the concentrations of up to 50 metabolites (Millington *et al.*, 1992, Chace *et al.*, 1999), the amount and complexity of the experimental data is quickly becoming unmanageable to be evaluated manually. Therefore, machine learning techniques have been suggested to discover and mine novel data in metabolic networks and to construct screening models for metabolic disorders in newborns with high predictive power (Mendes, 2002, Neville *et al*., 2003, Purohit *et al*., 2003, Baumgartner *et al*., 2004).

Task definition

Focusing on two representative inborn errors of metabolism – phenylketonuria (PKU), an amino acid disorder, and medium-chain acyl-CoA dehydrogenase deficiency (MCADD), a fatty acid oxidation defect – six well-established supervised machine learning techniques were evaluated to determine the "best" screening model according to the following criteria:

- *discriminatory performance* of the learning algorithm based on pre-classified, selected and clinically validated sub-databases of PKU and MCADD newborns, and controls,
- *diagnostic prediction* of constructed classifiers with respect to optimising sensitivity and minimizing the number of false positive results considering a larger database of approximately 100,000 controls.

In particular, we compared the classification capabilities of three directly interpretable decision rules (discriminant analysis, logistic regression analysis and decision trees), which represent the data relations in an explicit way, e.g. in a formula or in a tree-like structure, and three not directly interpretable techniques (k-nearest neighbours, artificial neural networks and support vector machines), which can not easily be described in terms of the original variables or attributes. Correlation based feature-selection methods were applied which aim at removing irrelevant and redundant metabolites while retaining or improving the discriminatory power of our classification models.

System and Methods

Tandem mass spectrometry (MS/MS)

A mass spectrometer separates ions based on their mass/charge (m/z) ratios. Characteristic patterns of fragments and relative peak intensities in the resulting spectrum allow qualitative as well as quantitative determination of chemical compounds. By coupling two mass spectrometers, usually separated by a reaction chamber or collision cell, the modern tandem mass spectrometry (MS/MS) allows simultaneous analysis of multi-compounds in a high-throughput process (Millington *et al.*, 1984). MS/MS has been used for several years to identify and measure carnitine ester concentrations in blood and urine of children suspected of having inborn errors of metabolism. Indeed, acylcarnitine analysis is a superior diagnostic test for disorders of fatty acid oxidation because abnormal levels of related metabolites are detected before the patient is acutely ill (Millington *et al.* 1992). More recently, MS/MS has been used in pilot programs to screen newborns for these conditions and for disorders of amino- and organic-acid metabolism as well

(Liebl *et al.*, 2002, Liebl *et al.*, 2003). MS/MS thus permits very rapid, sensitive and, with internal standards, accurate quantitative measurement of many different types of metabolites by conversion of raw mass spectra into clinically meaningful results (concentrations).

Disease characteristics, metabolism and epidemiological aspects of investigated disorders

Phenylketonuria (PKU, OMIM #261600) is an amino acid disorder which is caused primarily by a deficiency of phenylalanine hydroxylase activity with blocked hydroxylation of phenylalanine to tyrosine or impaired synthesis or recycling of the biopterin (BN4) cofactor. Phenylalanine hydroxylase deficiency produces a spectrum of disorders including classic PKU, non-PKU hyperphenylalaninemia and variant PKU (Chace *et al.* 1993, Rashed *et al.*, 1995, Guldberg *et al.*, 1998). Untreated children with persistent severe PKU show impaired brain development. Signs and symptoms include microcephaly, epilepsy, mental retardation, and behaviour problems. Since the appearance of universal newborn screening, symptomatic classic PKU is infrequently seen. Its predicted incidence in screened populations of less than one in a million live births reflects those children not detected by newborn screening. Prevalence of phenylalanine hydroxylase deficiency in various populations shows different values: Turks (1:2,600), Caucasians (1:10,000), Japanese (1:143,000). In our experiments we focused on cases of classic PKU. The estimated incidence of classic PKU calculated from Bavarian newborn screening (NBS) data is approximately 1:14,000.

Medium-chain acyl-CoA dehydrogenase deficieny (MCADD, OMIM #201450) is a fatty acid oxidation defect which leads to an accumulation of fatty acids and a decrease in cell energy metabolism. Fatty acids that accumulate due to the erroneous metabolism of MCADD are C6-carnitine, C8-carnitine, C10-carnitine as well as C10:1, which is metabolised by four β-oxidation cycles of oleyl-carnitine (C18:1) (Van Hove *et al.*, 1993, Rashed *et al.*, 1995, Blau *et al.*, 2001, Rinaldo *et al.*, 2002). Patients with MCADD (clinically two forms of MCADD can be distinguished, i.e. “classic” and “mild” MCADD) appear normal at birth and usually present between three and 24 month of age in response to e.g. intercurrent and common infections. Instances of metabolic stress can lead to vomiting and lethargy, which may quickly process to coma and death. MCADD is a disease that is prevalent in Caucasians, especially those of Northern European descent. The overall frequency of the disorder has been estimated to range from 1:4,900 to 1:17,000 (variations related to the ethnic background of populations). Based on NBS programs worldwide, the incidence has been defined in Northern Germany (1:4,900), USA (1:15,700) and in Australia (1:25,000). In our experiments we only investigated the classic form of MCADD. The estimated incidence of classic MCADD calculated from Bavarian (Southern Germany) NBS data is approximately 1:10,000.

For the screening of inborn errors of metabolism physicians generally use decision rules or flags, which are based on so-called primary diagnostic metabolites. In 2000 the American College of Medical Genetics/American Society of Human Genetics Test and Technology Transfer Committee Working Group (ACMG/ASHG) published a guideline where these primary metabolites/markers are summarized. According to this document phenylalanine (Phe) and tyrosine (Tyr) are mentioned as established primary markers for PKU, C8-carnitine and C10:1-carnitine for MCADD. We have used the proposed markers as reference for discussion of our found metabolites.

Examined newborn screening data

Our experimental datasets were anonymously provided from the newborn screening program in Bavaria, Germany (Public Health Newborn Screening Center of the State of Bavaria, Oberschleissheim) between 1999 and 2002. A single blood sample, which has been taken within a few days after newborn’s birth, undergoes MS/MS analysis, the measured metabolic datasets have been saved in a database (file based DB, stage 2002).

For an objective train-and-test design cycle we focused on one representative disorder of the amino acid and one of the fatty acid oxidation metabolism, each of them showing a relatively high incidence in-between their group of disorders (classic PKU, n=43 cases and classic MCADD, n=63 cases). The number of cases of further screened metabolic disorders was unfortunately too small for

an useful examination (e.g. 3-OH long chain acyl CoA dehydrogenase deficiency, LCHADD, n=2; short chain Acyl CoA dehydrogenase deficiency, SCADD, n=1; very long chain Acyl CoA dehydrogenase deficiency, VLCADD, n=5; propionic acidemia, PA, n=6; and methylmalonic acidemia, MMA, n=5). Based on the limited number of PKU and MCADD cases we created a statistically representative control group from the NBS database (~600,000 entries, end of year 2002) using a rate of ~1:25 (disorder to controls). Therefore the PKU sub-database (train-and-test database) contains all 43 cases designated as confirmed classic PKU and a small number of randomly sampled controls (1241 cases i.e. each 500^{th} case from NBS controls, which represent all newborns without verified cases of known metabolic disorders), the MCADD sub-database contains all 63 cases newborns of classic MCADD and again 1241 controls. A much larger randomly sampled control group of 98,411 cases, representing one sixth of the NBS control database, serves to obtain reliable estimates of the false positive rates.

