Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 38 (2005) 89–98
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Modelling of classification rules on metabolic patterns including machine learning and expert knowledge

Christian Baumgartner[a,*], Christian Böhm[b], Daniela Baumgartner[c]

[a] Research Group for Biomedical Data Mining, Institute for Information Systems, University for Health Sciences, Medical Informatics and Technology, Innrain 98, A-6020 Innsbruck, Austria
[b] Institute for Computer Science, University of Munich, Oettingenstrasse 67, D-80538 Munich, Germany
[c] Department of Pediatrics, Innsbruck Medical University, Anichstrasse 35, A-6020 Innsbruck, Austria

Received 8 March 2004
Available online 11 November 2004

Abstract

Machine learning has a great potential to mine potential markers from high-dimensional metabolic data without any a priori knowledge. Exemplarily, we investigated metabolic patterns of three severe metabolic disorders, PAHD, MCADD, and 3-MCCD, on which we constructed classification models for disease screening and diagnosis using a decision tree paradigm and logistic regression analysis (LRA). For the LRA model-building process we assessed the relevance of established diagnostic flags, which have been developed from the biochemical knowledge of newborn metabolism, and compared the models' error rates with those of the decision tree classifier. Both approaches yielded comparable classification accuracy in terms of sensitivity (>95.2%), while the LRA models built on flags showed significantly enhanced specificity. The number of false positive cases did not exceed 0.001%.

Keywords: Machine learning; Classification rules; Metabolic patterns; Expert knowledge; Metabolic disorders

1. Introduction

Newborn screening permits the detection of metabolic disorders in newborns during the first few days of life prior to the manifestation of symptoms [1–3]. Due to recent innovations and refinements of the screening methodology using modern tandem mass spectrometry (MS/MS) more than 20 inherited metabolic disorders can be detected simultaneously from a single blood spot by quantifying concentrations of up to 50 metabolites [4,5]. Machine learning techniques offer an obvious and promising approach to examine high-dimensional metabolic data, where manual analysis is tedious and time-consuming due to the great number and complexity. The investigation of novel metabolic patterns, the construction of classification models with high diagnostic prediction and the discovery of new clues for unknown causal relations lead to a better understanding of mined data in metabolic networks and constitutes a significant contribution to preventive medicine [6].

Our goal was to investigate high-dimensional metabolic data with respect to three severe inborn errors of metabolism to construct classification models for disease screening and diagnosis. In particular, we focused on (i) phenylalanine hydroxylase deficiency (PAHD), an amino acid disorder, which includes cases of classic phenylketonuria (PKU, OMIM No. 261600 [7]) and hyperphenylalaninemia (non-PKU HPA, OMIM No. 264070) [8,9], (ii) medium-chain acyl-CoA dehydrogenase deficiency (MCADD, OMIM No. 201450), a fatty acid oxidation defect [10], and (iii) 3-methylcrotonyl CoA carboxylase deficiency (3-MCCD, OMIM No. 210200), an organic acid disorder [11,12].

* Corresponding author. Fax: +43 50 8648 673827.
E-mail address: christian.baumgartner@umit.at (C. Baumgartner).

1532-0464/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2004.08.009

For the model-building process we applied two directly interpretable classification algorithms, i.e., the C4.5 decision tree paradigm and binary logistic regression analysis (LRA), to a metabolome training dataset. Decision trees optimize classification accuracy by reducing the full feature dimensionality to a relevant feature subset according to the algorithms' internal feature selection strategy. The aim is to assess the relevance of metabolic knowledge for the model-building process by comparing C4.5s feature subset with established diagnostic flags which have been developed from the current biochemical knowledge of abnormalities in newborn metabolism. For this task we built LRA models on these flags and benchmarked their error rates with those of the decision tree classifier which does not require any a priori knowledge for tree construction.

2. Metabolic data

Metabolites analyzed by modern MS/MS employing appropriate internal standards can be quantified very rapidly, sensitively and accurately requiring only minimal sample preparation [1]. For MS/MS analysis a single blood sample, which has been taken within few days after the newborn's birth, is sufficient. This screening methodology creates a high-dimensional metabolic dataset of each newborn including concentration values of more than 40 metabolites (14 amino acids and 29 fatty acids, see Table 1).

Our experimental datasets were anonymously provided from the newborn screening program of the State of Bavaria, Germany, between 1999 and 2002. For our train-and-test design cycle we focused on one representative disorder of the amino acid, one of the fatty acid oxidation, and one of the organic acid metabolism, each of them showing a relatively high incidence in-between their group of disorders (PAHD, $n = 94$ cases including 43 cases of classic PKU and 51 cases of non-PKU HPA, classic MCADD, $n = 63$ cases, and 3-MCCD, $n = 22$ cases). Unfortunately, the number of cases of further screened metabolic disorders was too small ($n < 5$ cases) for useful examination.

