Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography B, 803 (2004) 131–135
JOURNAL OF CHROMATOGRAPHY B
www.elsevier.com/locate/chromb

Protein identification by liquid chromatography–mass spectrometry using retention time prediction

Magnus Palmblad[a,b,*], Margareta Ramström[a,c], Christopher G. Bailey[d], Sandra L. McCutchen-Maloney[e], Jonas Bergquist[c], Loreen C. Zeller[d]

[a] *The Ångström Laboratory, Division of Ion Physics, Uppsala University, Box 534, SE-751 21 Uppsala, Sweden*
[b] *Center for Accelerator Mass Spectrometry, Lawrence Livermore National Laboratory, Livermore, CA 94551-0808, USA*
[c] *Department of Analytical Chemistry, Institute of Chemistry, Uppsala University, Box 599, SE-751 24 Uppsala, Sweden*
[d] *Chemistry and Materials Science, Lawrence Livermore National Laboratory, Livermore, CA 94551-0808, USA*
[e] *Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551-0808, USA*

Abstract

Liquid chromatography has been coupled with mass spectrometry to improve the dynamic range and to reduce the complexity of sample introduced to the mass spectrometer at any given time. The chromatographic separation also provides information on the analytes, such as peptides in enzymatic digests of proteins; information that can be used when identifying the proteins by peptide mass fingerprinting. This paper discusses a recently introduced method based on retention time prediction to extract information from chromatographic separations and the applications of this method to protein identification in organisms with small and large genomes.

Keywords: Protein identification; Retention time prediction

1. Introduction

Mass spectrometry (MS) is one of the most important and versatile tools in proteomics. Proteins separated by electrophoresis or chromatography in one or more dimensions are commonly digested off-line by a proteolytic enzyme such as trypsin and the fragment peptides analyzed by matrix assisted laser desorption/ionization (MALDI) or electrospray ionization (ESI) time-of-flight (TOF) or ion trap (IT) mass spectrometry [1,2]. Capillary electrophoresis (CE) MS [3,4] or liquid chromatography (LC) MS [5–7] coupled to mass spectrometry through ESI have been used to identify proteins in complex biological samples such as whole cell lysates [4,8–11] and human body fluids [12–14]. The main purpose of combining a liquid separation method with an electrospray mass spectrometer is to reduce the complexity of sample introduced to the mass spectrometer at any given time, resulting in an enhanced dynamic (concentration) range.

When infusing a sample containing multiple components directly into the mass spectrometer, the dynamic range of the measurement is limited by ion suppression in the electrospray ionization process and in detection. Ion trap and Fourier transform ion cyclotron resonance (FTICR) mass spectrometers have limited ion storage capacities and hence limited dynamic range. The latter can be partially overcome by selectively loading the FTICR cell using a mass selective quadrupole [15], although such instruments have only recently become commercially available. In addition to an improved dynamic range, the separation itself provides information on the analytes. In reversed phase chromatography, the information describes the hydrophobicity of the peptides [16,17].

For a given measured tryptic peptide mass and measurement accuracy, there are a finite number of theoretical tryptic peptides from proteins in a sequence database that are within the mass measurement error of the measured mass [18,19]. If the mass accuracy is very high, i.e. errors below 1 ppm, very nearly that which can be achieved in high-field FTICR mass spectrometry under ideal conditions [20], there may exist a tryptic peptide of each protein for which there is only one candidate peptide within the mass measurement error, even if the candidates are calculated from all proteins in the organism [11,19]. These peptides can then be used as 'accurate mass tags' for protein identification [19]. In general,

* Corresponding author. Tel.: +1-925-422-8462; fax: +1-925-423-7884.
E-mail address: palmblad1@llnl.gov (M. Palmblad).

1570-0232/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2003.11.007

however, mass accuracy is insufficient to identify proteins based on a single tryptic peptide mass, requiring either several peptides or additional information on the peptides for unambiguous protein identification.

We have previously shown [21] how to combine information from chromatographic retention time and accurate mass measurement to improve protein identification by peptide mass fingerprinting and LC/FTICR. This was demonstrated on the human cerebrospinal fluid (CSF) proteome. Actual retention times are compared with those predicted for peptides generated from a sequence database. This approach has also been successfully applied to a large number of LC/FTICR datasets from microbial proteomes by Petritis et al. [22] who used normalized retention times and an artificial neural network (ANN) with the sigmoidal transfer and output functions. The improvement from adding non-linear nodes in a hidden layer was found to be small. In fact, the linear model in Palmblad et al. [21] and the ANN in Petritis et al. [22] were almost equivalent, since Petritis et al. mapped retention times to the (nearly) linear part of the sigmoid output function when training the network. Optimizing the part of the output function to which measured data is mapped, as well as using a larger number of input neurons, taking the amino acid sequence into account, should improve the accuracy of the predictor. The information from a liquid separation is complementary to that from mass spectrometry, as long as peptides are not separated by molecular weight, such as in size-exclusion chromatography or denaturing gel electrophoresis.