Experimental design of the classification analysis

The general scheme for constructing a screening model (classifier) of high-dimensional metabolic data is illustrated in figure 1. Starting from the NBS database, we first selected two sub-databases containing all available PKU and MCADD cases and a representative small number of randomized controls for training and cross (X)-validation. Thus, computational efficiency could be ensured without loosing the models' classification accuracy on reduced data during training phase. However, aim of the classification task is to achieve highest discriminatory performance by minimizing the number of false negative and false positive cases. Feature sub-selection algorithms led to a reduced number of metabolites relevant for the calculated classifier. Finally, the most sensitive models were re-evaluated by applying them on a larger database of approximately 100,000 controls in order to estimate a representative value for specificity.
[FIGURE 1]

Supervised machine learning techniques

Usually, for a supervised classification problem, the training data sets are in the form of a set of tuples $\{(y_1, x_{1,j}),\ldots, (y_n, x_{n,j})\}$ where y_i is the class label and x_{ij} is the set of attributes (metabolites) for the instances. The task of the learning algorithm is to produce a classifier (model) to classify the instances into the correct class. The used classification and feature selection algorithms are described shortly in the section "algorithms".

Algorithms

Classification algorithms

Discriminant analysis (DA)

Both discriminant analysis and logistic regression analysis construct a separating hyperplane between the two datasets. This hyperplane is described by a linear discriminant function $z = f(x_1,\ldots x_n) = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c$ which equals to zero at the hyperplane if two preconditions are fulfilled: (i) multivariate normal distribution in both datasets and (ii) homogeneity of both covariance matrices. For discriminant analysis, the hyperplane is defined by the geometric means between the centeroids (i.e. the centres of gravity) of the two data sets. To take different variances and covariances in the data sets into account, the variables are usually first transformed to standard means ($\mu=0$) and variance ($\sigma^2=1$) and the Mahalanobis distance (an ellipsoid distance determined from the covariance matrix of the data set) is preferred to the Euclidean distance. (McLachlan, 1992).

Logistic regression analysis (LRA)

Similar to DA logistic regression analysis constructs a linear separating hyperplane between the two datasets which have to be distinguished by the classifiers. Additionally, a logistic function *p*

$$p = \frac{1}{1+e^{-z}} \quad (1)$$

is used to consider the distance from the hyperplane as a probability measure of class membership, where *p* is the conditional probability of the form $P(z=1| x_1,...,x_n)$ and *z* the logit of the model. The class membership to both classes is indicated by a cut-off value (p=0.5 by default). LRA uses a maximum likelihood method which maximises the probability of getting the observed results given the fitted coefficients (Hosmer *et al*., 2000).

Decision trees (DT)

Decision trees are rooted, usually binary trees, with simple classifiers placed at each internal node and a class label at each leaf. For most DT algorithms, these simple classifiers associated with the internal nodes are comparisons between an input variable and a fix value. The algorithm most often used to generate decision trees is ID3 (Quinlan, 1986) or its successors C4.5 and C5.0, respectively (Quinlan, 1993). This algorithm selects the next node to place in the tree by computing the information gain for all candidate features. Information gain (IG) is a measure how well the given feature A separates the remaining training data by expecting a reduction of entropy E, a measure of the impurity in the data (Mitchell, 1997).

$$E(S) = \sum_{c \in C} -\frac{|S_c|}{|S|} \cdot \ln \frac{|S_c|}{|S|} \quad (2)$$

$$IG(S,A) = E(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} \cdot E(S_v) \quad (3)$$

S represents the data collection, |*S*| its cardinality, *C* is the class collection, S_c the subset of *S* containing items belonging to class *c*, *V(A)* is the set of all possible values for feature A, S_v is the subset of *S* for which A has value *v*. We used the C4.5 algorithm with reduced-error-pruning option to avoid overfitting of training data.

k-nearest neighbour classifier (kNN)

A k-NN classifier defines decision boundaries in an n-dimensional space which separate different sample classes from each other in the data. The learning process consists in simply storing the presented data. All instances correspond to points in an n-dimensional space and the nearest neighbours of a given query are defined in terms of e.g. the standard Euclidean distance. The probability of a query *q* belonging to a class *c* can be calculated as follows:

$$p(c \mid q) = \frac{\sum_{k \in K} w_k \cdot 1_{(kc=c)}}{\sum_{k \in K} w_k}, \; w_k = \frac{1}{d(k,q)} \quad (4)$$

where *K* is the set of nearest neighbours, *kc* the class of *k* and *d*(*k*,*q*) the Euclidean distance of *k* from *q*. Larger values of *K* consider more neighbours, and therefore smooth over local characteristics, smaller values leads to limited neighbourhoods (Mitchell, 1997). In general, *K* can only be determined empirically. For our data representation we proposed *K* values of 1, 3 and 5.

Artificial neural networks (ANN)

An artificial neural network is an information processing paradigm that is inspired by the biological nervous systems, such as the brain. The network consists of several layers of neurons, which are the input, hidden and output layers. An input layer takes the input and distributes it to the hidden layers which do all the necessary computation and output the results to the output layer.

The standard algorithm which we used is a multi-layered ANN trained using backpropagation and the delta rule. This algorithm attempts to minimize the squared error between the network output values and the target value for these outputs (Bishop, 1995; Mitchell, 1997). The ANN was designed using a single layer of hidden units with (number of attributes + number of classes)/2 hidden units. Note that too many or too few hidden units can lead to over- or underestimation of training data. We chose 500 epochs to train through and a learning rate of 0.3, the amount the weights to be updated.

Support Vector Machines (SVM)
The basic idea of a SVM classifier is that the data vectors can be separated by a hyperplane. In the simplest case of a linear hyperplane there may exist many possible separating hyperplanes. Among them, the SVM classifier seeks the separating hyperplane that produces the largest separation margin between the two classes. Such a scheme is known to be associated with structural risk minimization to find a learning machine that yields a good trade-off between low empirical risk and small capacity.
In the more general case in which the data points are not linearly separable in the input space, a non-linear transformation is used to map the data vector $\boldsymbol{x}$ into a high-dimensional space prior to applying the linear maximum-margin classifier. To avoid over-fitting in this higher dimensional space, a SVM uses kernel functions (polynomial and Gaussian radial basis kernels are the most common) in which the non-linear mapping is implicitly embedded. With the use of a kernel, the decision function in a SVM classifier has the following form:

$$f(x) = \sum_{i=1}^{L_S} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x}) + b \qquad (5)$$

where $K(\cdot,\cdot)$ is the kernel function, $\boldsymbol{x}_i$ are the so-called support vectors determined from training data, L_S is the number of support vectors, y_i is the class indicator associated with each $\boldsymbol{x}_i$, and α_i, the Lagrange multipliers. Additionally, for a given kernel it is necessary to specify the cost factor c, a positive regularization parameter that controls the trade-off between complexity of the machine and the allowed classification error (Cortes *et al.*, 1995, Vapnik, 1998). We used the SVM with its simplest case of a linear hyperplane and with polynomial kernels of degree 2 and 3. The cost factor c was set to 100 for all three SVM settings.