Based on the given number of PAHD, MCADD, and 3-MCCD cases we sampled a statistically representative control group from the newborn screening (NBS) database (~600,000 entries, end of year 2002) using a rate of ~1:15–50 (disorder to controls). The PAHD sub-database thus contains all 94 cases designated as confirmed PAHD and 1241 randomly sampled controls, i.e., each 500th case from NBS controls. The MCADD and 3-MCCD sub-databases consist of all 63 and 22 cases designated as classic MCADD and 3-MCCD, and again 1241 controls. In our study population "controls" represent individuals without verified cases of known metabolic disorders.

A much larger randomly sampled control database of 98,411 cases, i.e., one sixth of the NBS control database, serves to estimate the real specific power of our constructed models. Table 2 gives a short clinical overview of the investigated disorders (enzyme defects, symptoms, proposed treatments, and diagnostic tests) [7,13–15].

3. Process of data analysis

3.1. Overview of data mining steps

The data analysis process constructing classification models on high-dimensional metabolic data is illustrated in Fig. 1. We constructed classification models using C4.5 decision tree paradigm and LRA. Both models were trained and 10-fold-cross validated according to a two-class problem on a training dataset containing n

Table 1
Overview of metabolites measured from MS/MS analysis

Amino acids (symbols)	Fatty acids (symbols)	Fatty acids (symbols)
Alanine (Ala)	Free carnitine (C0)	Hexadecenoyl-carnitine (C16:1)
Arginine (Arg)	Acetyl-carnitine (C2)	Octadecenoyl-carnitine (C18:1)
Argininosuccinate (Argsuc)	Propionyl-carnitine (C3)	Decenoyl-carnitine (C10:2)
Citrulline (Cit)	Butyryl-carnitine (C4)	Tetradecadienoyl-carnitine (C14:2)
Glutamate (Glu)	Isovaleryl-carnitine (C5)	Octadecadienoyl-carnitine (C18:2)
Glycine (Gly)	Hexanoyl-carnitine (C6)	Hydroxy-isovaleryl-carnitine (C5-OH)
Methionine (Met)	Octanyl-carnitine (C8)	Hydroxytetradecadienoyl-carnitine (C14-OH)
Ornitine (Orn)	Decanoyl-carnitine (C10)	Hydroxypalmitoyl-carnitine (C16-OH)
Phenylalanine (Phe)	Dodecanoyl-carnitine (C12)	Hydroxypalmitoleyl-carnitine (C16:1-OH)
Pyroglutamate (Pyrglt)	Myristoyl-carnitine (C14)	Hydroxyoleyl-carnitine (C18:1-OH)
Serine (Ser)	Hexadecanoyl-carnitine (C16)	Dicarboxyl-butyryl-carnitine (C4-DC)
Tyrosine (Tyr)	Octadecanoyl-carnitine (C18)	Glutaryl-carnitine (C5-DC)
Valine (Val)	Tiglyl-carnitine (C5:1)	Methylglutaryl-carnitine (C6-DC)
Leucine + Isoleucine (Xle)	Decenoyl-carnitine (C10:1)	Methylmalonyl-carnitine (C12-DC)
	Myristoleyl-carnitine (C14:1)	

Fourteen amino acids and 29 fatty acids are analyzed from a single blood spot using MS/MS. The concentrations are given in μmol/L.

Table 2
Clinical review of investigated metabolic disorders

Disorder	Enzyme defect	Symptoms if untreated	Treatments	Diagnostic tests
Phenylalanine hydroxylase deficiency (PAHD) including: classic PKU + non-PKU HPA	Phenylalanine hydroxylase, biopterin synthesis disorders	Microcephaly, mental retardation, seizures, autistic-like behavior, and fair-light complexion, hair color and eye color; "mousy/musty" odor (PKU) Low risk of developing intellectual, neurological and neurophysiological impairment (non-PKU HPA)	Phenylalanine restriction, tyrosine supplementation (special medical diet), tetrahydrobiopterin supplementation in some (PKU), dietary treatment discussed (non-PKU HPA)	Plasma amino acids, urine pterin studies, bloodspot DHPR assay (PKU + non-PKU HPA) molecular genetic testing for carrier status
Medium chain acyl CoA dehydrogenase deficiency (MCADD)	Deficiency of medium chain acyl CoA dehydrogenase	Fasting intolerance, hypoglycemia, hyperammonemia, acute encephalopathy, cardiomyopathy, liver failure	Avoidance of fasting, aggressive intervention when hypoglycemia impending, carnitine supplementation, regulation of dietary fat intake	Plasma acylcarnitine profile, urine organic acids, plasma carnitine, serum chemistry panel, molecular genetic testing
3-Methylcrotonyl CoA carboxylase deficiency (3-MCCD)	Deficiency of the enzyme 3-methylcrotonyl CoA carboxylase	Metabolic acidosis and hypoglycemia, some may be asymptomatic	Low protein diet, Leucine restricted diet, carnitine supplementation, glycine supplementation	Plasma acylcarnitine profile, urine organic acids, plasma amino acids, plasma carnitine, serum chemistry panel, molecular genetic testing

PKU, phenylketonuria; HPA, hyperphenylalaninemia; DHPR, dihydropteridine reductase. See sources [7,13–15].

cases of disorder (PAHD: $n = 94$, MCADD: $n = 63$ and 3-MCCD: $n = 22$) and a randomly sampled control group ($n = 1241$). The decision tree classifier optimises classification accuracy by reducing full metabolite dimensionality, while the LRA model was constructed on the biochemical knowledge—using diagnostic flags as input variables—with the goal to assess the metabolic knowledge for the model-building process. Finally, the models were validated on a larger dataset of ~100,000 controls to estimate the false positive rate representative for a real screening population.