2. Experimental

To assess the performance of the peptide retention time predictor for protein identification, proteins extracted from human cerebrospinal fluid obtained from healthy donors were digested with trypsin (Roche Diagnostics GmbH (Mannheim, Germany), sequencing grade) as described by Bergquist et al. [12]. *Yersinia pestis* membrane and cytosol protein fractions were obtained and similarly digested. The predictor for the column and gradient used to analyze the samples of interest was calibrated using standards consisting of bovine serum albumin (BSA) (Sigma, St. Louis, MO, USA) digested in-house, and BSA and 11 other large proteins obtained as tryptic digests (Michrom BioResources, Auburn, CA, Part No. 910/00012/35).

Reversed-phase chromatography was performed as described previously [21] (BSA and CSF tryptic digests) or with a direct pumping Eksigent Technologies (Livermore, CA) gradient NanoFlow LC system using 10 cm long, 75 μm i.d., 365 μm o.d., 5 μm BioBasic® C18 packing column pulled to a 15 μm tip (New Objective Inc., Woburn, MA). The dual pumps delivered stable flows at 300 nl/min in 90 min linear gradients from 5 to 60% organic solvent (90% ACN, 0.1% HAc) versus 0.1% HAc in H_2O. Sample volumes of 1–5 μl, containing approximately 20–50 ng of protein digest, were injected using a 10 μl sample loop. For the analysis of BSA and CSF digests, the LC system was connected to a Bruker Daltonics APEX II 9.4 T FTICR mass spectrometer (Bruker Daltonics, Billerica, MA, USA) through the modified nanospray interface [14] on an Analytica ESI source (Analytica, Branford, CT, USA) [23]. For the large protein digest mixture and the *Y. pestis* membrane proteins, the Eksigent LC system was connected to a different Bruker APEX II 9.4 T FTICR mass spectrometer through a homebuilt nanospray interface on an Apollo (Bruker Daltonics, Billerica, MA, USA) ESI source. For the purpose of retention time prediction and protein identification, these experimental configurations are very similar.

Predicted retention times of peptides based on amino acid composition were compared with the measured values from the chromatographic separations [24–29]. Typically, the experimentally measured masses of the peptides were within 5–7 ppm of those predicted from the protein database (*Homo sapiens*, *Y. pestis*, standard proteins) [23]. The likelihoods of all matching peptide masses and retentions were summed to a total likelihood score, which was used to discriminate between true and random protein matches. The non-randomness of enzymatic digestion of proteins was also factored in, as described previously [12].

We used a simple linear model of peptide behavior in reversed phase, similar to that used by Hodges and co-workers [26–28]. Tryptic peptides from standard proteins as well as abundant, previously identified proteins in human cerebrospinal fluid and *Y. pestis* samples were used to calculate a retention coefficient for each amino acid according to

$$t_{\mathrm{calc}} = \sum_{i=1}^{20} n_i c_i + t_0 \quad (1)$$

where c_i are the retention coefficients for the 20 amino acids, n_i the number of each amino acid and t_0 compensates for void volumes and a delay between sample injection and acquisition of mass spectra. Eq. (1) compensates for variability in void volume and time offsets in t_0 and implicitly for gradient slopes as a common factor in all c_i. Alternatively, all datasets can first be normalized with respect to each other [22]. The coefficients c_i correspond to the weights in the 20-0-1 neural network tested by Petritis et al. [22].

The c_i and t_0 were optimized by least-squares fitting t_{calc} to measured retention times of 100–200 standard protein peptides or peptides from abundant proteins in CSF and *Y. pestis* fractions, putatively identified by accurate mass measurement and high relative intensities in the mass spectra. All software was written in C, and run on standard single-processor personal computers under the Cygwin API [30].