Feature extraction algorithms
Feature subset selection is the process of identifying and removing as much irrelevant and redundant information as possible. This reduces the dimensionality of the data and may allow learning algorithms to operate faster and more efficiently (Mitchell, 1997).
We propose the *filter* approach using *gain ratio* and *relief*, two representatives of correlation based extraction techniques coupling an applicative correlation measure with a heuristic search strategy: As described previously the effectiveness of a feature in classifying the training data can be quantified using the given entropy E (equation 2). Using equation 3 (information gain, IG) the expected reduction of entropy caused by partitioning the data according to feature A can be measured. Thereby, IG favors features with many different values over those with few values which is not always desired. The concept of *gain ratio (GR)* overcomes this problem by introducing an extra term SI taking into account how the feature A splits the data.

$$GR(S, A) = \frac{IG(S, A)}{SI(S, A)} \text{ , with}$$

$$SI(S, A) = -\sum_{i=1}^{d} \frac{|S_i|}{|S|} \cdot \ln \frac{|S_i|}{|S|} \qquad (6)$$

where S_i are d subsets of data resulting from partitioning S by the d-valued feature A. For the special case where the SI term can be 0, GR(S, A) is set to IG(S, A).
Relief is a feature weighting algorithm that is sensitive to feature interactions. It evaluates the merit of a feature by repeatedly sampling an instance and considering the value of the given feature for the nearest instance of the same class (nearest hit) and different class (nearest miss). Equation 7 represents the weight updating formula:

$$W_A = W_A - \frac{\mathrm{diff}(A,R,H)^2}{m} + \frac{\mathrm{diff}(A,R,M)^2}{m} \quad (7)$$

where W_A is the weight for attribute A, R is a randomly sampled instance, H is the nearest hit, M is the nearest miss, and m is the number of randomly sampled instances. The function *diff* calculates the difference between two instances for a given attribute (Kira *et al*., 1992; Kononenko, 1995). The number of nearest neighbours was selected to be 10.

Evaluation of classifier's accuracy and validation
We evaluated the discriminatory power of the investigated techniques constructing a classification (confusion or contingency) table for our two class problem stating true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The evaluation measure most frequently used in classification is accuracy (Acc) which describes the proportion of correctly classified instances: Acc = (TP+TN)/(TP+FP+TN+FN). Measures which consider more precisely the influence of the class size are sensitivity (S_n) or recall, specificity (S_p), positive predictive value (PPV) or precision and negative (NPV) predictive value. S_n = TP/(TP+FN) measures the fraction of actual positive instances that are correctly classified; while S_p = TN/(TN+FP) measures the fraction of actual negative examples that are correctly classified. The PPV (or the reliability of positive predictions) is computed by PPV=TP/(TP+FP), the NPV is defined as NPV=TN/(TN+FN) (Salzberg, 1999).
Five classification algorithms (DT, LRA, kNN, ANN and SVM) and the feature extraction techniques *gain ratio* and *relief* used in this study were obtained from the WEKA machine learning package (http://www.cs.waikato.ac.nz/~ml/weka). WEKA is a publicly available, widespread and comprehensive tool set which guarantees high comparability of our results. DA and statistical analysis were performed with the software package ADE-4 (http://pbil.univ-lyon1.fr/ADE-4). The packages were used to investigate the models' discriminatory power on full and reduced data dimensionality. An established methodology to evaluate the robustness of the classifier is to perform a cross validation on the classifier. 10-fold-cross validation has been proved to be statistically good enough in evaluating the classification accuracy of the models (Witten *et al*., 2000).

Experiments

Descriptive statistics of metabolic data
Table 1 summarizes all metabolites measured by mass spectrometry: 14 amino acids representing the spectrum of metabolites involved in investigated amino acids disorders and 29 fatty acids (acylcarnitines) involved in the metabolism of fatty acid oxidation defects. The mean concentrations and respective standard deviations (sd) of all metabolites are given in μmol/L. Differences of metabolite concentrations between disorder and control group were performed with unpaired significance testing.

Comparison of classification methods examined on the full metabolite dimensionality
In order to investigate the discriminatory performance in a high-dimensional feature space, we first examined all presented supervised machine learning algorithms on selected sub-databases considering full amino acid dimensionality (PKU sub-database) and full fatty acid dimensionality

(MCADD sub-database), respectively. Overview of full amino and fatty acid dimensionality see table 1. The effectiveness of the classifiers is summarized in table 2 a-b.
These results revealed that most of the classifiers (without DA and k-NN) applied on the PKU database performed well in terms of classification accuracy ($S_n \geq 95.3\%$, $S_p \geq 99.8\%$, Acc $\geq 99.7\%$). Except the DT learner, all classifiers indicated an optimal specificity of 100%. Thereby, not directly interpretable algorithms such as ANN and SVM-2 yielded a minor advantage in sensitivity compared to the other ones (table 2a). Running our experiments on the MCADD sub-database, all six algorithms showed reduced classification accuracy (cf. S_n and Acc compared to the PKU results, table 2b). This tendency may arise from the induced classifiers being able to characterize the negative samples as our training set contains twice as much higher feature dimensionality compared to the PKU data (29:14 metabolites). In general, the DA and k-NN learners demonstrated decreased classification accuracy for both PKU and MCADD datasets. LRA, ANN, DT, SVM-1 (linear hyperplane) and SVM-2 (polynomial kernel, degree 2) led to better discrimination and, accordingly, classification accuracy indicated by high S_n ($\geq 95.3\%$) and high S_p ($\geq 99.8\%$) in PKU data, and minor reduced S_n ($\geq 92.1\%$) but also superior S_p of $\geq 99.6\%$ in MCADD data.

Feature extraction and metabolic patterns
Feature extraction methods identify redundant metabolites which can be removed leading to simplified classification models. We applied two correlation based filter techniques, *gain ratio* and *relief,* in order to identify most significant metabolites. Figure 2 a-b summarizes the metabolic patterns resulting from both techniques. Black bars indicate the established diagnostic markers.

[FIGURE 2]

According to the sequence of the amino acids group obtained by the gain ratio filter (figure 2a), Glu, Argsuc and Arg showed a high impact in PKU data in addition to Phe. However, these results correspond just partly with the abnormal PKU metabolism, as solely Phe and Tyr are used as conventional diagnostic metabolites for screening for PKU. Little differences were observed in the relief ranking where again Phe, which shows a highly accumulated concentration, was top-placed followed by the acids Xle, Glu, Val and Gly. The diagnostic marker Tyr, which is significantly diminished in PKU metabolism (cf. table 1), ranked at an irrelevant position in both filter approaches. In addition to Phe, Arg and Glu yielded strongly increased concentration levels, but accompanied by high variances (cf. descriptive results) in PKU data. However, they show no significant concentration changes in the control group. These observations can not be directly explained by the deficiency of phenylalanine hydroxylase activity, but seem to be an interesting secondary effect of metabolism which currently is in discussion with our clinical and biochemical experts.
For MCADD data the gain ratio ranker figured out a strong dominance of octanyl-carnitine (C8), followed by C10:1, which corresponds well to the established diagnostic markers. It is also of interest to note that the result of the relief algorithm yielded similar ranking results in the order of the first six fatty acids. In addition to C8 and C10:1 also C10 and C2 (medium and short chain fatty acids) with elevated and C12DC and C18:1 (long chain fatty acids) with diminished concentration levels correspond quite well with the abnormal MCADD metabolism.