Although the use of separate datasets for the model-building process is preferred when enough data is available (see larger control database), the number of cases designated as a metabolic disorder was too small to separate them into training and test sets so that cross-validation was proposed. We also tested the classifiers trained from the different disorders against each other by validating each model with all cases of disorder which the model was not built on (e.g., the PAHD model was validated on all cases of MCADD and 3-MCCD to demonstrate the model's specific behaviour for the screened disease it was developed for).

3.2. Classification algorithms

Decision trees. Decision trees are rooted, usually binary trees, with simple classifiers placed at each internal node and a class label at each leaf. These simple classifiers associated with the internal nodes are in the majority of algorithms comparisons between an input variable and a fix value. Decision trees are generally trained by means of a top down growth procedure, which starts from the root node and greedily chooses a split of the data that maximizes some cost function, usually a measure of the class purity of the two subgroups defined by the split. After choosing a split, the subgroups are mapped to the two child nodes. This procedure is then recursively applied to the children, and the tree grows until some stopping criterion is met. If the resulting tree is too complex (and, therefore, often overfitted) some of the branches can be pruned. The algorithm most often used to generate decision trees is C4.5 and C5.0, respectively. These algorithms select the next node to place in the tree by computing the information gain (IG) for all candidate features. Information gain 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 [16–20].

$$E(S) = \sum_{c \in C} -\frac{|S_c|}{|S|} . \ln \frac{|S_c|}{|S|} \tag{1}$$

$$\mathrm{IG}(S, A) = E(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} . E(S_v). \tag{2}$$

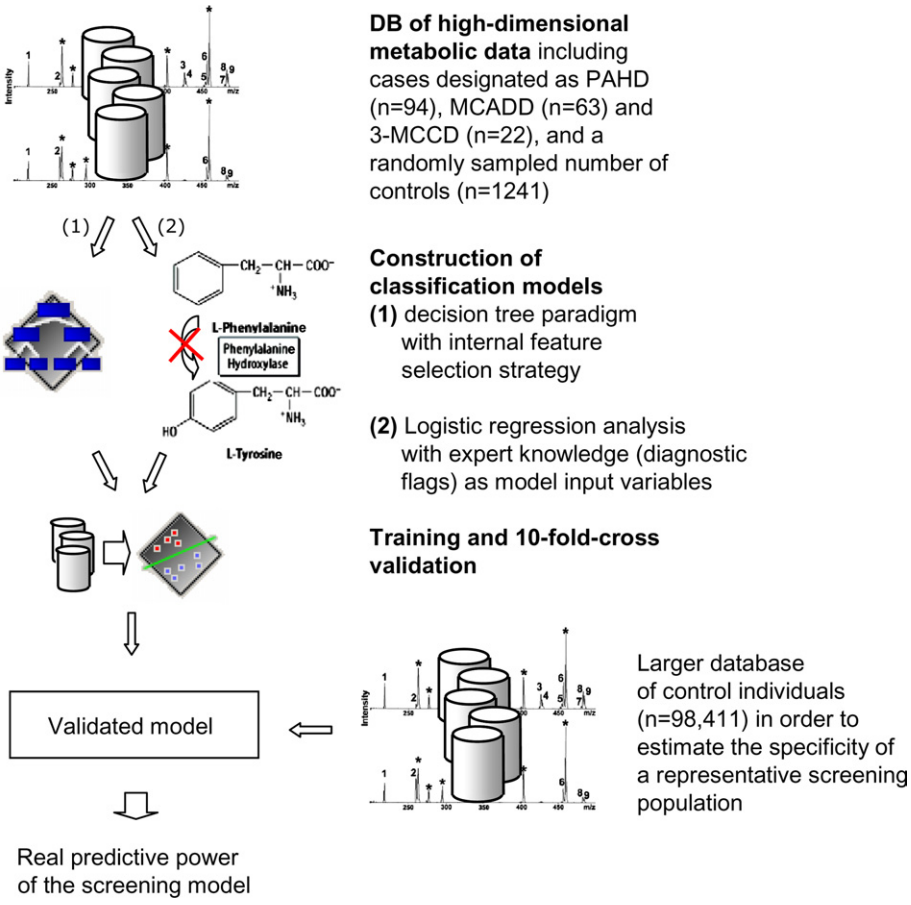

Fig. 1. Experimental design of model-building on high-dimensional metabolic data. Several intermediate data mining steps are performed resulting in a classification model with high diagnostic prediction.

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. For our experiments we used C4.5 for tree construction with pruning option.