3. Results and discussion

Fig. 1 shows an LC/FTICR mass chromatogram of a *Y. pestis* membrane fraction tryptic digest. Typically,

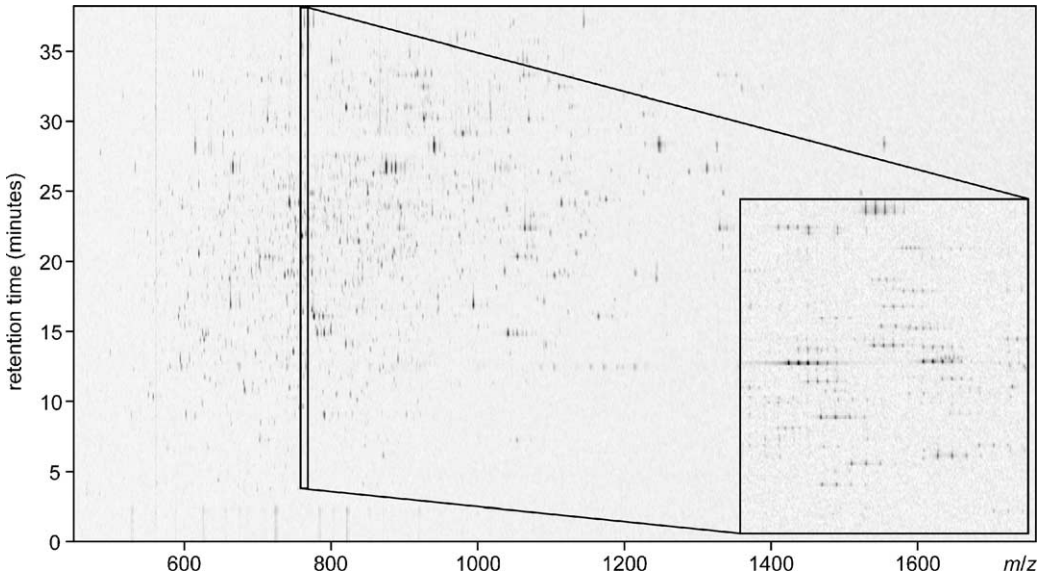

Fig. 1. A mass chromatogram of a *Yersinia pestis* membrane fraction tryptic digest showing 9285 peaks from 1828 individual peptides. The inset shows the region *m*/*z* 760–770 with many overlapping isotopic clusters and several near-isobars.

10,000–100,000 individual peaks were reduced to 1000–10,000 unique peptide masses. The redundancy was caused by sampling the same peptide in several mass spectra, in multiple charge states and in multiple isotopic peaks. Tryptic peptides closer than 0.005 in *m*/*z* can be resolved in these instruments [31], but use of a chromatographic separation enhances the practical dynamic range for peptides close in *m*/*z*. The inset shows how several near-isobars can be separated and detected. The retention time predictor then assists in putatively identifying these peptides.

A training set containing a large number of peptides is required for c_i and t_0 to converge. The c_i values correlated with hydrophobicity as expected from literature, i.e. the more hydrophobic, the higher the retention coefficient c_i. The distributions of normalized retention coefficients derived from five different CSF runs are shown in Fig. 2. The t_0 values were usually near the observed void time. The overall positive c_i implies a size dependency of the retention of tryptic peptides, i.e. the longer the peptide, the longer the retention time on the column. This is also consistent with results shown in Fig. 1. Small and internal tryptic peptides are relatively hydrophilic as they all have a basic C-terminal residue (arginine or lysine). Any linear dependence on peptide length is implicitly encoded by the retention