Comparison of classifiers examined on reduced metabolite dimensionality
We applied five of six machine learning algorithms (DT learner was not examined on the reduced feature spectrum due to its internal feature selection strategy) to the established diagnostic markers (ACMG/ASHG statement, 2000), which served as a reference for employing a low dimensional metabolite space. Table 3a-b summarizes the classification accuracy for PKU's and MCADD's primary diagnostic metabolites Phe and Tyr, and C8 and C10:1, respectively. Despite the small nuances on differences in classification accuracy within the examined algorithms on full and reduced feature dimensionality LRA is top-ranked for MCADD ($S_n = 95.2\%$) and together with ANN and k-NN learners also best ranked for PKU ($S_n = 97.7\%$). By contrast, DA and SVMs with

polynomial kernels show considerably decreased S_n values compared to full metabolite dimensionality. However, the minor alterations of classification accuracy in both disorders are caused simply by the obvious statistically significant differences between the groups of disorder and controls (cf. descriptive results).

Screening models for classic PKU and classic MCADD

Models which may prove feasible for clinical routine have to ensure easy interpretation without loosing predictive power. Within this context, from all six investigated machine learning paradigms, LRA, 1-NN, 3-NN, ANN and SVM-1 gave promising classification results on reduced metabolite dimensionality. For the screening of classic MCADD and classic PKU, we trained the LRA model - a paradigm widely used in medical applications - on both metabolic sub-databases showing highest S_n of ≥ 95.2% (cf. table 3). In order to further optimise the model's discriminatory performance we computed the six top-ranked metabolites as investigated from feature selection methods and examined their possible pair-wise combinations. Including combinations of more than two metabolites did not further improve the classification accuracy. Table 4 summarizes the most sensitive screening models (PKU: S_n = 95.4-100%, MCADD: S_n = 95.2-96.8%). The total number of falsely negative classified newborns did not exceed 0-4.6% for PKU and 3.2-4.8% for MCADD. Promising models predominantly include two metabolites, but differing partly from the clinically applied diagnostic metabolites. Following the process of analysis depicted in figure 1, our most sensitive models were consecutively re-evaluated on a larger control database of 98,411 cases. As expected this procedure decreased specificities of the constructed models significantly. However, the models' real classification accuracy in terms of the false positive rate and the positive predictive value can now easily be extrapolated considering the estimated disorder's incidences.

All derived screening models for PKU contain Phe as the predominant metabolite which is consistent with its role in erroneous metabolism (Chace *et al.* 1993, Rashed *et al.*, 1995). Models including Phe alone or combined with Arg or Argsuc yielded the highest PPV of 70.7-71.9%, i.e. 16-17 FP cases (0.00017%) out of 98,411 controls. The latter constellations, however, cannot be directly explained from the PKU metabolism. The classifier which includes the established diagnostic markers Phe and Tyr show maximum S_n (100%), but its PPV drops off significantly (16.2%). However, combinations of Phe and further meaningful metabolites do not change classification accuracy significantly, Phe remains the key marker for PKU. Nevertheless, the role of Arg, Glu, Argsuc, Val or Xle in alternative pathways needs to be cross-checked in order to understand their individual influence on PKU metabolism.

MCADD models led to S_n values (95.2-96.8%) slightly decreased but still superior to the established clinical markers; S_p and PPV showed higher values compared to the PKU models. The most sensitive model (S_n = 96.8%) combining octanyl-carnitine (C8) and octadecenoyl-carnitine (C18:1) yields an excellent S_p of 99.992% and a PPV of 88.4%. In other words only a marginal fraction of 8 FP cases (0.00008% of all controls) is wrongly classified to be a classic MCADD patient. Considering the established diagnostic markers C8 and C10:1 in the model, PPV decreases to 54.6%, so that decision rules based on alternative combinations of C8 + C18:1 and C8 + C12DC or C8 alone seem to be the better markers to enhance discriminatory performance and thus to optimise classification accuracy.

In addition the derived classification models allow to calculate a conditional probability value of the form $P(disorder = 1 | metabolite_1, \ldots, metabolite_n)$. The logits of constructed LRA models ($z = a_0 + a_1 m_1 + \ldots + a_n m_n$) are presented in table 4 (final column).

Discussion

Machine learning techniques have great potential to increase our knowledge in functional metabolomics, an area which is still in the early stages of comprehensive investigation. Focusing on inborn errors of metabolism from newborn screening data the metabolic patterns of a wide spectrum

of amino and fatty acid concentrations were examined in order to enhance diagnostics in an early stage of disorder.

Therefore we investigated the use of six different machine learning techniques for their suitability to construct classification models for two severe metabolic disorders, PKU and MCADD. To increase the classifier's effectiveness we reduced full metabolite dimensionality with two correlation based feature selection algorithms, gain ratio and relief, with their central hypotheses that good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other. Experiments showed that correlation based feature selection quickly removes irrelevant, redundant, and noisy features, and identifies relevant attributes as long as their relevance does not strongly depend on other features (Hall, 1999). In most cases, classification accuracy using reduced feature (metabolite) dimensionality equaled or increased accuracy using the entire metabolite spectrum as our experiments confirm.

The highest discriminatory performance was achieved with the LRA model, a directly interpretable technique, which proved readily applicable in the daily screening procedure. The resulting discriminant function can easily be cross-checked with already acquired patient data. Furthermore, the formulas can be used as a starting point for the detection of previously unknown causal dependencies in metabolic pathways.

For both disorders the computed sensitivity of the best LRA models ranged above 96.8%, the specificity exceeded 99.98%. By including novel constellations of metabolites into our models – as examined by the feature extraction procedures – specificity and PPV could be increased compared to the established screening metabolites. In case of PKU the PPV, which was determined on a larger sampled control database of 98,411 cases, improved up to 71.9% using solely Phe, and up to 70.7% by combining e.g. Phe and Arg. The PPV deteriorated to just 16.2% when considering Phe and Tyr, both of them being the metabolites predominantly altered in the abnormal PKU metabolism. Note that for the classification task solely single metabolite concentrations have been considered as model input variables. However, modelling e.g. a ratio of Phe/Tyr, which represent abnormal changes of Phe ($\uparrow$) and Tyr ($\downarrow$) due to the blocked hydroxylation of Phe to Tyr, the PPV can significantly be increased as already shown elsewhere (Chace *et al.*, 1998).

For MCADD the model's PPV increased significantly compared to the PPV of 54.6% for the established screening metabolites C8 and C10:1 resulting in PPV values of 88.4% for C8 and C18:1, 88.2% for solely C8 and 85.7% for combining C8 and C12DC, respectively. The false positive rate for MCADD most likely varies between screening programs because of differences in acylcarnitine analysis and profiling. Programs that screen for MCADD but not for other fatty oxidation disorders often limit their analysis to C8, the predominant, but not specific marker for MCADD which is elevated in several other disorders (e.g. medium/short chain 3-OH acyl CoA dehydrogenase deficiency or glutaric acidemia type II). Consideration of the disorders included in the differential diagnosis should minimize the false positive rate. Our presented models including novel combinations such as C8 + C18:1 and C8 + C12DC give additional information with respect to the aforementioned differential diagnostic challenges. However, the experimental confirmation is essential and is part of our ongoing investigations.

