

Improvement of Peptides Identification in Proteomics with the Use of New Analytical and Bioinformatic Strategies

Tomasz Baczek*

Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gdańsk, Poland

Abstract: Completion of the Human Genome Project enabled a better understanding of biological functions of organisms. However, these studies still provide a limited insight into the cellular processes. Nowadays, a comprehensive analysis and characterization of all expressed proteins, called proteomics, is the point of the interest. One of the important issues in proteomics is finding of analytical and bioinformatic strategies allowing unambiguous protein identification based on the searching of the peptide sequence databases. Some examples of bioinformatic strategies for analytical data processing obtained with the use of separation techniques and mass spectrometry analysis are given to demonstrate their usefulness in proteomics. First, the application of learning algorithms for the reliable evaluation of MS/MS spectra of peptides, which were separated and processed with reversed-phase liquid chromatography-tandem mass spectrometry is discussed. Detailed considerations of the use of artificial neural networks analysis to classify automatically peptide MS/MS spectra is provided and analyzed in the aspect of utility of another learning algorithms. Moreover, the usefulness of predictions of the reversed-phase liquid chromatography retention times of peptides in proteomic research is reported. In that case, quantitative structure-retention relationships (QSRR) analysis is considered in the view of the other approaches used in this field. Finally, the contribution of analytical information from the pI-based separation methods is considered as the additional source of peptide database matching constraint.

Keywords: Proteomics, Bioinformatics, Analytical chemistry, Review.

1. INTRODUCTION

The review describes some analytical and bioinformatic aspects of proteomics – globally actual, widely discussed subject of biological and bioanalytical interest. Nowadays, scientists are able to determine gene expression. Genomics gets spectacular, scientifically very significant successes. Scientists from noncommercial, public Human Genome Project (HGP) [1] and researchers from biotechnological company Celera Genomics [2] introduced independently the first identification of human genome code. However, the studies associated with genomics being very important scientifically seem to provide only a limited view of cellular processes. Life and death of the cells are strongly dependent on gene expression and activity of their products created on the basis of genetic information – proteins. A comprehensive analysis and characterization of all expressed proteins (proteome) has been given the name proteomics [3]. Proteins are the main catalysts of the biological functions and reflect the actual, not potential, like in the case of information comprised in genome, condition of the cell or organism in the certain moment of time. Unfortunately, there is a poor correlation between gene expression and protein expression. It complicates quite significantly proteomic analysis. The reason for that are the continuous changes in the cell on the level of proteins and the existence of the same protein in several forms because of, for example posttranslational

modifications (phosphorylation, glycosylation, etc.). In fact, the number of human genes is estimated in the range of 30-40 thousands [1], but the number of proteins can be even 10-20 times higher. The task to analyze all proteins complicates additionally the existence of the broad range of concentration, in which they can exist (6-10 orders of magnitude), high dissimilarities in the range of their physiological functions, the lack of the methods enabling the replication of proteins or peptides in analogous way as in the case of DNA replication with the use of polymerase chain reaction (PCR), as well as variety of their biological functions [4,5]. Proteome analysis is nowadays on the similar development as genome analysis in 1990s. Therefore, the further development of the new bioanalytical technologies and continuous improvement already available ones is necessary to obtain the level of proteome research observed in genomics.

Despite of the typical analytical difficulties as well as in the range of manipulations of the huge amount of information obtaining during proteomic research, the attempts of proteome characterization are nowadays realized in laboratory practice. The first important analytical tool used in proteomic research are techniques for the fractionation and separation of proteins and peptides. Besides, the separation process allows for the observation of the eventual differentiation in proteome during the comparative analysis. At last, separations of proteins enable also the selective extraction of the appropriate protein from the complex mixture [6,7]. Two-dimensional gel electrophoresis is recognized as the first and still the most popular separation technique in proteomics [8-11]. For example that technique in combination with mass

*Address correspondence to this author at the Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gdańsk, Poland; Tel: (48) (58) 3493260; Fax: (48) (58) 3493262; E-mail: tbaczek@amg.gda.pl

spectrometry enabled to determine and characterize several human health-related proteins in: lymphoblastoids [12], vitreous humor [13], bronchoalveolar lavage fluid [14] cerebrospinal fluid [15]. Two-dimensional gel electrophoresis helped to characterize proteins in dental tissues [16], human blood serum and plasma [17,18], and human allergens of protein origin [19]. However, the application of proteomic research for the evaluation of biological-medical processes in the view of rational design of new drugs (including orphan drugs) needs the application of separation techniques, which enable the analysis the significant number of samples in relatively short time. The main constrain of two-dimensional gel electrophoresis is impossible identification of the whole proteome with that method. Large and hydrophobic proteins have difficulties in the movement through the gel. Acidic and basis proteins are also poor resolved. Proteins existing in the small concentration in sample are usually below the limit of detection of that technique. That limitation is probably the most significant, because a lot of regulatory proteins, which could play potentially important role in disease and health processes and could be drug targets, are present in cells in very low concentration.

