Classification of Lactate Dehydrogenase of Different Origin by Liquid Chromatography-Mass Spectrometry and Multivariate Analysis

Dan Bylund, Jenny Samskog, and Karin E. Markides
Department of Analytical Chemistry, Uppsala University, Uppsala, Sweden

Sven P. Jacobsson
AstraZeneca R and D Södertälje, Pharmaceutical and Analytical R and D, Södertälje, Sweden

A method intended to serve as a multivariate quality control tool in the production of pharmaceutical proteins is presented. The method is based on multivariate analysis of peptide maps generated with liquid chromatography-mass spectrometry (LC-MS). Lactate dehydrogenase (LDH) from different species and tissues were used as model compounds in the study. The proteins were digested with Endoproteinase Lys-C before the LC-MS analysis. After data pretreatment of the peptide maps, successful classification of the LDHs were obtained by discriminant analysis with partial least squares regression and artificial neural networks. Further, principal component analysis was applied to visualize the relationships between the samples. (J Am Soc Mass Spectrom 2003, 14, 236–240) © 2003 American Society for Mass Spectrometry

In view of the human genome project and further understanding of biochemical processes, the pharmaceutical industry's interest in proteins and peptides as drug candidates has increased. This raises a need for analytical methods to monitor the purity and stability of such compounds. These methods should aim to confirm the amino acid sequence and detect the possible presence of posttranslational modifications, e.g., oxidation or glycosylation.

Peptide mapping (or fingerprinting) [1] is a traditional method in this area. The proteins are cleaved into smaller fragments, typically by tryptic digestion, and analysed with, e.g., liquid chromatography (LC), and ultraviolet detection [2]. The peak pattern of the chromatogram then serves to characterise the protein. Problems with this method are that no secure structural information is obtained and that drift in the chromatographic conditions may introduce changes in the pattern over time, thereby complicating the interpretation [3].

Recent developments in ionization technology and instrumentation have made mass spectrometry (MS) an important technique in the analyses of biomolecules. With matrix-assisted laser desorption ionization (MALDI) the mass of intact proteins and larger peptide fragments can be determined. Smaller fragments, however, are not suitable for traditional MALDI due to the high spectral background caused by matrix components in the lower *m/z* region. With electrospray ionization (ESI) the minor fragment can also be studied. Another advantage with ESI is that the technique is well suited for on-line coupling to LC, i.e., LC-MS. The drawback with ESI, compared to MALDI, is the low tolerance for salts in the sprayed solution. Peptides present at high concentration may also, due to suppression effects [4], decrease the ability to detect less abundant peptides. However, when LC-MS is applied, this problem is less significant.

Proteins analyzed with MS can be identified by the use of databases accessed on the Internet [5]. Further fragmentation by, e.g., collision induced dissociation can provide additional information which may be necessary especially for de novo sequencing [6]. However, in quality control of a production line, it is often minor variations that need to be considered [7]. Such small differences can be discovered by the use of chemometric tools. The data matrices generated with LC-MS can be modeled with multiway analysis methods, e.g., parallel factor analysis, into the mass spectra, elution profiles and concentrations of the sample components [8]. Alternatively, the data can be reduced into vectors, and the modelling based on multivariate analysis of the

Published online January 31, 2003

Address reprint requests to Dr. K. E. Markides, Department of Analytical Chemistry, Uppsala University, P.O. Box 599, 751 24 Uppsala, Sweden. E-mail: karin.markides@kemi.uu.se

1044-0305/03/$30.00
doi:10.1016/S1044-0305(02)00906-6

Received March 1, 2002
Revised December 10, 2002
Accepted December 11, 2002

derived mass spectra [9], chromatographic profiles [10], or lists of integrated peaks [11].

In this study, LC-MS was used to generate structurally informative peptide maps of lactate dehydrogenase (LDH) from different species and tissues. The dimensions of the data was then reduced, so that each sample was represented by a single mass spectrum. Thereby the need for reproducible retention times was circumvented at the expense of chromatographic information. Finally, multivariate analysis tools were applied to visualize the differences between the samples and classify the LDHs.