For the routine clinical screening LRA models proved particularly feasible because of their highly significant prognostic accuracy. The models permit to calculate the probability for the occurrence of the disorder by classifying the tested newborns according to a default cut-off level of $p = 0.5$. By employing sharper cut-offs (e.g. $0.25 \leq p \leq 0.75$, i.e. between the first and third quartile), this approach can be extended to a prognostic "alarm system" allowing a more effective response to cases of metabolic disorders detected during the screening procedure. Subsequent diagnostic clarification has only to focus on this "third" class of newborns in the interval [0.25, 0.75] which is highly suspicious for the screened disorders. However, the presented models showing high specificity do not always have optimal S_n of 100%. A feasible procedure for optimizing sensitivity is to change the default cut-off level of $p = 0.5$. The costs for elevating sensitivity by decreasing the default cut-off are subject of our current work. Preliminary results indicate that classification models showing optimised sensitivity of 100% have to accept a 2 to 3 fold increase of FP cases.

To sum it up, the top three machine learning techniques, LRA (as discussed above), SVM and ANN, delivered results of high predictive power when running on full as well as on reduced feature dimensionality. Although SVMs can effectively construct nonlinear decision boundaries by mapping training data into a higher-dimensional feature space (SVM-2, SVM-3), these polynomial SVM did not perform better in low dimensional feature spaces compared to the known linear techniques like LRA. Interestingly, SVM operating with a linear separating hyperplane (SVM-1) performed better than the polynomial ones. They led to results similar to the LRA classifier, an observation already described by other authors (Dreiseitl *et al.*, 2001). The C4.5 DT classifier, which selects features internally based on the information gain, showed good discriminatory performance, leading to the same S_n (95.3%) as LRA on PKU data (tree root = Phe, no child nodes) and slightly decreased S_n (-1.6%) on MCADD data (tree root C10:1, two child nodes C8 and C16). The third directly interpretable method, DA, operates on a separating linear hyperplane similar to LRA. As expected, the DA classifier discriminated worse in both, full and reduced feature spectrum, since an important precondition, the homogeneity of both covariance matrices, was not fulfilled (confer e.g. the data distribution of Phe in the PKU and control groups; see table 1). Out of the group of not directly interpretable techniques the ANN classifier performed best. Despite lacking direct interpretation of the knowledge representation, its ability to calculate non-linear decision boundaries emphasizes its diagnostic potential. The results of the k-NN algorithms (for all applied *k* values) were comparable with those running on established diagnostic metabolites, but significantly inferior (2 to 10 percentage points) to those running on the entire dimensionality of PKU and MCADD databases. However, *k* values larger than 5 generally led to a decrease of the classification accuracy due to smoothing effects of local data characteristics.

In conclusion, our results show that the use of machine learning paradigms, in particular the LRA model, is suitable to construct classifiers on high-dimensional metabolic data. Moreover, we could demonstrate that the screening model's high predictive power could be achieved by reducing the dimensionality of the parameter space using only 1-2 representative metabolites for PKU and MCADD. The mined results confirm some known patterns among the metabolites and reveal a number of novel patterns which may contribute towards a better understanding of newborn metabolism, and constitutes a significant contribution to the early recognition and therapy of metabolic diseases.

Acknowledgements

This study was generously supported by the Austrian Industrial Research Promotion Fund FFF (Grand No. HITT-10 UMIT).

Table 1: Metabolites of a single blood spot from MS/MS analysis

Amino acids (symbol)	**PKU**	**Controls**
Alanine (Ala)	421.8 ± 129.8	508.9 ± 210.7
Arginine (Arg)	333.0 ± 447.5	90.9 ± 49.7
Argininosuccinate (Argsuc)	1.17 ± 2.23	0.01 ± 0.02
Citrulline (Cit)*	24.7 ± 21.7	28.7 ± 39.9
Glutamate (Glu)	3498 ± 2485	235.9 ± 74.0
Glycine (Gly)	331.2 ± 140.0	624.2 ± 315.9
Methionine (Met)	23.5 ± 7.9	29.2 ± 12.9
Ornitine (Orn)*	80.4 ± 54.7	85.2 ± 60.7
Phenylalanine (Phe)	588.0 ± 240.4	57.9 ± 17.9
Pyroglutamate (Pyrglt)	32.1 ± 18.1	51.8 ± 31.6
Serine (Ser)	689.8 ± 362.2	400.6 ± 358.2
Tyrosine (Tyr)	58.1 ± 24.2	97.2 ± 64.2
Valine (Val)*	183.4 ± 71.8	170.6 ± 61.3
Leucine+Isoleucine (Xle)	193.1 ± 91.7	264.5 ± 107.7

Fatty acids (symbols)	**MCADD**	**Controls**
Free carnitine (C0)	26.416 ± 11.138	29.416 ± 12.087
Acetyl-carnitine (C2)	14.361 ± 7.245	6.661 ± 3.066
Propionyl-carnitine (C3)	3.209 ± 1.396	2.326 ± 1.205
Butyryl-carnitine (C4)*	0.524 ± 0.396	0.522 ± 0.309
Isovaleryl-carnitine (C5)	0.195 ± 0.122	0.160 ± 0.105
Hexanoyl-carnitine (C6)	1.990 ± 1.821	0.415 ± 0.516
Octanyl-carnitine (C8)	8.346 ± 6.558	0.223 ± 0.142
Decanoyl-carnitine (C10)	0.764 ± 0.501	0.079 ± 0.067
Dodecanoyl-carnitine (C12)*	0.166 ± 0.104	0.209 ± 0.206
Myristoyl-carnitine (C14)*	0.207 ± 0.104	0.198 ± 0.106
Hexadecanoyl-carnitine (C16)*	4.066 ± 1.615	4.413 ± 2.144
Octadecanoyl-carnitine (C18)*	0.954 ± 0.378	0.928 ± 0.394
Tiglyl-carnitine (C5:1)	0.031 ± 0.027	0.052 ± 0.067
Decenoyl-carnitine (C10:1)	0.805 ± 0.478	0.095 ± 0.059
Myristoleyl-carnitine (C14:1)	0.096 ± 0.048	0.122 ± 0.094
Hexadecenoyl-carnitine (C16:1)	0.158 ± 0.078	0.185 ± 0.104
Octadecenoyl-carnitine (C18:1)	0.743 ± 0.276	1.030 ± 0.401
Decenoyl-carnitine (C10:2)*	0.045 ± 0.029	0.051 ± 0.045
Tetradecadienoyl-carnitine (C14:2)	0.032 ± 0.019	0.055 ± 0.046
Octadecadienoyl-carnitine (C18:2)	0.108 ± 0.068	0.161 ± 0.115
Hydroxy-isovaleryl-carnitine (C5-OH)*	0.168 ± 0.101	0.159 ± 0.078
Hydroxytetradecadienoyl-carnitine (C14-OH)	0.016 ± 0.010	0.028 ± 0.024
Hydroxypalmitoyl-carnitine (C16-OH)*	0.021 ± 0.013	0.023 ± 0.016
Hydroxypalmitoleyl-carnitine (C16:1-OH)*	0.036 ± 0.021	0.043 ± 0.038
Hydroxyoleyl-carnitine (C18:1-OH)*	0.013 ± 0.010	0.016 ± 0.013
Dicarboxyl-butyryl-carnitine (C4-DC)*	0.137 ± 0.054	0.151 ± 0.077
Glutaryl-carnitine (C5-DC)	0.092 ± 0.056	0.047 ± 0.031
Methylglutaryl-carnitine (C6-DC)	0.072 ± 0.049	0.046 ± 0.040
Methylmalonyl-carnitine (C12-DC)	0.036 ± 0.036	0.096 ± 0.064

Concentrations (mean ± sd) of amino acids and fatty acids are denounced in μmol/L for PKU, MCADD and control group. Controls represent a randomised fraction of 1241 cases. The asterisk (*) indicates no significant ($P>0.05$) differences between both classes comparing by means of an unpaired significance test.