2. CHROMATOGRAPHIC TECHNIQUES IN PROTEOMICS

It is well known that efficient separation prior to mass spectrometry (MS) and proper database searching greatly facilitates identification of proteins [7]. Such high-resolution separation techniques as two-dimensional chromatography: ion exchange chromatography (IEC) combined with reversed phase liquid chromatography (RPLC) [20-25], size-exclusion chromatography (SEC) combined with RPLC [26], and RPLC combined with capillary zone electrophoresis (CZE) [27]) coupled to mass spectrometry, are currently under intensive development and evaluation in proteomics.

Currently, liquid chromatography is primarily used to separate complex mixtures of peptides and proteins whereas mass spectrometry is the method of choice for their identification. Liquid chromatography coupled to tandem mass spectrometry (LC/MS/MS) is a standard equipment set used to identify the components of protein complexes and subcellular compartments. This technique is capable to identify hundreds of biomolecules, like for example identification of the components of yeast and human ribosomes [20], or the whole yeast cells [21, 22].

Chromatographic techniques have several valuable features in comparison to one- or two-dimensional gel electrophoresis and can be successfully use for fractionation and further separation of proteins and peptides. Large capability of the columns in the modern high-performance liquid chromatography (HPLC) is compatible with the needs of effective preparative separations. It is possible also to concentrate the large volumes of samples onto the column without the dangerous waste in resolution of the individual analytes. On the other hand, nowadays available tandem mass spectrometers with electrospray ionization (ESI-MS/MS) are coupled with microcapillary high-performance liquid chromatography [28]. Miniaturization of the chromatographic columns in the case of LC-ESI-MS/MS

enables to increase selectivity and effectiveness of the separation of complex mixtures. It is caused by decreasing of the column dimension and lower flow rate of the mobile phase, what proportionally increase resolving power of the column. At flow rates of the mobile phase at the range of nL/min also effectiveness of the electrospray in ESI-MS/MS is increased, what generates additional increasing in sensitivity. Hence, using micro- or nanocapillary high-performance liquid chromatography coupled to ESI mass spectrometry the whole analysis of the sample with peptides is dramatically increased, including better effectiveness in resolution, recovery, ionization and detection limit [29].

The characterization of the whole set of proteins in the appropriate cell or tissue in the certain moment of time at the specific physiological or pathophysiological conditions is the main goal of the proteomic experiment. For that case the researchers first try to separate proteins, then they are treated with the appropriate proteolytic enzyme (e.g. trypsin), what provides the generation of peptides. Usually proteins can be digested according to a standard protocol. Briefly, appropriate amount of protein or proteins is denatured in a solution containing 7 M urea, 50 mM Tris and 3 mM dithiothreitol at 60°C for 60 min. After denaturation, the mixture is allowed to cool and iodoacetamide is added to a final concentration of 15 mM, and placed in the dark for 30 min at room temperature. After dilution with 50 mM ammonium bicarbonate until urea concentration is below 1M, trypsin is added at an enzyme:protein ratio of 1:50 (w/w). Then incubation at 37°C is performed overnight. Peptides can be then analyzed immediately with the use of mass spectrometry or further separated before the final identification on the basis of mass spectra (Fig. 1).

Proteolytic digestion with the use of trypsin generates the existence of about 50 peptides from one protein. It means that the final sample with all possible proteins from for example the cells of baker's yeast (*Saccharomyces cerevisiae*) with expected on the basis of genome about 6000 proteins can comprise of at least about 300 thousands of peptides. However, even the best single chromatographic separation system is not able to separate such a complex sample completely satisfactory. Therefore, one can meet nowadays the complex chromatographic systems, in which the separation system is based on different physicochemical properties of analytes. It enables finally to obtain the appropriate separation prior to the identification analysis [4]. For example, one can utilize ion-exchange chromatography, which reflects the first dimension in two-dimensional gel electrophoresis with the separation based on the differences between the charges on the analyte molecules, and reversed-phase chromatography, which reflects the second dimension and the separation is based on differences in hydrophobicity of the analytes (Fig. 2).

