Identification of optimal classification functions for biological sample and state discrimination from metabolic profiling data

Kyongbum Lee1,†, Daehee Hwang2,†, Tadaaki Yokoyama3, George Stephanopoulos2, Gregory N. Stephanopoulos2 and Martin L. Yarmush3,*

1Chemical and Biological Engineering, Tufts University, Medford, MA 02155, USA, 2Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA and 3Shriners Burns Hospital and Center for Engineering in Medicine, Massachusetts General Hospital, Harvard Medical School, GRB 1402, Boston, MA 02114, USA

Received on February 26, 2003; accepted on September 9, 2003
Advance Access publication January 29, 2004

ABSTRACT
Motivations: Classification of biological samples for diagnostic purposes is a difficult task because of the many decisions involved on the number, type and functional manipulations of the input variables. This study presents a generally applicable strategy for systematic formulation of optimal diagnostic indexes. To this end, we develop a novel set of computational tools by integrating regression optimization, stepwise variable selection and cross-validation algorithms.

Results: The proposed discrimination methodology was applied to plasma and tissue (liver) metabolic profiling data describing the time progression of liver dysfunction in a rat model of acute hepatic failure generated by d-galactosamine (GalN) injection. From the plasma data, our methodology identified seven (out of a total of 23) metabolites, and the corresponding transform functions, as the best inputs to the optimal diagnostic index. This index showed better time resolution and increased noise robustness compared with an existing metabolic index, Fischer’s BCAA/AAA molar ratio, as well as indexes generated using other commonly used discriminant analysis tools. Comparison of plasma and liver indexes found two consensus metabolites, lactate and glucose, which implicate glycolysis and/or gluconeogenesis in mediating the metabolic effects of GalN.

Contact: ireis@sbi.org

INTRODUCTION
With recent advances in high-throughput assay methods, various types of bioinformatics approaches have been proposed to extract useful empirical knowledge from large volumes of gene and protein expression data. For example, these kinds of analyses have been used for (1) classifying normal and cancer cells (Golub et al., 1999; Califano et al., 2000; Bichsel et al., 2001), (2) discovering novel subtypes of tumors (Hedenfalk et al., 2001), (3) identifying potential biomarkers for cancer prognosis (Alaiya et al., 2000; Kennedy, 2001), (4) delineating gene families by function (Xu et al., 2001; Rus et al., 2002) and (5) improving general understanding of cellular biochemical regulation (Holter et al., 2000). Nevertheless, the extent to which such information may be employed directly in practical clinical applications is currently limited, because: (1) the knowledge output as described by gene regulatory or protein interaction networks is in a form too convoluted for clinically meaningful interpretation, (2) dimension reduction analyses [e.g. principal component analysis (PCA) or Fischer’s discriminant analysis (FDA)] restrict description of the data structure to weighted linear combinations, which often result in loss of one-to-one correspondences between dimensional axes and physiological variables, (3) it is difficult to validate the statistical significance of results (e.g. regulation mechanisms elucidated by complex gene or protein network models) and (4) most gene and some protein expression data types are not well suited for medical uses such as routine diagnostics, because they are obtained via invasive procedures (e.g. tissue biopsies).

This study addresses these issues by developing a systematic method for extracting useful diagnostic information from multivariate biological data. In particular, the proposed method is applied to metabolic profiling data, which offers complementary information to gene and protein expression profiling data (Glassbrook et al., 2000; Glassbrook and Ryals, 2001). Metabolites are intermediates of biochemical
pathways that convert nutrient fuel into energy, maintain cellular homeostasis, eliminate harmful chemicals and provide building blocks for biosynthesis. Moreover, many metabolic intermediates are in constant exchange with the extracellular medium, and changes in extracellular metabolite concentrations may be correlated with changes in pathway activity. Hence, metabolites can be considered the most immediate descriptors of cell function. In recent applications, metabolic profiling has been used to study plant metabolic phenotypes (Fiehn et al., 2000a; Roessner et al., 2001) and discover novel gene functions (Trethewey, 2001). Metabolite profiling also has obvious applications to the investigation of metabolic disorders (Griffin et al., 2001). The potential for clinical diagnosis protocols based on multiple metabolite measurements has been recognized for some time (Goodman et al., 1977; Jellum, 1977), but the early efforts did not produce strategies or generally applicable models for systematic analysis. Recent works by Nicholson et al. have applied pattern recognition techniques, notably PCA and partial least squares (PLS), to ‘metabonomic’ data generated by nuclear magnetic resonance (NMR) spectroscopy to develop diagnostic screens for heart disease (Brindle et al., 2002) and drug toxicity (Nicholson et al., 2002; Coen et al., 2003). This paper expands the metabolic profiling tool space by developing a flexible discriminant analysis strategy that, unlike FDA or PLS-DA, constructs the sample classifier (diagnostic) using both linear and nonlinear variable transforms.