Table 2: Discriminatory performance of all six machine learning algorithms applied to full metabolite dimensionality

(a)

Classifier (PKU)	S_n (%)	S_p (%)	Acc (%)
DA	90.7	100	99.7
LRA	95.3	100	99.8
DT	95.3	99.8	99.7
1-NN	93.0	100	99.8
3-NN	90.7	100	99.7
5-NN	90.7	100	99.7
ANN	97.7	100	99.9
SVM-1	95.3	100	99.8
SVM-2	97.7	100	99.9
SVM-3	95.3	100	99.8

(b)

Classifier (MCADD)	S_n (%)	S_p (%)	Acc (%)
DA	88.9	100	99.5
LRA	93.7	98.8	98.5
DT	92.1	99.8	99.4
1-NN	88.9	99.4	98.9
3-NN	84.1	100	99.2
5-NN	82.5	100	99.2
ANN	92.1	99.7	99.3
SVM-1	93.7	99.6	99.3
SVM-2	93.7	99.8	99.5
SVM-3	93.7	99.8	99.5

Directly interpretable (DA, LRA, and DT) and not directly interpretable classifiers running on **(a)** 14 amino acids (PKU data) and on **(b)** 29 fatty acids (MCADD data). 1-NN, 3-NN and 5-NN represent the k-NN classifiers with a k value of 1, 3 and 5. Support vector machines (SVM) with a linear hyperplane are denounced as SVM-1, SVMs with polynomial kernels of degree 2 and 3 are abbreviated with the symbols SVM-2 and SVM-3.

Table 3: Discriminatory performance of all six machine learning algorithms applied to the established diagnostic markers

(a)

Classifier (PKU)	S_n (%)	S_p (%)	Acc (%)
DA	93	100	99.8
LRA	97.7	99.9	99.8
(DT)	(95.3)	(99.8)	(99.7)
1-NN	97.7	99.9	99.8
3-NN	97.7	100	99.9
5-NN	95.3	100	99.8
ANN	97.7	100	99.9
SVM-1	95.3	100	99.8
SVM-2	93	100	99.8
SVM-3	76.7	100	99.2

(b)

Classifier (MCADD)	S_n (%)	S_p (%)	Acc (%)
DA	71.4	100	98.6
LRA	95.2	99.8	99.6
(DT)	(92.1)	(99.8)	(99.4)
1-NN	93.7	99.8	99.5
3-NN	93.7	99.8	99.5
5-NN	92.1	99.8	99.5
ANN	92.1	99.8	99.4
SVM-1	93.7	99.9	99.6
SVM-2	84.1	99.9	99.2
SVM-3	60.3	100	98.1

Directly interpretable (DA, LRA, and DT) and not directly interpretable (k-NN, ANN, SVM) classifiers are including the primary diagnostic markers **(a)** Phe and Tyr for PKU and **(b)** C8 and C10:1 for MCADD. The DT learner was not examined on pre-selected metabolites due to its internal feature extraction strategy. Therefore the results are not directly comparable with the other classifiers and are depicted in brackets.

Table 4: Screening models based on LRA for classic PKU and classic MCADD

(a)

Metabolites (PKU)	S_n (%)	FN (# of cases)	S_p (%)	FP (# of cases)	PPV (%)	NPV (%)	Acc (%)	Logit of models $z=a_0+a_1m_1+\ldots+a_nm_n$
Phe, Xle	100	0	99.793	204	17.41	100	99.793	– 61.2577 + 1.8037·Phe – 1.4518·Xle
Phe, Tyr	100	0	99.775	222	16.23	100	99.775	– 211.2566 + 2.1318·Phe – 0.6224·Tyr
Phe, Val	97.67	1	99.895	103	28.966	99.999	99.894	– 11.8046 + 0.2248·Phe – 0.1210·Val
Phe, Arg	95.35	2	99.983	17	70.69	99.998	99.981	– 9.827 + 0.0462·Phe – 0.0035·Arg
Phe, Argsuc	95.35	2	99.984	16	71.93	99.998	99.982	– 10.167 + 0.0457·Phe – 0.340·Argsuc
Phe	95.35	2	99.984	16	71.93	99.998	99.982	– 10.1482 + 0.0455·Phe

(b)

Metabolites (MCADD)	S_n (%)	FN (# of cases)	S_p (%)	FP (# of cases)	PPV (%)	NPV (%)	Acc (%)	Logit of models $z=a_0+a_1m_1+\ldots+a_nm_n$
C8, C18:1	96.83	2	99.992	8	88.41	99.998	99.990	– 5.4917 + 5.7436·C8 – 2.1833·C18:1
C8	95.24	3	99.992	8	88.24	99.997	99.989	– 7.5362 + 5.7931·C8
C8, C12DC	95.24	3	99.990	10	85.71	99.997	99.987	– 4.8647 + 5.149·C8 – 40.4661·C12DC
C8, C10	95.24	3	99.989	11	84.51	99.997	99.986	– 7.6114 + 4.6649·C8 + 3.3668·C10
C8, C10:1	95.24	3	99.950	50	54.55	99.997	99.947	– 8.7572 + 4.2517·C8 + 10.888·C10:1

Screening models for **(a)** classic PKU and **(b)** classic MCADD. Sensitivity (S_n), specificity (S_p), positive predictive value (PPV), negative predictive value (NPV), accuracy (Acc), number (#) of false negatives (FN), number (#) of false positives (FP) and the logits (z) of the LRA models are denounced. The specificity of the models was re-evaluated on a randomly selected control database of 98,411 cases.