Experimental

Sample Preparation

The LDH variants (beef heart, hog muscle, pig heart, and rabbit muscle) were purchased from Boehringer Mannheim GmbH (Mannheim, Germany). Before digestion of these proteins, the protein buffer was exchanged to the digestion buffer, 6 M guanidine hydrochloride (Fluka Chemie AG, Buchs, Switzerland), by use of a Micron centrifugal filter, 10,000 MWCO, (Millipore Corp., Bedford, MA). 50 μl 2 mg/ml LDH was preincubated at 37 °C. After that 50 μl water, 200 μl 15 mM Tris(hydroxymethyl)aminomethane (Tris) (Merck, Darmstadt, Germany) pH 8.5, and 2 μl endoproteinase Lys-C (sequencing grade, Boehringer) was added to the solution and incubated at 37 °C during 4 h. Thereafter, 2 μl formic acid was added to the digested protein solution. Finally, 50 μl was desalted on a ZipTip (Millipore Corp.), eluted in 80% ACN, and dried by a SpeedVac concentrator (Savant Instruments Inc., Holbrook, NY). Before injection into the LC-MS, the peptides were redissolved in 20 μl solvent A (95% water and 5% methanol vol/vol).

LC-MS Analysis

The LC-MS system comprised a Rheos 2000 pump (Flux Instruments, Basel, Switzerland), a 1.0 μl external loop injector (Valco Instruments, Houston, TX), a PepMap C18 column (150 × 0.3 mm, 3 μm) from LC Packings (Amsterdam, The Netherlands), and an API 100 mass spectrometer (PE Sciex, Concord, ON, Canada) operated in the positive ion scan mode for *m/z* 250–1250 with a step size of 0.5 u and a total scan time of 2.5 s. The LC pump was operated at 60 μl/min, with a flow splitting device before the injector giving approximately 3 μl/min through the column. The solvents were A (95% water and 5% methanol) and B (20% water and 80% methanol), both acidified to 0.1% formic acid. Linear gradient elution was applied from 0 to 90% B within 40 min and followed by isocratic elution with 90% B for 5 min.

A total of 20 samples of LDH digests (4 beef heart, 5 hog muscle, 5 pig heart, and 6 rabbit muscle) were analyzed. The samples were run on the system in random order with replicates of both the digestion and desalting steps during a period extended over one year.

Data Analysis

Principal component analysis (PCA) was used to visualize the relationships between the samples, while the final classification was based on discriminative analysis with partial least squares regression (PLS) [12] or artificial neural networks (ANN) of the multi-layer feed-forward architecture [13]. A single model was used for classification of all the proteins, i.e., the PLS2 algorithm was applied. The ANNs were constructed with a single hidden layer of operating units (nodes) and four output nodes, corresponding to the four LDH proteins studied. In all nodes, sigmoidal transfer functions were applied according to

$$o_j = \frac{1}{1 + e^{-\sum_i o_i w_{ji} + \theta}} \quad (1)$$

where o_i and o_j are the outputs from layers I and J, respectively, w are the weights for the connections and θ is a bias term. The weights and the biases were initialized as small random numbers and then iteratively adjusted by the back-propagation learning rule [13] according to

$$\Delta w_{ji}(epoch) = \eta \delta_j o_i + \alpha \Delta w_{ji}(epoch - 1) \quad (2)$$

where η is the learning rate, α the momentum term, δ_j a function of the error in o_j, and *epoch* is an index for the number of iterations performed. In this work the following settings were used: $\eta = 0.1$, $\alpha = 0.5$, and the maximum number of iterations max(*epoch*) = 2000.

All calculations were performed with MATLAB (version 6.0, The MathWorks, Natick, MA) run on a standard PC with a Pentium II 400 MHz processor and 128 MB RAM. The Chemometrics Toolbox (The MathWorks) was used for PCA and PLS, while the codes for ANN and data preprocessing were written by DB.

Results and Discussion

A typical total ion chromatogram for LDH hog muscle is shown in Figure 1. Manual identification of the peaks and comparison with databases indicated sequence coverage of about 30% on average, where mainly the smallest and largest fragments were lost. A few unknown peaks were also observed, probably due to unspecific or incomplete cleavages.

The LC-MS data matrices for the retention interval 10–45 min were background adjusted by a combination of matched filtering and second derivatives [14], assuming the peptides to generate Gaussian peaks with a constant base width of 12 data points (as found for

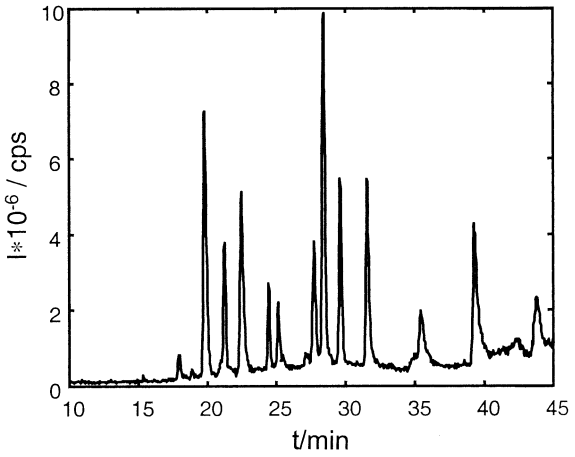

Figure 1. Total ion chromatogram for a Lys-C digest of lactate dehydrogenase from beef heart.