In that way, multidimensional liquid chromatography was created, which coupled to compatible ESI-MS/MS or matrix-assisted laser desorption ionisation time-of-flight mass spectrometry (MALDI-TOF-MS) is nowadays the most often used tool in proteomic research [20-27,30,31].

Several kinds of multidimensional liquid chromatography can be distinguished now. Special attention could be taken on [6]: i) multidimensional liquid chromatography with

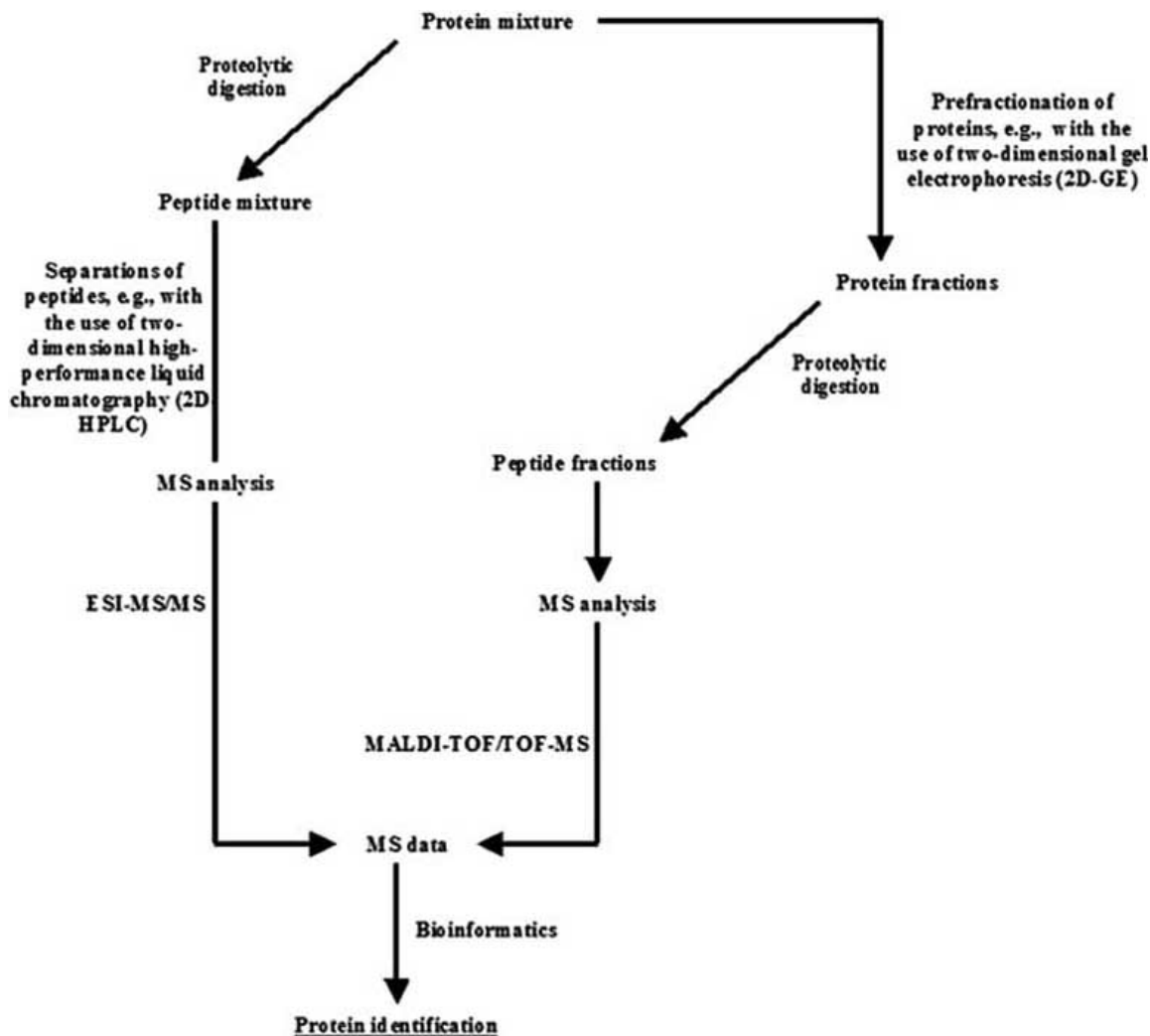


Fig. (1). General analytical strategies in proteomics based on the fractionation and separation techniques, mass spectrometry and bioinformatic data processing.