Logistic regression analysis. We constructed classification models on diagnostic flags using logistic regression analysis, which is widely used in medical applications. LRA constructs a linear separating hyperplane between two datasets (cases of disorder and controls) which have to be distinguished by the classifiers. This hyperplane is mathematically described by a linear discriminant function $z = f(x_1, \ldots, x_n) = b_1x_1 + b_2x_2 + \cdots + b_nx_n + c$. Here, $x_1, \ldots, x_n$ are the input variables. The coefficients $b_1, \ldots, b_n$ and the constant c have to be learned by the method.

The distance from the hyperplane is considered as probability measure of class membership based on a so-called logistic function $p = 1/(1 + e^{-z})$, where p is the conditional probability of the form $P(z = 1|x_1, \ldots, x_n)$ and z the logit (discriminant function) of the model. The class membership is indicated by a cut-off value ($p = 0.5$ by default, $p < 0.5$ classifies controls and $p \geqslant 0.5$ cases of disorder). LRA uses a maximum likelihood method which maximizes the probability of getting the observed results given the fitted coefficients [21].

3.3. Diagnostic flags

In NBS labs diagnostic flags are routinely used to pre-screen newborns that are highly suspicious for the screened disorders. The procedure how these flags have been modelled is briefly sketched by an example.

In the abnormal PAHD metabolism the essential amino acid Phe can not be metabolized to Tyr due to a blockade of the enzyme phenylalanine hydroxylase. Therefore, Phe shows strongly elevated concentration levels accompanied by slightly decreased Tyr concentrations. The diagnostic flags for PAHD contain the key marker Phe showing the most significant concentration changes and additionally the ratio of Phe/Tyr representing the block of the enzyme phenylalanine hydroxylase. For our experiments we used flags, which were developed by biochemical and medical experts of the Bavarian newborn screening program, as summarized in Table 3. NBS centres worldwide use such decision rules with slight modifications for disease screening [22,23].

Table 3
Statistical analysis of metabolic patterns investigated from decision tree and expert knowledge approach

Metabolic disorders	Metabolites filtered from decision tree approach						Diagnostic flags developed from expert knowledge					
	Metabolites	Controls (μmol/L)	CV	Disorder (μmol/L)	CV	Disorder/controls (%)	Flags	Controls (μmol/L)	CV	Disorder (μmol/L)	CV	Disorder/controls
Phenylalanine hydroxylase deficiency (PAHD)	Phe (root)	57.9 (17.1)	0.31	406 (242)	0.59	+601	Phe	*	*	*	*	*
	Tyr	97.2 (64.2)	0.66	67.6 (27.6)	0.41	−31	Phe/Tyr	0.74 (0.40)	0.54	7.2 (6.1)	0.84	+876
	Ala	509 (211)	0.41	444 (185)	0.42	−13						
Medium chain acyl CoA dehydrogenase deficiency (MCADD)	C8	0.22 (0.14)	0.64	8.35 (6.56)	0.79	+3649	C8	*	*	*	*	*
	C10:1 (root)	0.09 (0.06)	0.63	0.08 (0.48)	0.59	+750	C10:1	*	*	*	*	*
	C16	4.41 (2.14)	0.49	4.07 (1.61)	0.40	−8	C6	0.41 (0.52)	1.24	1.99 (1.82)	0.91	+380
							C8/C10	5.02 (6.09)	1.21	13.0 (10.4)	0.8	+159
							C8/C6	0.88 (0.89)	1.01	4.34 (2.16)	0.5	+390
3-Methylcrotonyl CoA carboxylase deficiency (3-MCCD)	C5-OH (root)	0.16 (0.08)	0.49	4.68 (3.15)	0.67	+2941	C5-OH	*	*	*	*	*
	C6	0.41 (0.52)	1.24	0.15 (0.11)	0.70	−63	C5-OH/C3	0.09 (0.07)	0.80	3.91 (4.27)	1.09	+4351

Mean (SD) concentration values of metabolites (μmol/L) and coefficient of variation (CV, which is standard deviation divided by the mean value of metabolites' concentrations) are given for both classes controls and disorder. Relative changes of markers' mean concentrations referenced to mean control values (Disorder/controls) are denounced in percent (%). The asterisk (*) represents equal values as already shown.

3.4. Classification accuracy and validation process

The discriminatory power of the models was evaluated from stating all true positive (TP), true negative (TN), false positive (FP), and false negative (FN) cases. We used sensitivity (S_n), specificity (S_p), positive predictive value (PPV) and accuracy (Acc = (TP + TN)/all cases) as measure of classification accuracy.

For both approaches the model-building process was validated using 10-fold-cross validation on our training datasets which has been proved to be statistically good enough in evaluating the classification accuracy of the models [18]. Traditionally, 10-fold-cross validation uses "train on 9-folds, test on onefold." For our experiments we used the publicly available, widespread, and comprehensive data mining tool set WEKA (http://www.cs.waikato.ac.nz/~ml/weka) to perform classification and cross-validation. The WEKA tool constructs LRA models based on the default cut-off of $p = 0.5$ [18].