In the present analysis, we used a moderate sized data-set (~1000 entries) obtained in an established animal model of injury as inputs to a new computational methodology to derive the optimal functional forms for a diagnostic index. The injury model (Shito et al., 2001) describes the progression of liver dysfunction, or fulminant hepatic failure (FHF), induced by injection of liver toxin δ-galactosamine (GalN). The computational methodology combines multivariate alternating conditional expectation MACE (Breiman and Friedman, 1985) with stepwise search algorithm and leave-one-out cross-validation LOOCCV (Hwang et al., 2002), which integrates the selection (number and type) of the best input variables and identification of optimal functional forms whereby the selected variables enter the diagnostic index. This computational method should also facilitate the physiological interpretation of the diagnostic index. To explore this notion, the study also compares the effects of GalN on the metabolic profiles of the whole body (plasma samples) and the liver (isolated perfused liver samples), which is the primary site of GalN action.

SYSTEMS AND METHODS

Animals

Male Sprague–Dawley rats (Charles River Laboratories, Boston, MA) weighing 150–200 g were housed in a temperature (25°C) and light-controlled room (12 h light–dark cycle). Animals were individually housed and allowed to adjust to their new surroundings for at least 2 days before receiving treatment. The animals were cared for in accordance with the National Research Council guidelines. Experimental protocols were approved by the Subcommittee on Research Animal Care, Committee on Research, Massachusetts General Hospital. The initiation of FHF by GalN injection (Arai et al., 2001; Shito et al., 2001) has been described in detail elsewhere. Briefly, FHF was caused by intraperitoneally administering GalN dissolved in normal saline at a dose of 1.4 g/kg to fasted rats 12 and 24 h after beginning the fast. Previous work showed that a single dose administration of GalN elicits reversible liver dysfunction, whereas a dual dose administration produces irreversible liver failure. The mortality rate of this treatment was 25% at 48 h and 83% at 168 h after the initial injection. Control rats were fasted, but received normal saline injections. Rats remained fasted throughout the study period, which lasted up to 36 h. Plasma samples were obtained via cardiac puncture at 1, 4, 8 and 12 h after the second GalN or saline injection. Livers were then isolated and perfused for 60 min, during which time the perfusion medium was periodically sampled.

Metabolite measurements

Biochemical assays measured 23 metabolite concentrations on plasma samples from 61 animals divided into groups according to treatment and time elapsed following the second GalN/saline injection: 1(n = 8), 4(n = 9), 8(n = 11) and 12 h control (n = 6), and 1(n = 5), 4(n = 9), 8(n = 7) and 12 h FHF (n = 6). The metabolites were glucose, urea, lactate, ammonia, ketone bodies and amino acids. Assays on perfused liver samples measured the same set of metabolites. Each perfusion produced eight time point samples, which were collapsed into a single uptake or release rate by linear regression of the concentration–time data. Glucose and lactate concentrations were measured using commercial
enzyme assay kits (Sigma Chemical Co., St Louis, MO) that are based on the methods of Trinder (1969) and Loomis (1961), respectively. Urea was measured by the urease method reported by Seary et al. (1961). Ketone bodies (acetoacetate and β-hydroxybutyrate) were measured using an enzymatic fluorometric method developed by Olsen (1971). Ammonia and amino acids were quantified by high-performance liquid chromatography (HPLC) following pre-column derivatization with 6-aminoquinolyl-N-hydroxysuccinimidyl carbamate (Cohen and De Antonis, 1994), which is a commercially available fluorescent label for amino acids and primary and secondary amines (Waters Corp., Milford, MA).