peptide standards in the same experimental set-up). For each run, the maximum intensity for each detection channel was then taken as the one-dimensional, mass spectral representation of the data. As mentioned, this reduction means that the chromatographic information is lost. However, the use of an LC separation step is of high importance for the quality of the mass spectra since the impacts of suppression effects and background problems are minimized. Finally, the obtained data vectors were adjusted to equal sum to reduce the influence from variations in instrumental sensitivity, resulting in a normalized matrix **X** of size 20 × 2000 (corresponding to the 20 samples analyzed and the 2000 detection channels used).

The average derived mass spectra of the four different LDH proteins are shown in Figure 2. It can be seen that most of the ions are present for at least two of the proteins. In theory, a 50% mass spectral overlap is expected for Lys-C digested LDH rabbit muscle and LDH hog muscle, while the overlap between LDH pig

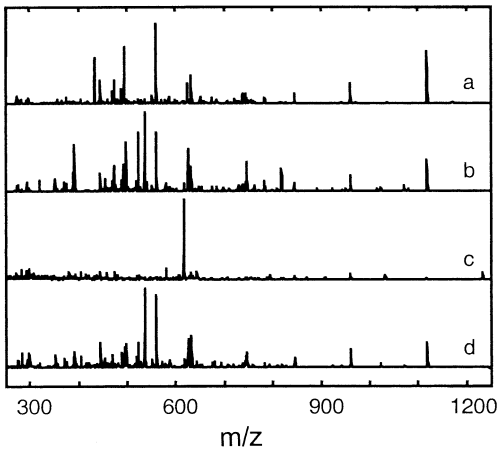

Figure 2. Derived mass spectra for Lys-C digests of lactate dehydrogenase from beef heart (**a**), hog muscle (**b**), pig heart (**c**), and rabbit muscle (**d**).

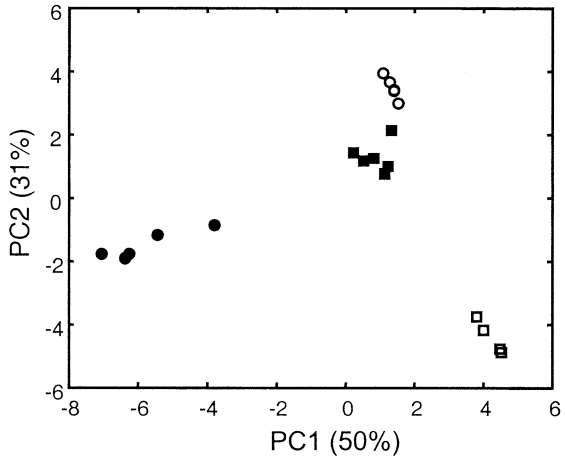

Figure 3. PCA score plot separating the four variants of lactate dehydrogenase studied, i.e., beef heart (open square), hog muscle (open circle), pig heart (filled circle), and rabbit muscle (filled square).

heart and the others should be approximately 20% (no sequence information was found for LDH beef heart). PCA was performed on **X** to determine the chemical rank and to further visualize the relationships between the samples. Six principal components (PCs) were found to be optimal according to a Malinowski F-test [15], and the score plot for PC1 versus PC2 (Figure 3) separated the samples into the different classes (i.e., the different LDHs). It can also be seen that the amino acid sequence similarity between LDH rabbit and hog muscle is reflected by a short distance between these two classes in the score plot.

For discriminative analysis, a response data matrix **Y** of size 20 × 4 was set up, in which the class memberships were binary coded by ones and zeros. However, since the ANN output is constrained between one and zero by the applied transfer function (eq 1), **Y** was scaled between 0.1 and 0.9 during training of the ANNs to avoid the extreme values. Full cross-validation (leave-one-out) was applied to determine the optimal number of components for the PLS2 model as well as the number of hidden nodes in the ANN models. With eight PLS components, all of the samples were correctly classified (Figure 4a). As the input to the ANNs, either the data matrix **X** or the PCA score matrix **T** for six PCs was used. (The construction of the PCA-ANN combination is shown in Figure 5.) In both cases, five hidden nodes were found to be optimal and gave correct classification of all the samples (Figure 4b and c).