4. Metabolic patterns

Metabolic patterns mined from the decision tree paradigm—the nodes of the generated tree represent the relevant feature subset—and basic statistical analysis are summarized in Table 3. The detected metabolites (tree nodes) are well comparable to the established diagnostic metabolites/flags representing the present metabolic knowledge. In particular, the mined metabolites Phe and Tyr (PAHD), C8 and C10:1 carnitines (MCADD) and C5-OH carnitine (3-MCCD) represent the established markers known from the disorders' metabolism [15]. However, the decision tree paradigm only selects single metabolites with highest information gain, whereas diagnostic flags also take interacting metabolites (ratios) into account. Easy statistical analysis as illustrated in Table 3 showed that the ratio Phe/Tyr, for instance, further increases the relative concentration changes compared to the single markers (Phe/Tyr +876% vs. Phe +601% and Tyr −31% alone, Table 3). However, all presented ratios (Phe/Tyr, C8/C10, C8/C6, and C5-OH/C3) revealed an enhanced coefficient of variation (CV), predominantly observed in the control groups. It is of interest to note that each disorder is characterized by one key metabolite i.e., Phe for PAHD, C8 for MCADD and C5-OH for 3-MCCD showing a significant concentration enhancement with a moderate CV (<1).

5. Screening models

Following the process of data analysis decision tree and LRA models were trained and 10-fold-cross validated on pre-selected training datasets. Classification

Table 4a
Classification accuracy of decision tree and LRA models trained and cross-validated on a small training set

Disorder (No. of cases)	Classification model	Model input variables	S_n (%)	FN (No. of cases)	S_p (%)	FP (No. of cases)	Acc (%)	Tree structure of C4.5 classifier/logit of LRA models $z = a_0 + a_1\, m_1 + \cdots + a_n\, m_n$	Standard error of coefficients	RMS error
PAHD ($n = 94$)	Decision tree classifier	Phe Tyr Ala	96.809	3	99.758	3	99.551	Phe < = 115.58: control Phe > 115.58 | Ty r < = 95.92: PAHD | Tyr > 95.92 | | Ala < = 686.13: PAHD | | Ala > 686.13: control		0.0678
	LRA built on diagnostic flags	Phe Phe/Tyr	96.809	3	99.758	3	99.551	−14.776 +0.0738 · Phe +2.2157 · Phe/Tyr	2.622 0.015 0.519	0.0598
		Phe*	93.617	6	99.678	4	99.251	−11.8681 +0.081 · Phe	1.429 0.011	0.0687
MCADD ($n = 63$)	Decision tree classifier	C8 C10:1 C16	95.238	3	99.517	6	99.310	C10:1 <= 0.37 | C8 <= 0.46: control | C8 > 0.46 | | C16 <= 3.16: MCADD | | C16 > 3.16 | | | C8 <= 1.48: control | | | C8 > 1.48: MCADD C10:1 > 0.37: MCADD		0.0813
	LRA built on diagnostic flags	C8 C10:1 C6 C8/C10 C8/C6	95.238	3	99.678	4	99.463	−8.5391 +7.1856 · C8 +11.0392 · C10:1 −1.6439 · C6 −0.1836 · C8/C10 −0.1774 · C8/C10	2.394 2.865 4.092 2.893 0.060 1.837	0.0658
		C8*	95.238	3	99.839	2	99.617	−7.5362 +5.7931 · C8	0.746 0.764	0.0595
3-MCCD ($n = 22$)	Decision tree classifier	C5-OH C6	95.455	1	99.839	3	99.683	C5–OH <= 0.41: Control C5–OH > 0.41 | C6 <= 0.37: 3-MCCD | C6 > 0.37: control		0.0563
	LRA built on diagnostic flags	C5-OH C5-OH/C3	90.909	2	99.919	1	99.762	−722.52 +1054.41 · C5-OH −248.09 · C5-OH/C3	3935.91 5719.10 1348.61	0.0486
		C5-OH*	95.455	1	99.919	1	99.942	−47.349 +63.205 · C5-OH	48.291 65.095	0.0404

n, number of PAHD, MCADD and 3-MCCD cases, No, number of FN and FP cases. Classification results are given in terms of sensitivity (S_n), specificity (S_p) and accuracy (Acc). The generated C4.5 tree structures as well as the logit z of the LRA models and the standard error of their coefficients are given. For both algorithms the root mean squared (RMS) error, which is a measure of success of numeric prediction, is shown. The asterisk (*) highlights those LRA models built on the disease's predominant metabolite (see Table 3, Disorder/controls). Tenfold-cross validation was applied to validate both classifiers.

accuracy in terms of sensitivity (S_n), specificity (S_p) and accuracy (Acc) is given in Table 4a. In addition to the parameters S_n, S_p and PPV we also denounce the generated tree structure as well as the logit of the LRA model and its standard error of the coefficients, and for both approaches the root mean squared (RMS) error, a measure of success of numeric prediction. Both the decision tree and LRA model for PAHD showed equal classification accuracy ($S_n = 96.809\%$, $S_p = 99.758\%$, and Acc = 99.551%). The MCADD models' error rates were comparable to the PAHD ones. Only the decision tree model yielded a slightly reduced specificity. The 3-MCCD models achieved the highest classification accuracy showing only small alterations in sensitivity and specificity.