**ALGORITHM**

Multi-variate alternating conditional expectation

MACE is a non-parametric procedure to estimate the optimal transformation \( \theta(Y) \) and \( \phi_1(X_1), \ldots, \phi_p(X_p) \) of the response variable \( Y \) (treatments of GalN in this study; 1 for normal and 2 for GalN injected samples) and the predictor variables (metabolites in this study) \( X_1, \ldots, X_p \) for multiple regression (Breiman and Friedman, 1985). These transformations aid in the interpretation of the relationship between response and predictor variables. Here, we used MACE to find the optimal transformations of metabolite variables that discriminate plasma or perfused liver samples from control and FHF rats. Algorithmically, MACE finds optimal transformations that maximize the correlation between response and predictor variables. In its original formulation, for the bivariate case (\( p = 1 \)), the optimal transformations of \( Y \) and \( X \) [denoted \( \theta^*(Y) \) and \( \phi^*(X) \)] satisfies

\[
\rho^* = \rho(\theta^*, \phi^*) = \max_{\theta, \phi} \rho(\theta(Y), \phi(X)), \quad (1)
\]

such that the transformations minimize the fraction of the variance not explained by a regression of \( \theta(Y) \) on \( \sum_{i=1}^p \phi_i(X_i) \).

The variance \( \sum_{i=1}^p \phi_i(X_i) \) is computed by:

\[
e^2(\theta, \phi) = \frac{E \left[ (\theta(Y) - \sum_{i=1}^p \phi_i(X_i))^2 \right]}{E\theta^2(Y)}. \quad (2)
\]

In this study, we looked for the optimal transformations of response and predictor variables for the purpose of discriminant analysis, rather than regression analysis. Therefore, the objective function maximizes the ratio of between-group variance to total variance instead of the correlation. Correspondingly, \( e^2 \) in Equation (2) is replaced with the ratio of within-group variance to total variance. This replacement permits a more stable stepwise search to find the best additive discriminant model. Let \( R = \sum_{i=1}^p \phi_i(X_i) \) and \( T = \theta(Y) \). Then, the ratio of within-group variance to total variance is defined by

\[
e^2 = \frac{\sum_{j=1}^c \left( \frac{\sum_{i \in \{j\}} (R_i - \bar{R}_j)^2}{N} \right)}{\sum_{k=1}^N (R_k - \bar{R})^2}, \quad (3)
\]

where \( \bar{R}_j \) is the mean of \( R \) in group \( j \) and \( c \) is the number of groups. Then, the ratio of between-group variance to total variance is \( \rho = 1 - e^2 \). For the sake of consistency, the between- and within-group ratios are denoted by the symbols used for their corresponding regression measures, correlation and the fraction of residuals.

**Stepwise variable selection**

MACE is combined with a stepwise selection method (Dillon and Goldstein, 1984) to find the best-fitting additive discriminant model (called the diagnostic index in this study). The original implementation of MACE (Breiman and Friedman, 1985) used the forward selection method (Dillon and Goldstein, 1984), but this study uses the stepwise selection method, because it reveals more clearly the independent nature of the variables selected to the discriminant model.

**Leave-one-out cross-validation**

In order to develop a discriminant model with a set of predictors that are robust with respect to sample variability and population noise, we used LOOCV (Hwang et al., 2002). The dataset was split into a test set consisting of eight randomly selected samples (one control and FHF sample for each of the four time points) and a training set composed of the remaining 53 samples:

1. A sample is taken out from the training set. Stepwise variable selection combined with MACE is performed on the remainder of the training set (53 samples) to find the best-fitting additive discriminant model.
2. A set of predictor variables \( (s_l) \) retained in the model and their partial \( F \)-ratios \( (r_l) \) are stored for the \( l \)-th iteration.
3. Steps 1 and 2 are continued until all training samples are taken out once.