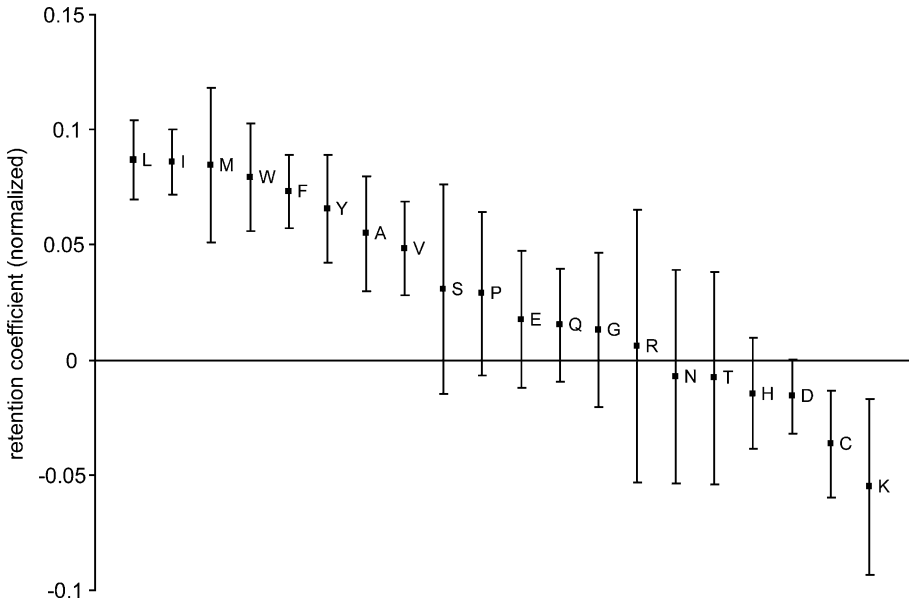

Fig. 2. Retention coefficients for the 20 amino acids, normalized to the sum of the absolute values of all retention coefficients, for five CSF samples with 77–142 peptides ($\mu = 103$). The cysteines (C) had been carbamidomethylated in all peptides used to train the predictor.

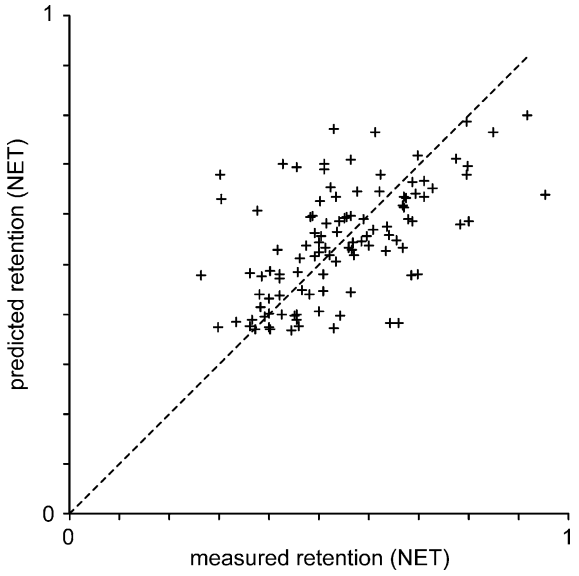

Fig. 3. Predicted vs. measured normalized elution time (NET) [22] of peptides from the 12-protein digest mixture. The predictor was trained on one dataset and evaluated on another, using masses present in both datasets. A significant fraction of the mass based peptide assignments are either erroneous or the elution times poorly predicted.

coefficients in this model as a constant term added to each c_i.

The inaccuracy of the predictor was found to be around 10% for internally calibrated datasets. This may be limited by the presence of false peptides in the training set, as also suggested by Petritis et al. [22]. The retention times for the same peptides in the same sample ran multiple times were within 2–3%. Fig. 3 shows the predicted versus measured normalized elution times (NETs) [22] for peptides from the 12-protein digest. The predictor was trained on one dataset and evaluated on another. The two datasets were acquired under identical conditions.

The predictor increased the number of proteins identified at any given level of confidence. In both CSF and *Y. pestis* samples, the set of tentatively identified proteins at a given probability of having one or more false positive(s) in the set approximately doubled when using the predictor. These probabilities were estimated by comparing the likelihood scores with the distribution of scores of a reference database. This reference database consisted of around 40,000 protein sequences obtained from all publicly available and fully sequenced archaea, which are evolutionarily distant from both *Y. pestis* and *H. sapiens*. The differences in protein size distributions between domains and organisms could be taken into account by randomly filtering out sequences with a size dependent frequency. This should provide a conservative estimate on the statistical significance, as there are some highly conserved proteins with a significant sequence homology between the protein in the reference database and the organism of interest. This is different from the accurate mass tag approach [19], which requires extensive validation using MS/MS and precise knowledge of many experimental parameters, such as the frequency of non-tryptic peptides, modified peptides, peptides from other organisms (i.e. contaminating peptides), database sequence errors, mass measurement accuracy as a function of mass or m/z, etc. Given sufficient information on these, the accurate mass (and time) tag approach is superior, but without such information, the confidence in a given set of protein identifications is hard to estimate.