However, to estimate specificity and positive predictive value (PPV) representative for a real screening population we validated our models on a larger control database of approximately 100,000 cases. All validation results based on a real screening population are summarized in Table 4b. As expected, the specificity of both models decreased, while the number of false positive cases of the decision tree models was 6–16th fold reduced compared to that of the LRA/flag approach. For instance, considering Phe, and Phe/Tyr as LRA model input variables, the most specific PAHD model was established showing a S_p of 99.905% and a PPV of 49.5%, respectively. In other words, only a fraction of 93 FP cases (i.e., 0.0009% out of 98,411 controls) was wrongly classified. Testing the PAHD model with all MCADD and 3-MCCD cases no additional false negative cases were observed, because abnormal PAHD metabolism only alters amino acid, but not fatty acid concentrations. In analogy, no false negative cases were observed when checking the MCADD or 3-MCCD models with all PAHD/3-MCCD or PAHD/MCADD cases, respectively.

For MCADD our results yielded a different picture. Constructing a model on the diagnostic flags (model inputs are three single metabolites and two ratios) the number of FP cases could be reduced from 915 to 55 false positive cases compared to the decision tree classifier. The best classification accuracy ($S_n = 95.2\%$, $S_p = 99.992\%$, PPV = 88.2%) was obtained by building the model solely on the predominant marker C8 carnitine which basically can be explained by the dramatic concentration changes of C8 caused by the erroneous fatty acid metabolism of MCADD (cf. increase of mean concentration level of +3649%, Table 3). The flags C6, C10:1, C8/C10, and C8/C6 yielded a promising impact (mean concentration changes 159–750%) and moderate CV in the MCADD class (<0.92), but they were not able to further increase specificity. However, the false positive rates for MCADD between different screening programs most likely vary because of differences in acylcarnitine analysis and profiling. C8, the predominant, but not specific marker for MCADD, which is elevated in several other disorders (e.g.,

Table 4b
Classification accuracy of decision tree and LRA models validated on a larger control database

Disorder (No. of cases)	Classification model	Model input variables	S_n (%)	FN (No. of cases)	S_p (%)	FP (No. of cases)	PPV (%)	Acc (%)
PAHD ($n = 94$)	Decision tree classifier	Phe Tyr Ala	96.809	3	99.404	587	13.41	99.401
	LRA built on diagnostic flags	Phe Phe/Tyr	96.809	3	99.905	93	49.46	99.903
		Phe*	95.745	4	99.885	113	44.34	99.881
MCADD ($n = 63$)	Decision tree classifier	C8 C10:1 C16	95.238	3	99.070	915	6.15	99.068
	LRA built on diagnostic flags	C8 C10:1 C6 C8/C10 C8/C6	95.238	3	99.944	55	52.17	99.941
		C8*	95.238	3	99.992	8	88.24	99.989
3-MCCD ($n = 22$)	Decision tree classifier	C5-OH C6	95.455	1	99.627	367	5.41	99.626
	LRA built on diagnostic flags	C5-OH C5-OH/C3	90.909	2	99.955	44	33.33	99.955
		C5-OH*	95.455	1	99.957	42	33.33	99.956

n, number of PAHD, MCADD, and 3-MCCD cases, No. number of FN and FP cases. Classification results are given in terms of sensitivity (S_n), specificity (S_p), positive predictive value (PPV), and accuracy (Acc). S_p, PPV and Acc were re-evaluated on a larger database of ~100,000 controls to estimate specificities of a representative screening population. S_n remain unchanged (see results Table 4a). The asterisk (*) highlights the LRA models built on the disease's predominant metabolite (see Table 3, Disorder/controls).

medium/short chain 3-OH acyl CoA dehydrogenase deficiency), can be used to pre-screen several fatty oxidation disorders. Both presented models including additional markers to C8 are specific for MCADD which is important when physicians consider differential diagnostic aspects.

The classification models for 3-MCCD differed significantly from that of PAHD and MCADD as both approaches showed the lowest classification error rates. The LRA model built on the predominant marker C5-OH, however, yielded the best classification accuracy. The ratio C5-OH/C3 seems to be redundant.

6. Procedure to optimize sensitivity

In newborn screening the declared aim is to optimise sensitivity to 100% and to increase specificity as far as possible. This assignment is of high importance as ethical arguments and cost effectiveness related to an erroneous diagnosis are concerned. However, the presented models showing high specificity do not always have optimal S_n of 100%. Exemplarily, Fig. 2 illustrates a procedure to optimize sensitivity of our most specific PAHD model (constructed on Phe and Phe/Tyr) by changing the default cut-off level from $p = 0.5$ stepwise to $p = 0.25$, 0.15, 0.1, and 0.01. By reducing the cut-off value to $p = 0.15$, all PAHD cases can be correctly classified ($S_n = 100\%$, arrow in Fig. 2), while the number of FP cases increases from 93 to 179. Therefore, S_p demounted from 99.905 to 99.818%. However, this procedure to optimize sensitivity of the regression model was performed only on the full training dataset, thus giving too optimistic results in general. For cross- or leave-one-out validation, which is recommended, if training sets are small, the entire model-building process including optimisation of the cut-off value, and classification has to be repeated in each cross-validation training subset. Currently, the validation procedure is in progress, these results have to be presented later.