fraction collection, *ii*) multidimensional liquid chromatography with one column comprising two different stationary phases, *iii*) multidimensional liquid chromatography with column switching. The first method was used for example to determine proteins in yeast cells (*Saccharomyces cerevisiae*). Proteins were labeled with ICAT reagent (isotope-coded affinity tags). Three-dimensional chromatography was used finally to separate generated protein digest. First, peptides were fractionated on cation-exchange column. Next, isolation of the labeled peptides was performed with affinity column. Prior to identification with the use of mass spectrometry peptides were separated with reversed-phase column [30, 31]. Multidimensional liquid chromatography with one column comprising two different stationary phases is called as MudPIT (multidimensional protein identification technology) [20, 21]. In that technology, there is strong cation-exchange stationary phase prior to reversed-phase stationary phase packed in one column. Chromatography system is combined directly to mass spectrometry. With the use of that technology 1484 [21] and 1504 [22] proteins of *Saccharomyces cerevisiae* were detected and identified.

Column switching with the use of novel silica-based restricted access materials (RAM) with ion-exchange functionalities in the first dimension and reversed-phase column in the second dimension was used to protein mapping of biological samples of human hemofiltrate as well as of cell lysates originating from a human fetal fibroblast cell line [23].

Mass spectrometry in the last 20 years reached very high level of scientific and technological development and is now very sensitive and reliable tool for the analysis of biomolecules. That technique is very useful in proteomics because of the possibility to perform some very important analyses. First of all, thanks to mass spectrometry it is possible to achieve very accurate measurements of molecular mass for peptides and proteins. Mass spectrometry is nowadays the best method for molecular mass measurement of proteins and peptides besides the methods based on the migration of biomolecules in polyacrylamide gels in gel electrophoresis. However, even the most accurate measurements of molecular mass value of the certain protein or peptide with the use of MS spectrum has the limited

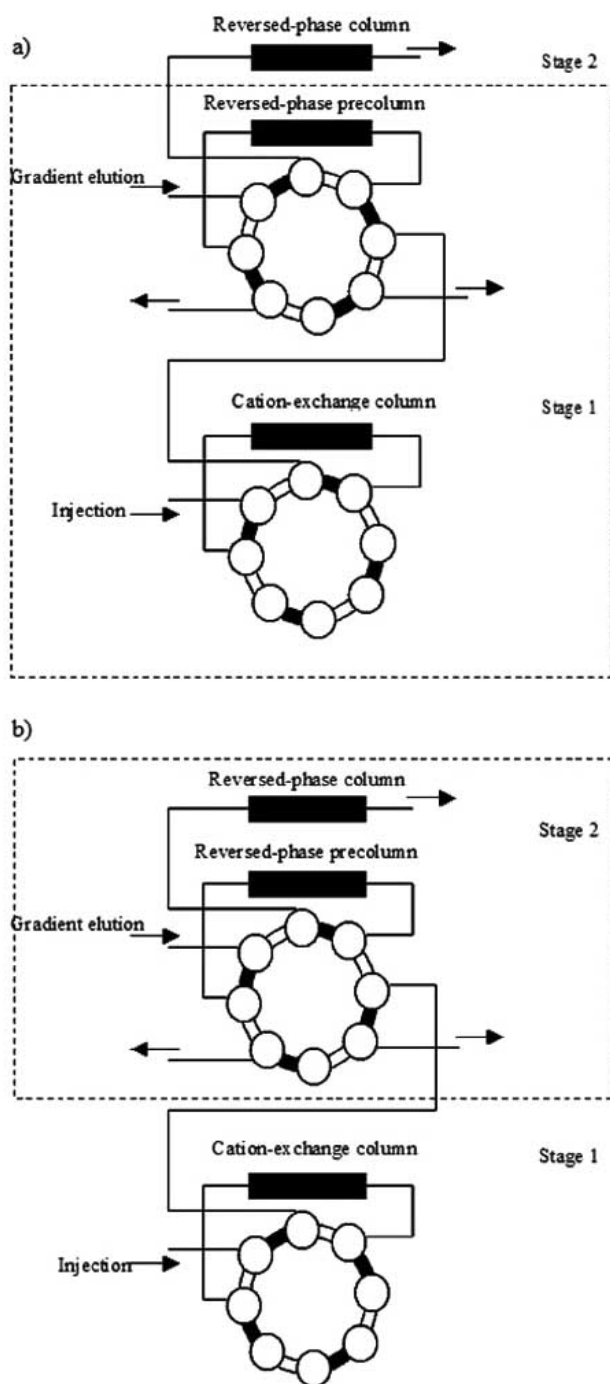


Fig. (2). Schematic representation of two-dimensional liquid chromatography: a) valves in position for loading of the sample onto the cation-exchange column; b) valves in position for deposition of the sample from reversed-phase precolumn after the separation in stage 1 onto the reversed-phase column (stage 2).

application (especially considering very complex mixtures of proteins, which are of proteomics interest), because it can be insufficient for unambiguous identification of the protein or peptide. On the other hand, mass spectrometry can be also useful in sequence analysis of peptides applying MS/MS mass spectra. Analysis with the use of MS/MS spectra enables to identify the certain peptide on the basis of their

amino acid sequence, what provides the higher probability of the identification of the origin protein [32,33].