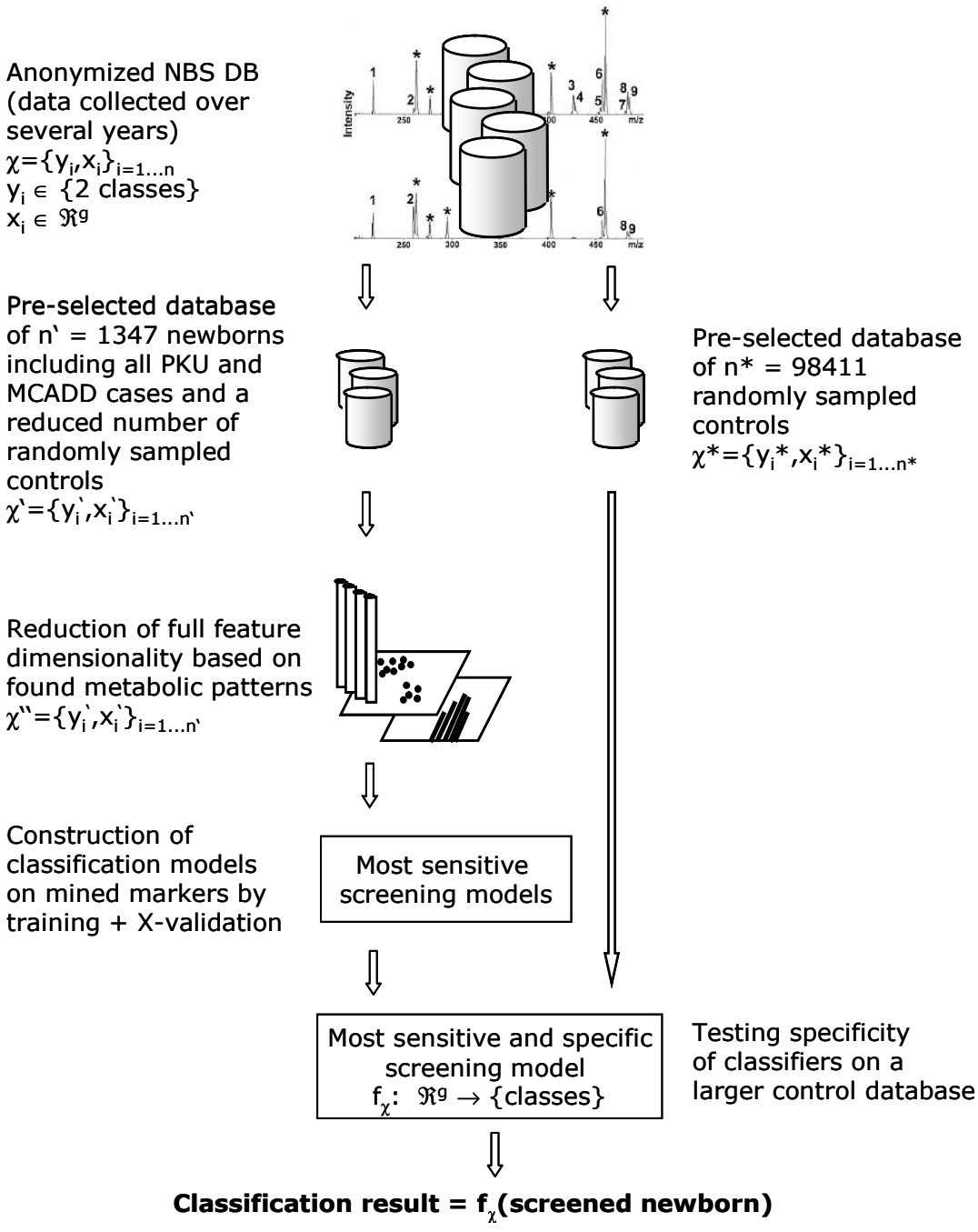

Figure 1: General process of data analysis for constructing a screening model on high dimensional metabolic data. Starting from the newborn screening database (NBS DB) several intermediate data mining steps resulted in a classification model with optimised sensitivity and specificity. χ describes all available tuples of the database containing the measured metabolites (x_i) and the flag for the class membership (y_i). f_χ is the formula for the final model.

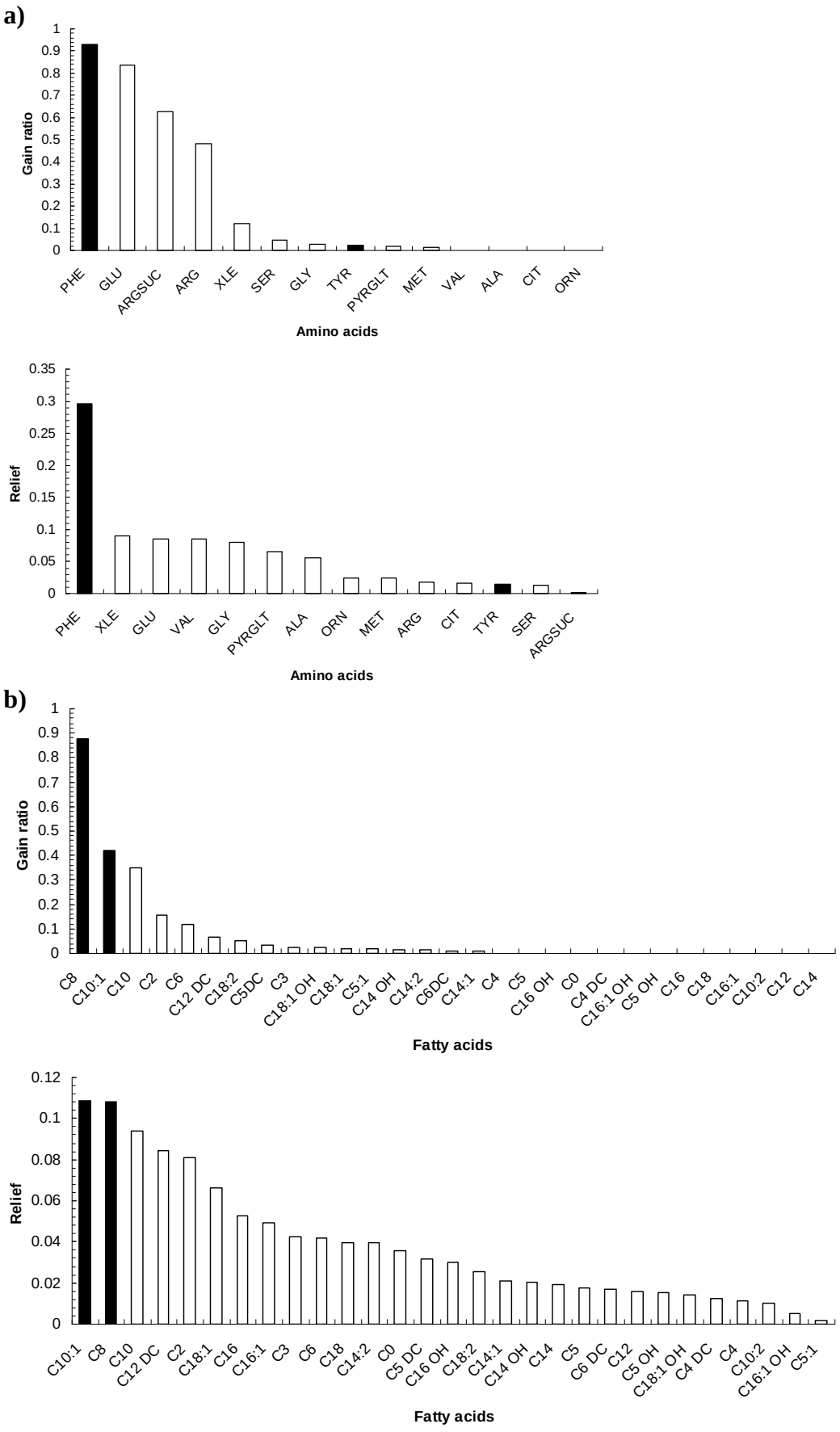

Figure 2: Metabolic patterns examined with gain ratio and relief filter algorithms on **(a)** full amino acid dimensionality (14 acids, PKU) and on **(b)** full fatty acid dimensionality (29 acids, MCADD). Black bars indicate the established primary diagnostic markers for PKU and MCADD (ACMG/ASHG statement, 2000).