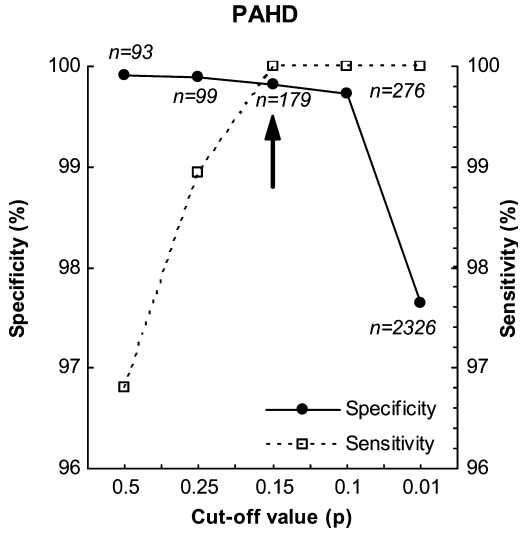

Fig. 2. Procedure to optimize sensitivity of the LRA model for PAHD constructed on the input variables Phe and Phe/Tyr: costs to optimize sensitivity by reduction of specificity are illustrated. The changes of sensitivity (%), specificity (%) and absolute number of FP cases (n) are given for cut-off values of $p = 0.5$, 0.25, 0.15, 0.1, and 0.01. The arrow indicates 100% sensitivity and a lower specificity of 99.818% ($n = 179$ FP cases) at $p = 0.15$ compared to $p = 0.5$ by default showing a sensitivity of 96.809% and a specificity of 99.905% ($n = 93$ FP cases). The cut-off optimization was performed on full training data (not validated).

7. Discussion

To satisfy the ever growing need for effective screening and diagnostic tests MS/MS provides a very high throughput and has the potential to be highly accurate. The complexity of analysed high-dimensional metabolic data using MS/MS requires machine learning and data mining techniques to support the classification of disease and the identification of potentially useful diagnostic markers. In turn, the identification of key metabolites could shed light on the nature of the disorder.

However, there are many data mining techniques for the processing and general learning of high-dimensional data in proteomics/metabolomics. Current research focuses on the detection of regions of interest in matrix-assisted desorption/ionization-time of flight mass spectra to mine differences in the protein pattern between healthy and diseased persons using established supervised and unsupervised methods [24,25]. Our experiments were not directly applied to the protein mass spectra, but were performed on data by the conversion of raw mass spectra into clinically meaningful results (amino acid and fatty acid concentrations) [1]. We investigated two different approaches to build classification models on provided newborn screening data with high diagnostic prediction. Machine learning offers the advantage that markers are found without any other a priori assumptions or conditions, and our results correspond well to the established biochemical knowledge [8–12]. In our approach we used the decision tree paradigm to identify potentially useful metabolites (nodes of tree) by computing the information gain for all candidate features. However, filter based selection techniques like information gain or correlation-based filters also yield promising results, but solely select single attributes as subsets [26,27]. Building LRA models on single metabolites and ratios, which reflect the interaction of single metabolites in newborn metabolism (cf. PAHD: Phe $\uparrow \cap$ Tyr $\downarrow \Rightarrow$ Phe/Tyr), the number of false positive cases could be diminished significantly (up to 6–16 times) compared to the decision tree classifier.

Therefore, knowledge on abnormal newborn metabolism modelled as ratios of interacting attributes thus provides an important contribution to enhance a model's specific power, while sensitivities in both approaches remain unchanged.

However, no classification model in both approaches achieved 100% sensitivity. In this context we briefly sketched a procedure showing the best trade-off between optimal sensitivity and specificity that can be accepted by adjusting the cut-off value in LRA. As mentioned before, further validation studies are warranted.

In addition to the decision tree classifier we used LRA, which shows highest discriminatory performance as we could demonstrate in prior experiments comparing various classification algorithms on metabolic data [27]. Some of them are well accepted in current proteomic research as other authors described [28]. However, both paradigms can be classified as directly interpretable techniques, which represent their data relations in an explicit way like a probabilistic model (LRA) or a tree structure, and so they find more acceptance in a clinical ambience.

8. Conclusion

Data mining in MS/MS data enables us to identify disease state metabolites without any a priori knowledge. The consideration of biochemical knowledge for the model-building process by combining interacting disease state metabolites revealed a further increase of the classifiers' discriminatory performance and lead to a significant increase of the specific power of our screening models. Our models achieved sensitivity values >95.2%. The number of FP cases in all three disorders did not exceed 0.001%.

The presented approach, which considers mining techniques and expert knowledge for the model-building process, permits the construction of classification rules with high diagnostic prediction.

Acknowledgments

We thank Dr. A.A. Roscher from Dr. von Hauner Children's Hospital, University of Munich, Germany for providing anonymous newborn screening data. This study was generously supported by the Austrian Industrial Research Promotion Fund FFF (Grant No. HITT-10 UMIT).