At the completion of LOOCV, \( N \) sets of predictors \( (s_l) \) and their partial \( F \)-ratios \( (r_l) \) are obtained. In order to identify the best set of predictors for the training set, we used the averaged partial ratios and the conditional expectation of partial ratio for predictor variables as measures from the \( g \) sets of \( s_l \) and \( r_l \). They are referred to as power and local power in Table 1. Let \( S \) be the overall set of predictors selected during a total of \( N \) LOOCV procedures. First, the averaged partial ratio of \( X_k \) during \( N \) iterations of LOOCV is computed by:

\[
g_k = E (r_k | X_k \in S) = \frac{\sum_{X_k \in g} r_{kl}}{N}, \quad (4)
\]

where \( r_{kl} \) is the partial ratio of \( r_k \) in the \( l \)-th iteration, when \( X_k \) is selected in the \( l \)-th iteration. Next, the conditional expectation \( (c_{kh}) \) of \( X_k \) in a particular position \( h \) of the stepwise selection is computed by:

\[
c_{kh} = E (r_k | X_k \in o(h)) = \frac{\sum_{X_k \in o(h)} r_{kl}}{q_{kh}}, \quad (5)
\]
where \( o(h) \) is a predictor that is most frequently selected in the \( h \)-th position by the stepwise selection method during the \( N \) LOOCV procedures, and \( q_{kh} \) is the number of times that \( X_k \) is selected in the \( h \)-th position. For example, if \( o(1) \) is \( X_7 \) and \( X_7 \) is selected 37 times as the first variable to enter \( (q_{kh} = 37) \), then the conditional expectation of \( X_7 \) in the first position \( (c_{71}) \) is the average of the corresponding \( 37 \) partial ratios. If \( o(h_j) \) is the variable already included in \( o(h_i: i < j) \), the \( o(h_j) \) is defined by the variable that is most often selected, but not included in \( o(h_i: i < j) \). For example, if \( X_7 \) of \( o(1) \) and \( X_{15} \) are most and second most often selected in the fifth position, respectively, \( o(5) \) is \( X_{15} \), not \( X_7 \). The predictor variables that will be included in the discriminant model should have \( g_k \) and \( c_{kh} \) values larger than the cutoff value of 0.01.

**IMPLEMENTATIONS AND RESULTS**

**Optimal formulation of diagnostic index**

We obtained metabolic profiles on plasma samples from control and injured rats undergoing progressive FHF. Control and injured rats were injected with normal saline and saline dissolved GalN, respectively. The metabolic profiles consisted of plasma concentration data for 23 major primary metabolites (called predictor variables \( x_i \); see Methods section) obtained at various time periods after the GalN insult (1, 4, 8 and 12 h in both controls and FHF samples). From these data, the proposed criteria for outputs of LOOCVs (see Methods section) identified an optimal subset of seven metabolites and their transforms such that a diagnostic index consisting of the sum of these transforms yields maximal discrimination between control and FHF samples at the earliest time point after the GalN injection. These seven metabolites are called a basis set of metabolites. The most discriminating metabolites are, in order, histidine (HIS), arginine (ARG), lactate, alanine (ALA), isoleucine (ILE), methionine (MET) and glucose (Table 1).

Figure 2A shows the concentrations of these metabolites at various times after the second saline or GalN injection. Plasma concentrations of HIS and MET in the FHF animal increased significantly over control by 4 h after the second GalN injection, and remained elevated thereafter. FHF increased lactate concentration and decreased ARG and glucose concentrations. Finally, ALA and ILE concentrations changed little throughout the study period, except a sudden increase in ALA concentration between 8 and 12 h. The decrease in glucose and increase in lactate concentrations are noteworthy, because they consistently point to impaired glucose metabolism.

The transforms of these predictor variables and the progressive improvement in discriminatory power (i.e. the ratio of between-group variance to total variance; see Methods section) are shown in Figure 2C. The scatter plots of metabolites \( x_i \) and treatments of GalN \( y; 1 \) for controls and \( 2 \) for FHF) in Figure 2B show that the untransformed metabolites do not discriminate individually the source samples to any significant degree. In contrast, the scatter plots of the sum of transformed \( x_i(I_p) \) and the transformed \( y \) (response score; \( T(y) \) (Fig. 2D) clearly show the progressively improving discrimination of the source samples with the inclusion of each additional metabolite in the regression model:

\[
I_p = \sum_{i=1}^{P} T(x_i). \tag{6}
\]