2.1. Artificial Neural Networks in Proteomic Research

An important issue in proteomics is finding algorithms allowing unambiguous protein identification based on the searching of a sequence database using mass spectrometry data. Peptides of various molecular mass from enzymatic digestion of a protein are compared to the experimental data in the so-called peptide mass fingerprinting approach [34]. In another case, MS/MS data of one or more peptides are used to confirm the protein identification in the so-called MS/MS ions search approach. The experimental data are compared with the calculated peptide mass or fragment ion mass values in a sequence database and then corresponding mass values are counted or scored in a way that the peptide or protein to be identified matches best the data from the database [34].

One of the most widely used program in proteomics for identification of proteins is a correlation algorithm Sequest. It was developed in 1994 by Yates and co-workers [35-37]. The basis of the Sequest is the assumption that amino acid sequences of peptides can be inferred from tandem mass spectra. The Sequest algorithm automates inferring process by enumerating candidates from the database that match the observed peptide's mass. The sequences are quickly checked against the spectrum by a preliminary scoring algorithm and first nonmatches are removed. A more extensive cross-correlation scoring algorithm is switched then and evaluates the sequence-derived theoretical spectra. It compares them against the observed spectrum, and the sequences are ranked on the basis of such scoring.

Output information in the Sequest program is generated for peptides noted in the given database for which theoretical spectra match well the given experimental spectrum [35-37]. The generated collection of the statistics helps to classify each match. The difference between normalized cross-correlation functions for the first and second ranked results (C_n) is used to indicate a correctly selected peptide sequence. The cross-correlation score between the observed peptide fragment mass spectrum and the theoretically predicted one (X_{corr}), the preliminary score based on the number of ions in the MS/MS spectrum that match the experimental data (S_p), the rank of the particular match during the preliminary scoring (RS_p) and the ions value (I) describing how many of the detected (observed) ions match the theoretical ions for the peptide are then listed. The final Sequest analysis is normally followed by additional manual interpretation of the MS/MS spectra.

For additional evaluating of the Sequest database search results Anderson *et al.* [38] created learning algorithm called the support vector machine (SVM). SVM was designed to distinguish between the correctly and incorrectly identified peptides based on recognition of subtle patterns in a complex data set. Using appropriate training set for MS/MS data, the SVM analysis allowed a better match between these MS/MS data and the theoretical peptide sequences. Manual examination of spectra of peptides with low SVM-calculated scores was still recommended to identify the noisy or poorly fragmenting spectra that might compromise peptide identification. To circumvent or reduce manual interpretation

Table 1. Exemplary data describing peptides used during ANN analysis [39].

Amino acid sequence	pI	dz/dpH	H	MW	CH	X _{corr}	C _n	S _p	RS _p	I	"Good" / "Bad" 1/0
-.MAKDLLPKQAANESLK.D	8.2	-0.2353	43.10	1855.19	2	1.73	0.30	124.3	19	0.34	0
-.MASNAARVVATAKDFDK.V	8.3	-0.1936	40.30	1796.04	2	1.92	0.69	137.4	17	0.31	1
-.MDFYTTDINKNVPLFSK.G	5.7	-0.2192	21.00	2133.45	2	1.76	0.70	301.0	2	0.35	0
-.MEQINSNSRK.K	8.5	-0.1378	44.20	1207.34	2	2.26	0.33	994.4	2	0.67	1
-.MFNCLTKLVILVCLKYVAK.A	10.3	-1.5358	-34.90	2314.89	2	1.56	0.71	92.7	25	0.19	1
-.MGSISRLLK.K	11.0	-0.3808	1.80	1168.44	2	2.00	0.24	193.2	23	0.44	1
-.MSDDDYMNSDDDNDAEKR.Y	3.7	-4.3706	104.60	2137.12	2	1.66	0.66	144.5	21	0.26	1
R.GNPTVEVELTTEK.G	4.2	-1.9601	52.80	1417.54	1	2.49	0.53	180.6	1	0.38	0
K.AADALLKVNQIGTLESSEIK.A	6.11	-0.1463	23.70	2085.43	3	3.82	0.40	959.4	1	0.39	1
-.MYPVDAVLTK.I	5.6	-0.1665	1.40	1251.48	2	1.50	0.15	112.3	88	0.35	1
-.TQFTDIDKLAIVSTIR.I	5.6	-0.2653	14.00	1708.94	3	3.58	0.54	549.0	1	0.50	1

in that aspect, artificial neural network (ANN) analysis has been also recently proposed [39].