References

American College of Medical Genetics/American Society of Human Genetics Test and Technology Transfer Committee Working Group. (2000) Tandem mass spectrometry in newborn screening. *Genet Med*, **2**, 267-269.

Baumgartner C, Baumgartner D, Böhm C. (2004) Classification on high dimensional metabolic data: Phenylketonuria as an example. *IASTED Proc. 2nd Int. Conf. on Biomedical Engineering (BioMED 2004),* Innsbruck, Austria, pp 357-360.

Bishop, C.M.(1995) *Neural networks for pattern recognition*, Oxford university press, Oxford.

Blau, N., Thony, B., Cotton, R.G.H., Hyland, K. (2001) Disorders of tetrahydrobiopterin and related biogenic amines. In: Scriver CR, Kaufman S, Eisensmith E, Woo SLC, Vogelstein B, Childs B (eds) *The Metabolic and Molecular Bases of Inherited Disease*, 8 ed. McGraw Hill, New York, Ch. 78.

Chace, D.H., Millington, D.S., Terada, N., Kahler, S.G., Roe, C.R., Hofman, L.F. (1993) Rapid diagnosis of phenylketonuria by quantitative analysis for phenylalanine and tyrosine in neonatal blood spots by tandem mass spectrometry. *Clin Chem*, **39**, 66 –71.

Chace, D.H., Sherwin, J.E., Hillman, S.L., Lorey, F., Cunningham, G.C. (1998) Use of phenylalanine-to-tyrosine ratio determined by tandem mass spectrometry to improve newborn screening for phenylketonuria of early discharge specimens collected in the first 24 hours. *Clin Chem*, **44**, 2405-2409.

Chace, D.H., DiPerna, J.C., Naylor, E.W. (1999) Laboratory integration and utilization of tandem mass spectrometry in neonatal screening: a model for clinical mass spectrometry in the next millennium. *Acta Paediatr (Suppl)*, **88**, 45-47.

Cortes, C., Vapnik, V.(1995) Support vector networks. *Machine learning*, **20**, 273-297.

Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., Binder, M. (2001) A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *J Biomed Inf*, **34**, 28-36.

Guldberg, P., Rey, F., Zschocke, J., Romano, V., Francois, B., Michiels, L., Ullrich, K., Hoffmann, G.F., Burgard, P., Schmidt, H., Meli, C., Riva, E., Dianzani, I., Ponzone, A., Rey, J., Guttler, F. (1998) A European multicenter study of phenylalanine hydroxylase deficiency: classification of 105 mutations and a general system for genotype-based prediction of metabolic phenotype. *Am J Hum Genet*, **63**, 71-79

Hall, M.A. (1999) Correlation-based feature selection for machine learning. PhD thesis at the University of Waikato, New Zealand.

Hosmer, D.W., Lemeshow, S. (2000) *Applied logistic regression*, 2nd edition, Wiley, New York.

Kira, K., Rendell, L.A. (1992) A practical approach to feature selection. In *Machine Learning: Proceedings of the Ninth International Conference*, Aberdeen, Scotland, pp 249-256.

Kononenko, I. (1995) On biases in estimating multi-valued attributes. In *IJCAI'95*, Montreal, Canada, pp 1034–1040.

Liebl, B., Nennstiel-Ratzel, U., von Kries, R., Fingerhut, R., Olgemoller, B., Zapf, A., Roscher, A.A. (2002) Very high compliance in an expanded MS-MS-based newborn screening program despite written parental consent. *Prev Med,* **34**, 127-131.

Liebl, B., Nennstiel-Ratzel, U., von Kries, R., Fingerhut, R., Olgemoller, B., Zapf, A., Roscher, A.A. (2002) Expanded newborn screening in Bavaria: tracking to achieve requested repeat testing. *Prev Med,* **34**, 132-137.

Liebl, B., Nennstiel-Ratzel, U., Roscher, A.A., von Kries, R. (2003). Data required for the evaluation of newborn screening programmes. *Eur J Pediatr*, **162** Suppl 1, 57-61.

McLachlan, G.J. (1992) *Discriminant analysis and statistical pattern recognition*. Wiley, New York.

Mendes, P. (2002) Emerging bioinformatics for the metabolome. *Brief Bioinform,* **3**, 134-145.

Millington, D.S., Roe, C.R., Maltby, D.A. (1984) Application of high resolution fast atom bombardment and constant B/E ratio linked scanning to the identification and analysis of acylcarnitines in metabolic disease. *Biomed Mass Spectrom,* **11**, 236-241.

Millington, D.S., Terada, N., Kodo, K., Chace, D.H. (1992) A review: carnitine and acylcarnitine analysis in the diagnosis of metabolic diseases: advantages of tandem mass spectrometry. In: Matsumoto, I., editor. *Advances in chemical diagnosis and treatment of metabolic disorders*, Vol 1. New York, John Wiley & Sons, pp 59-71.

Mitchell, T.M. (1997) *Machine Learning*, McGraw-Hill Boston, MA.

National Center for Biotechnology Information. Online Mendelian Inheritance in Man (OMIM), http://www3.ncbi.nlm.nih.gov/Omim.

Neville, P., Tan P.Y., Mann, G., Wolfinger, R. (2003) Generalizable mass spectrometry mining used to identify disease state biomarkers from blood serum. *Proteomics*, **3**, 1710-1715.

Purohit, P.V., Rocke, D.M. (2003) Discriminant models for high-throughput proteomics mass spectrometer data. *Proteomics,* **3**, 1699-1703.

Quinlan, R.J. (1986) Induction of decision trees, *Machine Learning*, **1**, 81-106.

Quinlan, R.J. (1993) *C4.5: Program for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Rashed, M.S., Ozand, P.T., Bucknall, M.P., Little, D. (1995) Diagnosis of inborn errors of metabolism from blood spots by acylcarnitines and amino acids profiling using automated electrospray tandem mass spectrometry. *Pediatr Res,* **38**, 324-331.

Rinaldo, P., Matern, D., Bennett, M.J. (2002) Fatty acid oxidation disorders. *Annu Rev Physiol,* **64**, 477-502.

Salzberg, S. (1999) On comparing classifiers: a critique of current research and methods. *Data mining and knowledge discovery*, **1**, 1-12.

Van Hove, J.L., Zhang, W., Kahler, S.G., Roe, C.R., Chen, Y.T., Terada, N., Chace, D.H., Iafolla, A.K., Ding, J.H., Millington, D.S. (1993) Medium-chain acyl-CoA dehydrogenase (MCAD) deficiency: diagnosis by acylcarnitine analysis in blood. *Am J Hum Genet*, **52**, 958–966.

Vapnik, V. (1998) *Statistical Learning Theory*, Wiley, New York.

Witten, I.H., Frank, E. (2000) *Data Mining - Practical machine learning tools and techniques with java implementations*. Morgan Kaufmann, San Francisco.