Table 1
Set of 12 proteins (Swiss-Prot TrEMBL accession number and description) identified in a *Y. pestis* membrane protein extracts using LC-FTICR

SPTr AC	Description
Q8ZAR6	Isocitrate lyase
Q8D014	Hypothetical protein
Q8D056	Outer membrane porin A
Q8D1D4	Integral membrane peptidase
Q8ZFI8	Probable *N*-acetylmuramoyl-L-alanine amidase
Q8D0Z7	Outer membrane protein X
Q8D0Y2	Lipoprotein-34
Q8ZC69	Peptidyl-prolyl *cis–trans* isomerase D
Q8CKQ7	Hypothetical protein
Q8ZF72	Putative cystine-binding periplasmic protein
Q8ZI46	Topoisomerase IV subunit B
Q8ZB96	Putative exported protein

Typically, 10–30 proteins could be identified with a small number of expected false positives in human cerebrospinal fluid and in the *Y. pestis* samples. Table 1 shows 12 proteins repeatedly identified in *Y. pestis* membrane fractions using this method.

Endogenously high abundant proteins, such as human serum albumin (HSA) and transferrin in CSF and abundant proteins in the *Y. pestis* membrane fractions, were used to calibrate the predictor to each dataset. If only a few peptides in the sample were known, these peptides could be mapped onto the same peptides in the dataset used to train the predictor and the retention coefficients scaled accordingly. Alternatively, the retention times themselves can be normalized to each other using a few peptides present in all datasets [22]. The retention coefficients are likely to be dependent on chromatographic conditions such as mobile phase composition and pH. The pH directly influences the charge of peptide side chain groups and termini and hence the hydrophobicity (charged residues are more hydrophilic) and retention coefficients in reversed-phase chromatography [24,29].

In addition to a large number of peptides in the training set, each amino acid must be present in several peptides in this set. The models used so far are not the only conceivable models for predicting reversed-phase chromatographic retention of peptides. More sophisticated approaches would account for the actual sequence or even the predicted secondary structure [32]. For instance, residues predicted to interact with the surroundings would be given higher weights in Eq. (1).

4. Conclusions

Information on physicochemical properties of peptides, such as hydrophobicity, that can be predicted from the peptide sequence assists protein identification by peptide mass fingerprinting. The information used is already available in all LC/MS experiments, whether on-line via electrospray ionization, or in LC-MALDI off-line. The approach may be equally applicable to other types of separations, although prediction of electrophoretic migration is not as straightforward as reversed-phase chromatographic retention [29,33–35]. The predictor described here has been successfully incorporated in analysis software and used for screening clinically sampled cerebrospinal fluid as well as membrane protein fractions from bacteria grown under various conditions.

Acknowledgements

The authors are grateful for funding from the Swedish Natural Sciences Research Council (Grant K-1618/1999), the Swedish Research Council (Grant 13123, 621-2002-5261, 629-2002-6821), the Swedish Society for Medical Research, the Knut and Alice Wallenberg Foundation and LLNL LDRD 01-SI-002. The authors also wish to thank Ardeshir Amirkhani at the Department of Analytical Chemistry and Gloria Murphy and Arlene Gonzales of the Biology and Biotechnology Research Program for technical assistance. The work was performed in part under the auspices of the US Department of Energy by University of California Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