The ANN analysis is a method of data analysis which is supposed to reflect the work of human brain [40,41]. In chemistry and related fields of research the interest in neural network computing has been noted since 1986 [42-58]. ANN is a compact group of connected, ordered layers, which are able to process information. There are three kinds of layers in ANN: input layer, one or more hidden layers and output layer. Elements of the network are the elementary units called artificial neurons. These elements are connected with each other with a different connection strength. The connections are called synaptic weights. In weights the whole information on the network is encoded. Those weights are in fact the numbers, which determine the strength of the stimuli coming to the neurons. The most important feature of ANN, which determines the specificity of that computational method, is the process of learning. Learning of the networks is realized by the changes of the values for all synaptic weights with the use of a specific algorithm. The most commonly used learning algorithm is the so-called back-propagation training algorithm. In the process of learning with that algorithm the network uses the error between the current and the desirable output to improve the values of synaptic weights [39-41].

To prove the ability of ANN to reduce manual interpretation of MS/MS spectra a tryptic digest of proteins from yeast extract was analyzed by LC-ESI-MS/MS [39].

Each individual peptide was characterized by several features (Table 1): the observed data (peptide molecular mass, MW, and peptide charge, CH), the Sequest program-calculated statistics data (X_{corr}, C_n, S_p, RS_p and I) and the parameters calculated for individual peptides based on their structural formulas: isoelectric point value (pI), the change in protein charge with pH at the pI (dz/dpH), and hydrophobicity (H).

It was tested whether the trained and validated ANNs exhibit high enough sensitivity and specificity regarding an accurate, high-throughput assignment of the Sequest data in accordance with their manual interpretation. It has been demonstrated that ANNs were capable to process efficiently large sets of data generated during the analysis of complex mixtures of proteins. The ANN constructed in that study predicted in a reliable manner whether the MS/MS spectrum for considered peptide was "good" or "bad" thus replacing the need of manual interpretation of huge amounts of MS/MS spectra typically considered in proteomics.

2.2. Predictions of HPLC Peptides' Retention

Chromatographic retention time can be considered as a chemical structure dependent parameter, which is constant for a given separation conditions (mobile phase composition, stationary phase, temperature, pH). Therefore, prediction of the retention time for a given peptide structure combined with MS/MS data could be helpful for improving the confidence of peptide identifications and increasing the

number of correct identifications. There are observed attempts to use information from liquid chromatography during proteomic analysis [59,60]. For that case the amino acid composition of peptides was taken into account. A number of reports have already been published [61-68] in which chromatographic behavior of peptides in reversed-phase liquid chromatography was considered. First, in the paper by Meek [61] the derivation of specific values ("retention coefficients") that represent the contribution to retention of each of the common amino acids and end groups was demonstrated. "Retention coefficients" were derived directly from HPLC data for all amino acids and end groups. Consequently, the retention time of a peptide were predicted from the sum of "retention coefficients" for each amino acid and end group. Similar strategy, but with different numbers of "retention coefficients" was presented by Browne *et al.* [62], Casal *et al.* [63], Guo *et al.* [64,65]. Mant *et al.* [66] considered also the polypeptide chain length as the

Those all above described approaches were based on simple, additive, amino-acid-composition-of-peptide-based relationships. Another approach for the prediction of HPLC retention times of peptides was proposed lately [69] and was based on quantitative structure-retention relationships (QSRR) [42,70]. QSRR are statistically derived relationships between the chromatographic parameters and descriptors characterizing molecular structure of analytes, here peptides.

QSRR was derived allowing prediction of reversed-phase HPLC retention of peptides [69]. To quantitatively characterize the structure of a peptide, and to predict its gradient retention time the following descriptors were employed: logarithm of the sum of retention times of the amino acids composing the peptide, $\log Sum_{AA}$, logarithm of Van der Waals volume of the peptide, $\log VDW_{Vol}$, and logarithm of its calculated *n*-octanol-water partition coefficient, $\log P$:

$$t_R = 8.02 (\pm 2.04) + 14.86 (\pm 0.93) \log Sum_{AA} - 5.77 (\pm 1.16) \log VDW_{Vol} +$$

$$p = 1 \times 10^{-4} \quad p = 6 \times 10^{-29} \quad p = 3 \times 10^{-6}$$

$$+ 0.28 (\pm 0.06) \log P \quad \text{(Eq. 1)}$$

$$p = 3 \times 10^{-6}$$

$$n = 101; R = 0.963; F = 411; s = 0.97; p < 5 \times 10^{-55}$$

additional descriptor along with the contribution of amino acids into the retention of peptides. Additionally, the influence of different amino acid sequence on peptide retention was studied by Houghten and DeGraw [67]. On the other hand, Zhou *et al.* [68] observed also the presence of a preferred binding domain in an amphipathic α -helical peptide as producing greater retention than might be predicted based on amino acid composition.