The transform of the response variable \( y \) was trivial, yielding 1 and −1, respectively, for FHF and control. With the seven metabolites, the final discriminatory power was 0.936. The shapes of the transform functions for the predictor variables are described in Figure 2C. The equations, shown in Table 1, were derived by fitting \( x_i \) and \( T(x_i) \) to various function families, including linear, logarithmic, exponential and polynomial functions (Friedman and Stuetzle, 1982). For each series of \( x_i \) and \( T(x_i) \), the best function estimate was chosen to minimize the residual (Table 1). Substituting for each \( T(x_i) \) in Equation (6) with Table 1 entries, the final diagnostic index

<table>
<thead>
<tr>
<th>Rank</th>
<th>Metabolite</th>
<th>Power</th>
<th>Local power</th>
<th>Fitted equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HIS</td>
<td>0.4767</td>
<td>0.6148</td>
<td>( T(x) = 0.43 + 0.98x )</td>
</tr>
<tr>
<td>2</td>
<td>ARG</td>
<td>0.1533</td>
<td>0.1907</td>
<td>( T(x) = -0.07 - 0.57x )</td>
</tr>
<tr>
<td>3</td>
<td>LAC</td>
<td>0.1555</td>
<td>0.0587</td>
<td>( T(x) = 0.17 + 0.51x )</td>
</tr>
<tr>
<td>4</td>
<td>ALA</td>
<td>0.0249</td>
<td>0.0452</td>
<td>( T(x) = -0.33 - 0.53x + 0.21x^2 - 0.27x^3 )</td>
</tr>
<tr>
<td>5</td>
<td>ILE</td>
<td>0.0219</td>
<td>0.0201</td>
<td>( T(x) = -0.05 - 0.21x^3 + 0.40x^4 + 0.45x^5 )</td>
</tr>
<tr>
<td>6</td>
<td>MET</td>
<td>0.0273</td>
<td>0.015</td>
<td>( T(x) = 0.31 \log(2.91x + 2.96) - 0.04 )</td>
</tr>
<tr>
<td>7</td>
<td>GLC</td>
<td>0.0543</td>
<td>0.0161</td>
<td>( T(x) = 0.02 \exp(-3.07x + 0.96) - 0.29 )</td>
</tr>
</tbody>
</table>

Metabolites were selected and ordered by MACE, stepwise variable selection and LOOCV. Power and local power refer to position (selected order) independent and dependent contribution, respectively, of the entering variable to overall discrimination as measured by the within-group to between-group variances (see Methods section). Variables were not allowed to enter unless they increased the final discrimination by at least 1%. This cut-off value selected a total seven metabolite predictor variables for final plasma discriminant model (diagnostic index).

**Abbreviations:** HIS, histidine; ARG, arginine; LAC, lactate; ALA, alanine; ILE, isoleucine; MET, methionine; GLC, glucose.

Table 1. Plasma discriminant model variables
Optimization for biological sample discrimination

Fig. 2. Plasma metabolic profiles and optimized diagnostic index. MACE, stepwise variable selection and LOOCV identified metabolite predictor variables and transforms that progressively improved the discriminatory power of the final diagnostic index. (A) Mean plasma concentrations of the selected metabolites, in order of decreasing discriminatory power, top-to-bottom (Table 1). Controls are represented by open square boxes and FHF by closed square boxes. Error bars are SD. Treatment and time point sample sizes varied from five to eight. (B) Scatter plots of untransformed response ($y$) and scaled predictor ($x_i$) variables. Last row shows the linear transform of $y$ to $Ty$. (C) Scatter plots of transformed predictors ($Tx_i$) and $x_i$. Dashed lines represent estimates of $Tx_i$ obtained by fitting various function families to the data. (D) Scatter plots of $Ty$ versus $Tx$ showing progressively improving discrimination. Numbers are ratios of between-group variance to within-group variance.

In general, the transforms (six out of the seven) increased or decreased monotonically over the present range of predictor variables, suggesting a simple relationship between the raw and transformed predictor variables. One exception was the transform of ILE ($x_4$), which was described by a fifth order polynomial.