References

[1] Chace DH, DiPerna JC, Naylor EW. Laboratory integration and utilization of tandem mass spectrometry in neonatal screening: a model for clinical mass spectrometry in the next millennium. Acta Paediatr (Suppl) 1999;88:45–7.

[2] Liebl B, Nennstiel-Ratzel U, von Kries R, Fingerhut R, Olgmoller B, Zapf A, et al. Expanded newborn screening in Bavaria: tracking to achieve requested repeat testing. Prev Med 2002;34:132–7.

[3] Liebl B, Nennstiel-Ratzel U, Roscher AA, von Kries R. Data required for the evaluation of newborn screening programmes. Eur J Pediatr 2003;162(Suppl. 1):57–61.

[4] Millington DS, Terada N, Kodo K, Chace DH. 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; 1992. p. 59–71.

[5] Rashed MS, Ozand PT, Bucknall MP, Little D. Diagnosis of inborn errors of metabolism from blood spots by acylcarnitines and amino acids profiling using automated electrospray tandem mass spectrometry. Pediatr Res 1995;38:324–31.

[6] Mendes P. Emerging bioinformatics for the metabolome. Brief Bioinform 2002;3:134–45.

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

[8] Adler C, Ghisla S, Rebrin I, Heizmann CW, Blau N, Curtius HC. Suspected pterin-4a-carbinolamine dehydratase deficiency: hyperphenylalaninemia due to inhibition of phenylalanine hydroxylase by tetrahydro-7-biopterin. J Inherit Metab Dis 1992;15:405–8.

[9] Chace DH, Millington DS, Terada N, Kahler SG, Roe CR, Hofman LF. Rapid diagnosis of phenylketonuria by quantitative analysis for phenylalanine and tyrosine in neonatal blood spots by tandem mass spectrometry. Clin Chem 1993;39:66–71.

[10] Van Hove JL, Zhang W, Kahler SG, Roe CR, Chen YT, Tereda N, et al. Medium-chain acyl-CoA dehydrogenase (MCAD) deficiency: diagnosis by acylcarnitine analysis in blood. Am J Hum Genet 1993;52:958–66.

[11] Bannwart C, Wermuth B, Baumgartner R, Suormala T, Wiesmann UN. Isolated biotin-resistant deficiency of 3-methylcrotonyl-CoA carboxylase presenting as a clinically severe form in a newborn with fatal outcome. J Inherit Metab Dis 1992;15:863–8.

[12] Blau N, Thony B, Cotton RGH, Hyland K. Disorders of tetrahydrobiopterin and related biogenic amines. In: Scriver CR, Kaufman S, Eisensmith E, Woo SLC, Vogelstein B, Childs B, editors. The metabolic and molecular bases of inherited disease. 8th ed. New York: McGraw Hill; 2001. chapter 78.

[13] Children's Health System and University of Washington, GeneTests. A medical genetics information resource, Available from: http://www.geneclines.org.

[14] California Department of Health Services, Newborn Screening Program MS/MS Research Project, Available from: http://www.dhs.ca.gov/pcfh/gdb/html/PDE/MSMSMainPage.htm.

[15] American College of Medical Genetics/American Society of Human Genetics Test and Technology Transfer Committee Working Group. Tandem mass spectrometry in newborn screening. Genet Med 2000;2:267–9.

[16] Mitchell TM. Machine learning. Boston, MA: McGraw-Hill; 1997.

[17] Langley P. Selection of relevant features in machine learning. In: Proceedings of the AAAI fall symposium on relevance. New york: AAAI Press; 1994. p. 140–4.

[18] Witten IH, Frank E. Data mining—practical machine learning tools and techniques with java implementations. San Francisco: Morgan Kaufmann; 2000.

[19] Quinlan RJ. Induction of decision trees. Machine learning 1986;1:81–106.

[20] Quinlan RJ. C4.5: program for machine learning. San Mateo, CA: Morgan Kaufmann; 1993.
[21] Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York: Wiley; 2000.
[22] Health professionals guide to newborn screening. Wisconsin state laboratory of hygiene. Available from: www.slh.wisc.edu/newborn/guide.
[23] Chace DH, Sherwin JE, Hillman SL, Lorey F, Cunningham GC. 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 1998;44:2405–9.
[24] Lee KR, Lin X, Park DC, Eslava S. Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method. Proteomics 2003;3:1680–6.
[25] Neville P, Tan PY, Mann G, Wolfinger R. Generalizable mass spectrometry mining used to identify disease state biomarkers from blood serum. Proteomics 2003;3:1710–5.
[26] Baumgartner C, Baumgartner D, Böhm C, Classification on high dimensional metabolic data: Phenylketonuria as an example. In: Proceedings of the Second International Conference on Biomedical Engineering (BioMED 2004), Innsbruck, Austria, ACTA Press; 2004, p. 357-60.
[27] Baumgartner C, Böhm C, Baumgartner D, Marini G, Weinberger K, Olgemöller B, et al. Supervised machine learning techniques for the classification of metabolic disorders in newborns. Bioinformatics 2004 [in press].
[28] Purohit PV, Rocke DM. Discriminant models for high-throughput proteomics mass spectrometer data. Proteomics 2003;3:1699–1703.