In the case of proteomic research, Palmblad *et al.* [59,60] reported prediction of retention times for tryptic peptides. Those predictions were based again on the idea of "retention coefficients". The applied algorithm was tested within the use of tryptic digests of well characterized proteins. Its accuracy was established on the basis of the differences between predicted and experimental retention of peptides identified by mass spectrometry. The most important is that the accuracy of predictions was promising to distinguish between true and false protein matches.

The approach based on ANNs has been proposed also for the prediction of peptide elution times by Petritis *et al.* [41]. Predictive capability of ANN was tested by using large sets of confidently identified peptides of proteomes of two microorganisms. The predicted retention time were demonstrated as useful tool to increase the confidence of peptide identifications. The development of the initial ANN model was based again on the assumption that peptide elution times depends just on amino acid compositions.

where: *n* is the number of peptides used in the studies, *R* is multiple correlation coefficient, *F* is the value of the *F*-test of significance, *s* is standard error of estimate, and *p* is the significance level of the equation or the individual terms of the equation.

The first descriptor, $\log Sum_{AA}$, was based on a set of empirical data for 20 natural amino acids. The next two descriptors, $\log VDW_{Vol}$ and $\log P$ were calculated from a structural formula with the use of molecular modeling methods. The predicted gradient retention times were in good agreement with the experimental data, determined for a structurally diversified series of 101 peptides (Fig. 3).

2.3. Data Processing in pI-Based Separation Methods

Today, the most widely used strategy for analyzing complex protein mixtures is two-dimensional gel electrophoresis [12-19,71-74]. Fractionation of proteins in the first dimension is based on the isoelectric point value of the protein and in the second dimension - on the molecular mass value of the protein. Isoelectric focusing (IEF) is an electrophoretic method that separates proteins and peptides according to their isoelectric points (pI). Proteins and peptides are amphoteric molecules and they carry either positive, negative, or zero net charge, depending on the surrounding pH. The net charge of a protein is the sum of all the negative and positive charges of its amino acid side chains and amino- and carboxyl-termini. The isoelectric point is the specific pH at which the net charge of the protein

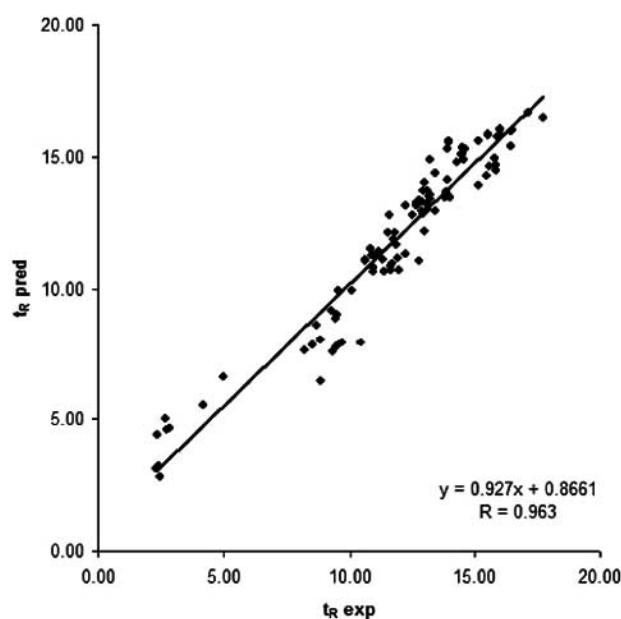


Fig. (3). Correlation between the calculated by QSRR and the experimental retention times for a set of 101 peptides studied [69].

or the peptide is zero. The pI value is one of the most significant characteristics of proteins and peptides. Being a very valuable physicochemical structural descriptor in proteomics it can be used as the additional constraint during the identification process of proteins [75,76].