The quality of the predictions were measured by the root mean squared error of cross validation (RMSECV), defined as

$$RMSECV = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{m}} \quad (3)$$

where $\hat{y}$ are the predicted values and m is the number of

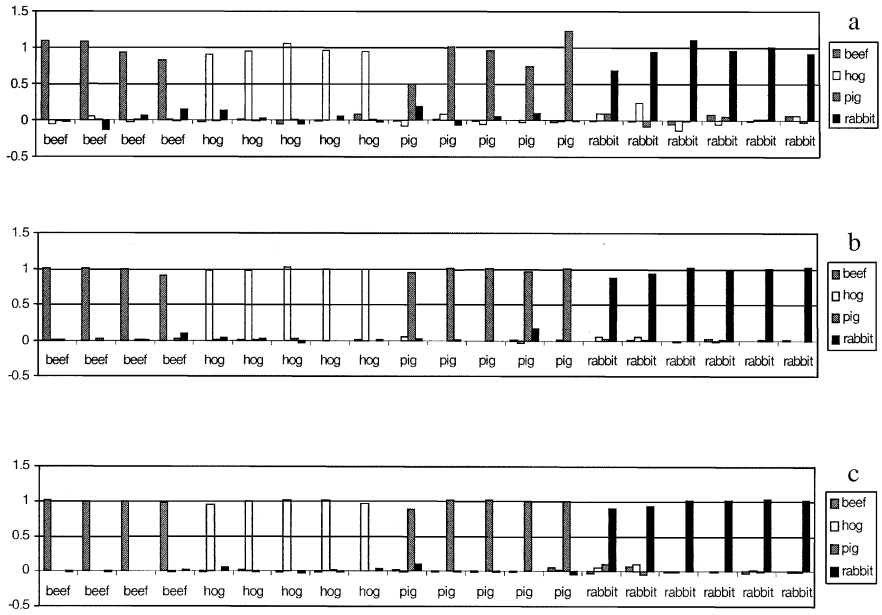

Figure 4. Cross-validation results for the classification of the lactate dehydrogenase samples by discriminative analysis with PLS2 (**a**), ANN (**b**), and PCA-ANN (**c**). The bars correspond to the output of models with the object excluded from the calibration set. 100% accuracy in classification was obtained for all models.

samples. For the ANNs, each $\hat{y}$ was determined five times in order to account for the variability introduced from the fact that the output from an ANN is dependent on the calibration procedure (input weights etc.). The RMSECV values of the optimal ANN models were found to be significantly lower than for the PLS2 model (Table 1). Even though it is somewhat controversial to give a quantitative measure on a qualitative analysis, this indicates that the cut-off limits (i.e., the decision limits for classification) can be set closer to one and zero for the ANN models compared to the PLS model. This fact, which must be a result of the ANNs capability to model non-linearities and/or variable interactions, can also be seen in Figure 4. When selecting the modeling tool, one should be aware that ANNs of the type used in this work are quite sensitive to outliers [16] and often require a high number of calibration samples in order to avoid problems with overfitting. The combination of PCA and ANN can, in some cases, compensate for these problems [17]. Other alternatives are to apply robust training procedures [16] or networks less prone to overfitting.

An advantage with PLS is that the contribution from the variables can be evaluated from loading plots (Figure 6). Similar information can be derived from the weight matrices in the ANN models [18], but not as straightforward as in the PLS case. Another advantage with the projection methods (PLS and PCA) is that new samples not belonging to any of the modeled classes can be identified from unusually high residuals [19].

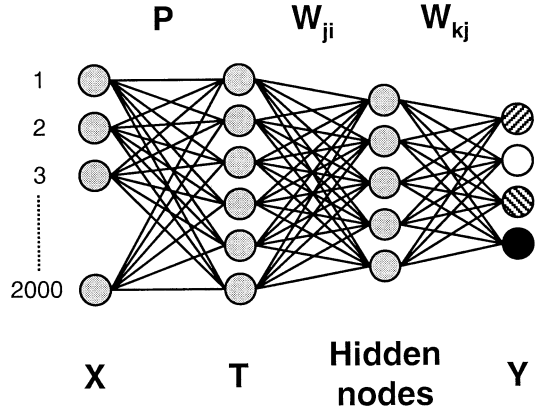

Figure 5. Schematic of the PCA-ANN combination. The 2000 variables in **X** is reduced by PCA into six score values in **T**. By adjusting the weights (**W**) and the number of hidden nodes (five was found to be optimal), the ANN is then trained to recognize the four different LDH variants.

Conclusion

A successful classification of LDH with a robust multivariate quality control tool has been presented. However, before any general conclusions can be made, the method must be applied to larger data sets and more

Table 1. Root mean squared error of cross validation (RMSECV) for the discriminant analysis models used in LDH classification

Model	RMSECV
PLS2	0.103
ANN	0.040±0.032[a]
PCA-ANN	0.034±0.017[a]

[a]95% confidence interval (n = 5).