Application to diagnostics

The diagnostic index ($I$) described in Table 1 was evaluated for the training data, and plotted against the corresponding sample identification numbers (Fig. 3A). Consistent with our objective, the control and FHF samples clearly separated.
A diagnostic index (I) described in Table 1 is evaluated for the training data and plotted against corresponding sample identification numbers. The sample group memberships are indicated by the symbol legend. L1, L4, L8 and L12 are FHF samples taken at 1, 4, 8 and 12 h after GalN injection. H1, H4, H8 and H12 are corresponding control samples. FHF are plotted as closed and control as open symbols. The separation between control and FHF samples was achieved even for the earliest time point, thus enabling us to detect FHF from metabolite measurements only 1 h after the second GalN injection.

In Figure 3B, new test samples, which were not involved in formulating the diagnostic index, were predicted using the confidence limits derived from the training set. As with the training set, all the test samples were correctly classified.

The performance of our method was compared against that of a clinical index, Fischer’s ratio (Fischer et al., 1975), which correlates increasing severity of hepatic dysfunction with decreasing molar ratio of branched chain amino acids to aromatic amino acids (BCAA/AAA). This ratio was evaluated for the training data used to develop our optimal diagnostic index and plotted against corresponding sample identification numbers in Figure 4A. In contrast to our optimal index, which clearly separated control from FHF samples even at the 1 h time point, the BCAA/AAA ratio did not discriminate significantly between control and FHF samples at any time point. The largest difference between control and FHF mean BCAA/AAA scores was observed for the 12 h samples, but even at this time point, statistical significance could not be inferred (at the level of p < 0.05).

We also compared our diagnostic index against other popular models used for sample classification: self-organizing map (SOM; Fig. 4B), PCA (Fig. 4C) and method of PLS (Fig. 4D). SOM (Fig. 4B) did not separate control and FHF samples correctly except for the 12 h time point. PCA and PLS (Nguyen and Rocke, 2002) yielded comparable results, and performed better than the previous two methods, as they discriminated between control and FHF samples for both the 8 and 12 h time points. However, neither method was able to discriminate clearly between control and FHF samples for the earlier time points (1 and 4 h). In all cases, the separation afforded by MACE was superior. The two best discriminated groups were the control and FHF samples obtained at the latest time point (12 h), consistent with the time course of GalN insult established using plasma levels of liver enzymes (Shito et al., 2001).

**Metabolic profiling at the tissue level**

To investigate the relationship between changes in cellular metabolic activity at the site of initial disturbance (by GalN action) and the metabolic state of the whole body, we compared the optimized diagnostic indexes for liver and plasma. Liver metabolic profiles were obtained using the isolated perfused liver. This experimental system removed systemic influences and afforded measurements on metabolic rates, which more directly reflect pathway activities than concentrations.

MACE and stepwise-selection identified the following basis metabolites for the liver, listed in order of their discriminatory power: lactate output, glutamine (GLN) uptake, β-hydroxybutyrate (3-OHB) output, ammonia (AMM) output, glucose output and ornithine (ORN) output. Figure 5A shows the time course of changes to the uptake or output of these metabolites. Lactate output was elevated throughout, whereas GLN uptake and glucose output steadily decreased. The trends were not as simple for 3-OHB, AMM and ORN output, as the time courses for control and FHF intersected at one or more time points.

The elevated lactate and depressed glucose outputs pointed to impaired glucose metabolism in the liver, consistent with observations on the plasma data. Moreover, both the plasma and perfused liver diagnostic indexes included lactate and glucose, with lactate as the more important variable. On the
Optimization for biological sample discrimination

Fig. 4. Diagnostic index comparison. A clinical index (A) and commonly used discriminant models (B, C and D) are compared against our optimized diagnostic index. The sample group memberships are indicated by the symbol legend as in Figure 3A. (A) Fischer’s ratio correlates decreasing molar BCAA/AAA ratios with increasing liver dysfunction severity. (B) Unsupervised SOM. (C) PCA (FDA). (D) Method of PLS. Note that PCA and PLS plots show sample projections onto the first Principal Component (PC) and the first Latent Variable (LV) for sake of fair comparisons.

On the other hand, HIS, ARG, ALA, ILE and MET in the plasma set were replaced by GLN, 3-OHB, AMM, ORN and tyrosine (TYR) in the liver set. Although lactate appeared in both indexes, it played a more critical role in the liver diagnostic index than the plasma index, as it accounted for 62 and 15% of discrimination of the liver and plasma samples, respectively.