References

[1] S.D. Patterson, R. Aebersold, Electrophoresis 16 (1995) 1791.
[2] A. Shevchenko, O.N. Jensen, A.V. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, H. Boucherie, M. Mann, Proc. Natl. Acad. Sci. U.S.A. 93 (1996) 14440.
[3] R.D. Smith, J.A. Loo, C.J. Barinaga, C.G. Edmonds, H.R. Udseth, J. Chromatogr. 480 (1989) 211.
[4] R.D. Smith, L. Pasa-Tolic, M.S. Lipton, P.K. Jensen, G.A. Anderson, Y. Shen, T.P. Conrads, H.R. Udseth, R. Harkewicz, M.E. Belov, C. Masselon, T.D. Veenstra, Electrophoresis 22 (2001) 1652.
[5] C.M. Whitehouse, R.N. Dreyer, M. Yamashita, J.B. Fenn, Anal. Chem. 57 (1985) 675.
[6] C.C. Stacey, G.H. Kruppa, C.H. Watson, J. Wronka, F.H. Laukien, J.F. Banks, C.M. Whitehouse, Rapid Commun. Mass Spectrom. 8 (1994) 513.
[7] R.D. Voyksner, in: R.B. Cole (Ed.), Electrospray Ionization Mass Spectrometry, Wiley, New York, 1997, pp. 323–341.
[8] P.K. Jensen, L. Pasa-Tolic, K.K. Peden, S. Martinovic, M.S. Lipton, G.A. Anderson, N. Tolic, K.K. Wong, R.D. Smith, Electrophoresis 21 (2000) 1372.
[9] Y. Shen, N. Tolic, R. Zhao, L. Pasa-Tolic, L. Li, S.J. Berger, R. Harkewicz, G.A. Anderson, M.E. Belov, R.D. Smith, Anal. Chem. 73 (2001) 3011.
[10] T.P. Conrads, K. Alving, T.D. Veenstra, M.E. Belov, G.A. Anderson, D.J. Anderson, M.S. Lipton, L. Pasa-Tolic, H.R. Udseth, W.B. Chrisler, B.D. Thrall, R.D. Smith, Anal. Chem. 73 (2001) 2132.
[11] R.D. Smith, G.A. Anderson, M.S. Lipton, C. Masselon, L. Pasa-Tolic, Y. Shen, H.R. Udseth, Omics 6 (2002) 61.
[12] J. Bergquist, M. Palmblad, M. Wetterhall, P. Håkansson, K.E. Markides, Mass Spectrom. Rev. 21 (2002) 2.
[13] M. Wetterhall, M. Palmblad, P. Håkansson, K.E. Markides, J. Bergquist, J. Prot. Res. 1 (2002) 361.
[14] M. Ramström, M. Palmblad, K.E. Markides, P. Håkansson, J. Bergquist, Proteomics 3 (2003) 184.
[15] M.E. Belov, E.N. Nikolaev, G.A. Anderson, H.R. Udseth, T.P. Conrads, T.D. Veenstra, C.D. Masselon, M.V. Gorshkov, R.D. Smith, Anal. Chem. 73 (2001) 253.
[16] J. Frenz, W.S. Hancock, W.J. Henzel, C. Horváth, in HPLC of Biological Macromolecules: Methods and Applications, Marcel Dekker, New York, 1990, p. 145.
[17] J.L. Cornette, K.B. Cease, H. Margalit, J.L. Spouge, J.A. Berzofsky, C. DeLisi, J. Mol. Biol. 195 (1987) 659.
[18] R.A. Zubarev, P. Håkansson, B.U.R. Sundqvist, Anal. Chem. 68 (1996) 4060.
[19] T.P. Conrads, G.A. Anderson, T.D. Veenstra, L. Pasa-Tolic, R.D. Smith, Anal. Chem. 72 (2000) 3349.
[20] J.E. Bruce, G.A. Anderson, J. Wen, R. Harkewicz, R.D. Smith, Anal. Chem. 71 (1999) 2595.
[21] M. Palmblad, M. Ramström, K.E. Markides, P. Håkansson, J. Bergquist, Anal. Chem. 74 (2002) 5826.
[22] K. Petritis, L.J. Kangas, P.L. Ferguson, G.A. Anderson, L. Pasa-Tolic, M.S. Lipton, K.J. Auberry, E.F. Strittmatter, Y. Shen, R. Zhao, R.D. Smith, Anal. Chem. 75 (2003) 1039.
[23] M. Palmblad, K. Håkansson, P. Håkansson, X. Feng, H.J. Cooper, A.E. Giannakopulos, P.S. Green, P.J. Derrick, Eur. Mass Spectrom. 6 (2000) 267.
[24] J.L. Meek, Proc. Natl. Acad. Sci. U.S.A. 77 (1980) 1632.
[25] D.C. Guo, C.T. Mant, R.S. Hodges, J. Chromatogr. 386 (1987) 205.
[26] M.T. Hearn, M.I. Aguilar, C.T. Mant, R.S. Hodges, J. Chromatogr. 438 (1988) 197.
[27] R.S. Hodges, J.M. Parker, C.T. Mant, R.R. Sharma, J. Chromatogr. 458 (1988) 147.
[28] C.T. Mant, N.E. Zhou, R.S. Hodges, J. Chromatogr. 476 (1989) 363.
[29] V. Sanz-Nebot, I. Toro, F. Benavente, J. Barbosa, J. Chromatogr. A 942 (2002) 145.
[30] http://www.cygwin.com.
[31] M. Palmblad, M. Wetterhall, K. Markides, P. Håkansson, J. Bergquist, Rapid Commun. Mass Spectrom. 14 (2000) 1029.
[32] B. Rost, J. Struct. Biol. 134 (2001) 204.
[33] P.D. Grossman, J.C. Colburn, H.H. Lauer, Anal. Biochem. 179 (1989) 28.
[34] A. Cifuentes, H. Poppe, Electrophoresis 18 (1997) 2362.
[35] M. Castagnola, D.V. Rossetti, M. Corda, M. Pellegrini, F. Misiti, A. Olianas, B. Giardina, I. Messana, Electrophoresis 19 (1998) 2273.