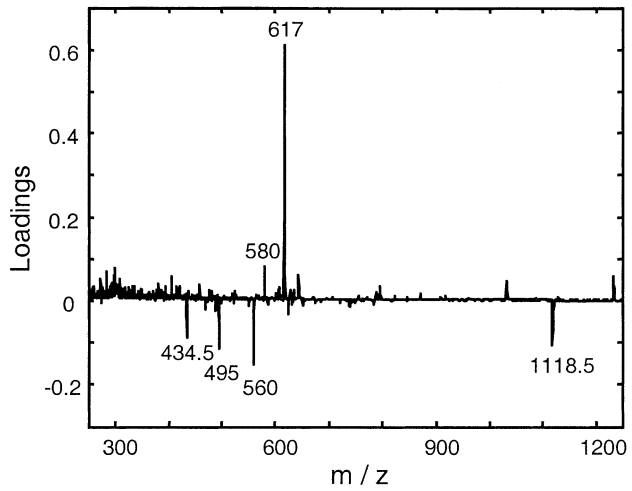

Figure 6. The second loading vector of the discriminative PLS model, revealing that, e.g., *m/z* 617 and 560 are important variables when discriminating among the LDH variants in the second PLS2 component.

complicated classification problems (e.g., the detection of posttranslational modifications), and our work will continue in this direction. The method should also gain from improvements in the digestion step to get a better sequence coverage, and by the use of high resolution MS to enhance the selectivity. Further, it should be of interest to also utilize the chromatographic information in the data analysis.

Acknowledgments

This work was financed by the Swedish Natural Research Council project K-5104-706 and the Swedish Foundation for Strategic Research. The authors gratefully acknowledge Rudolf Kaiser of AstraZeneca, Södertälje, for materials provided and for helpful discussions.

References

1. Garnick, R. L.; Solli, N. J.; Papa, P. A. *Anal. Chem.* **1988,** *60,* 2546–2557.
2. Hoff, E. R.; Chloupek, R. C. *Method. Enzymol.* **1996,** *271,* 51–68.
3. Malmquist, G. *J. Chromatogr. A* **1994,** *687,* 89–100.
4. King, R.; Bonfiglio, R.; Fernandez-Metzler, C.; Miller-Stein, C.; Olah, T. *J. Am. Soc. Mass Spectrom.* **2000,** *11,* 942–950.
5. Jensen, O. N.; Podtjelejnikov, A. V.; Mann, M. *Anal. Chem.* **1997,** *69,* 4741–4750.
6. Aebersold, R.; Goodlett, D. R. *Chem. Rev.* **2001,** *101,* 269–295.
7. Lundell, N.; Schreitmüller, T. *Anal. Biochem.* **1999,** *266,* 31–47.
8. Bylund, D.; Danielsson, R.; Malmquist, G.; Markides, K. E. *J. Chromatogr. A* **2002,** *961,* 237–244.
9. Harrington, P.B.; Voorhees, K. J.; Basile, F.; Hendricker, A. D. *J. Am. Soc. Mass Spectrom.* **2002,** *13,* 10–21.
10. Tan, B. J.; Hardy, J. K.; Snavely, R. E. *Anal. Chim. Acta* **2000,** *422,* 37–46.
11. Eide, I.; Neverdal, G.; Thorvaldsen, B.; Shen, H.; Grung, B.; Kvalheim, O. *Environ. Sci. Technol.* **2001,** *35,* 2314–2318.
12. Ståhle, L.; Wold, S. *J. Chemom.* **1987,** *1,* 185–196.
13. Smits, J. R. M.; Melssen, W. J.; Buydens, L. M. C.; Kateman, G. *Chemom. Intell. Lab. Syst.* **1994,** *22,* 165–189.
14. Danielsson, R.; Bylund, D.; Markides, K. E. *Anal. Chim. Acta* **2002,** *454,* 167–184.
15. Malinowski, E. R. *J. Chemometrics* **1988,** *2,* 49.
16. Walczac, B. *Anal. Chim. Acta* **1996,** *322,* 21–29.
17. Defernez, M.; Kemsley, E. K. *Analyst* **1999,** *124,* 1675–1681.
18. Andersson, F. O.; Åberg, M.; Jacobsson, S. P. *Chemometrics Intell. Lab. Syst.* **2000,** *51,* 61–72.
19. Wise, B. M.; Gallagher, N. B. *J. Proc. Cont.* **1996,** *6,* 329–348.