The liver transform functions exhibited more complex shapes (Fig. 5C). Six out of the seven plasma transform functions (Fig. 2C) were monotonic, and could be described with a linear, exponential or polynomial function of order 3 or less. In contrast, most of the basis metabolites in the liver diagnostic index had to be fitted using stepwise polynomial regression. Five out of the six liver transforms showed non-monotonic behavior, and were described by higher (>3) power polynomials. Transforms for LAC, GLN and AMM were biphasic, and those for 3-OHB and ORN oscillatory, suggestive of a complex relationship between the raw and transformed variables.

DISCUSSION AND CONCLUSION

This study presents a novel computational methodology for optimally discriminating biological systems based on metabolite measurements, which has potential clinical applications, such as diagnostics. We demonstrate the methodology on a moderate sized (2 × 61 × 23) dataset describing the progression of liver failure in a rat model of GalN toxemia developed previously (Shito et al., 2001). Some detailed information already exists regarding GalN toxicity and liver metabolism (Ozeki et al., 1982; Usui et al., 1985; De Oliveira et al., 1992; Arai et al., 2001), providing an opportunity to check the discrimination results against physiological knowledge. The data were collected at two levels, whole body...
Fig. 5. Perfused liver metabolic profiles and optimized diagnostic index. After obtaining plasma samples by cardiac puncture, livers were isolated and perfused to obtain corresponding liver specific metabolic rate profiles. (A) Mean metabolic rates, in order of decreasing discriminatory power, top-to-bottom. Controls are represented by open square boxes and FHF by closed square boxes. Error bars are SD. Treatment and time point sample sizes varied from five to eight. See Figure 2 for (B)-(D) captions. (E) The diagnostic index (I) is evaluated for the training data (see Methods section) and plotted against corresponding sample identification numbers. The sample treatment group memberships are indicated by the symbol legend as in Figure 3A.
Optimization for biological sample discrimination

Optimization for biological sample discrimination (plasma) and tissue (isolated liver). The plasma data was used to construct an additive discriminant model (the diagnostic index) composed of linear and non-linear transforms of the basis metabolites. The diagnostic index correctly classified, even at the earliest observed time point (1 h), both the training and test samples with 99% confidence. This represented a significant improvement over existing metabolic indexes, Fischer’s BCAA/AAA molar ratio as well as indexes generated using other popular discriminant analysis tools (SOM and FDA).

The systematic development of a diagnostic index from even moderate sized (~10s–100s of variables) data is a complicated task, when little or no mechanistic information is at hand regarding the relationship structure (i.e. functional dependence) between the measured metabolites and the physiological state under investigation, which is often the case for many diseases of multi-factorial origin. For example, with only five functional (e.g. periodic, exponential, polynomial, power and logarithmic) families considered, the number of all possible functional form combinations for a given set of \( k \) metabolites is \( 5^k \), even in the simplest case of an additive model. A key advantage of the present method is that it simultaneously selects both the best subset of basis metabolites and identifies the necessary transforms of these metabolites for optimal discrimination, without \textit{a priori} ordering of the variables and any assumptions about the type of transform functions. Thus, both parsimony and flexibility are achieved.

The stepwise search algorithm guarantees that the metabolite inputs to the diagnostic index define the most compact, and therefore maximally independent, set of metabolic characteristics that differentiate one physiological state from another. An intuitive, geometric interpretation of these basis metabolites is given by an analogy with vector spaces. In this view, the basis metabolites are considered to be nonlinear bases spanning a particular physiological discrimination space. In this study, we refer to the hyperspace spanned by the seven basis metabolites as a ‘liver dysfunction space’. Similar interpretations have been suggested for projection-based data analysis methods such as PCA (in essence a singular value decomposition) and FDA, where linear combinations of variables constitute bases spanning a multi-dimensional hyperspace. However, unlike PCA or FDA, our method uniquely associates each basis with a single transformed metabolite rather than a linear combination of several metabolites. This type of one-to-one basis mapping directly facilitates physiological interpretation of discrimination characteristics.

Figure 6 shows a three-dimensional liver dysfunction space, spanned by \( T_x(HIS) \), \( T_x(ARG) \) and \( T_x(LAC) \), where the FHF and control group separation occurs mainly along the positive directions of the \( T_x(HIS) \), \( T_x(ARG) \) and \( T_x(LAC) \) axes. Thus, elevated HIS and LAC, and decreased ARG (ARG transform is linear and negative) concentrations are characteristic of this pathology. It is important to note that the aforementioned liver dysfunction space, as defined by the plasma diagnostic index, is a subspace of the total attainable physiological space. By definition, this subspace captures significant metabolic changes associated with FHF as observed on plasma samples.
Physiological subspaces can also be defined for various subsystems within an organism that are subject to the same pathology (in this case liver dysfunction), but are described by different discriminant models. In this study, a second liver dysfunction subspace was defined by the diagnostic index derived from the perfused liver data. The plasma and liver subspaces shared two common basis components, lactate and glucose, which can be used to construct a ‘consensus’ subspace (Fig. 6B). In this space, plasma and liver sample projections obtained by their respective glucose and lactate transforms trace two distinct trajectories. In the case of plasma, \( T_x(LAC) \) monotonically decreased with increasing \( T_x(GLC) \) regardless of sample origin (control or FHF). On the other hand, FHF samples traced a steeper curve, suggestive of higher sensitivity of lactate to changes in glucose concentration. In the case of the perfused liver, \( T_x(LAC) \) again decreased with increasing \( T_x(GLC) \) for the control samples, but showed biphasic behavior for the FHF samples. Moreover, the perfused liver trajectory spanned a larger range of transformed variable values than the plasma trajectory. This is unlikely due to differences in measurement error, as the same animals and biochemical assays were used for both plasma and perfused liver samples. A more likely explanation is that the effects of GalN on glucose–lactate metabolism in the liver are dampened by extra-hepatic contributions in the whole body. Conceptually, we consider such ‘smoothing’ to result from a ‘time integral’ of tissue-level metabolic activity, which manifests as regulatory interactions between various tissues aimed at maintaining whole body-level metabolic homeostasis (Fig. 6C).

The monotone nature of control sample trajectories indicates that both whole body and liver glucose and lactate metabolic activities are synchronized, presumably because the liver is the primary regulator of glucose homeostasis in the body. Our subspace analysis indicates that GalN disrupts this synchronization, and together with the concentration and rate profiles of lactate and glucose (Figs 2A and 5A) point to metabolic pathways connecting glucose and lactate as likely sites of GalN action in the liver. This interpretation is supported by several published reports. For example, administration of fructose 1,6-bisphosphate reduced liver cell death during GalN-induced FHF (De Oliveira et al., 1992). Other studies found a diminishing effect of GalN on the reserve power of glycolysis and fructose 1,6-diphosphatase activity (Ozeki et al., 1982).

In conclusion, our methodology identified a robust and efficient discriminant model from metabolic profiling data, which has shown superior diagnostic power compared with a clinical index as well as other statistical models. A novel feature of our method is that it simultaneously selects the best subset of input variables and their linear or nonlinear transforms for optimal sample discrimination. The method is systematic and general, and could be applied to other diseases, especially those of multi-factorial origin, such as diabetes and obesity, as well as other types of data, such as protein expression profiles. Given enough repeated observations, we expect to develop separate discriminant models (diagnostic indexes) for particular time points, or stages, of disease progression, which will enable us to construct a dynamic model for monitoring disease progression. Moreover, a larger dataset would more representatively sample the population distribution, and lessen the probability of over-fitting, which give rise to transforms with little or no physical meaning. This was likely the case when a fifth-degree polynomial was selected as the optimal transform for the plasma basis metabolite ILE (Table 1). In light of these limitations, future studies could benefit from (1) improved tools for high-throughput measurements of metabolites in complex biological fluids (Fiehn et al., 2000b; Griffin et al., 2002) and (2) improved knowledge of metabolic pathways, so that more comprehensive datasets, including derivative quantities like metabolic fluxes, are routinely available for systematic analysis.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Albert K. Lee for useful discussions. This work was supported by NIH grant no. GM58125, the Shriners Hospitals for Children and the Tufts University Faculty Fund.

REFERENCES