Preparative isoelectric focusing analogously to the first step in 2D-PAGE, can be performed not only on an immobilized pH gradient (IPG) strip or in a tube gel, but also in solution. pH gradient in this method is created with the certain immobiline mixture immobilized into polyacrylamide gel or with the use of ampholytes in the certain pH range [75]. When voltage is applied across the focusing cell stable pH gradient is generated. It makes the possibility to separate

the proteins or peptides according to their isoelectric point. However, in the case of proteins, some of them demonstrate a tendency to aggregate and precipitate during focusing [4]. pI-based separations for proteome analysis have been employed several times for the fractionation of proteins. In that case Rotofor [77], multicompartement electrolyzer [78], microscale solution isoelectrofocusing device (μ sol-IEF) [79], off-gel isoelectric focusing [80] or multicompartement electrolyzer with polyacrylamide gel beads [81] were used. Another pI-based method – chromatofocusing – has been used as well [82]. However, again for fractionation of proteins. On the other hand, the physical and chemical properties of peptides, derived from enzymatic digestion, are less diverse than those of the original proteins. Most peptides resulting from enzymatic digestion are readily soluble in water or water/organic mixtures. Despite of a wide use of isoelectric focusing-based methods for protein fractionation, reports on the separation of peptides are rather limited. They include the utility of either capillary isoelectric focusing (cIEF) [83,84] or in-solution isoelectric focusing (sIEF) [75,76,85]. New HPLC strategies, e.g., gradient HPLC [86] could have also a potential value for separation of peptides. Also in that set of proteomic strategies two-dimensional separation system comprising for example sIEF and RPLC can be considered [75,76].

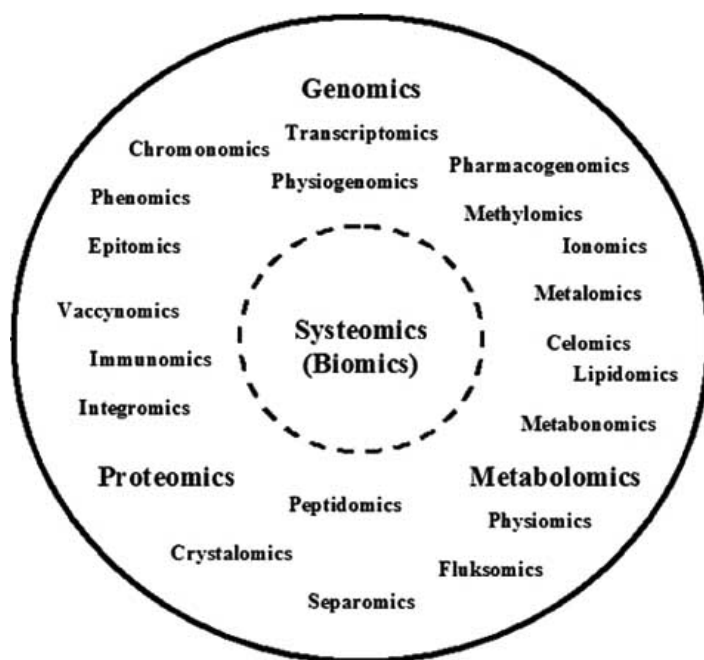
sIEF method was proposed as the useful alternative fractionation method of protein digests [75]. It was considered as the first dimension for 2-D proteomics separations. It was demonstrated for the fractionation of BSA digest and mixture of five proteins digest. Additionally, sIEF in combination with ZipTip fractionation was used as an easy-to-use and analytically efficient extension of sIEF used previously alone. ZipTip pipette tips for sample preparation (Millipore, Billerica, MA) contains 10 μ L of C18 resin fixed at their tip. They allow to desalt and concentrate peptides during step-fractionation of complex peptide mixture. sIEF in combination with ZipTip fractionation were proved to be useful for five proteins mixture with the

Table 2. Results of identification of tryptic digest of five proteins mixed in the same concentrations based on MS/MS spectra and Mascot database searching [75].

Separation system	Protein identified	Peptides identified (Sequence coverage)
MALDI-TOF/TOF-MS	Ovalbumin	2 (8%)
	-Lactoglobulin	2 (11%)
	-Casein	2 (6%)
	Myoglobin	1 (13%)
SIEF +MALDI-TOF/TOF-MS	Ovalbumin	4 (14%)
	-Lactoglobulin	3 (26%)
	-Casein	2 (12%)
	Myoglobin	9 (63%)
SIEF +ZipTips +MALDI-TOF/TOF-MS	BSA	2 (3%)
	Ovalbumin	4 (13%)
	-Lactoglobulin	6 (33%)
	-Casein	2 (19%)
	Myoglobin	13 (75%)

Table 3. Number of proteins identified with the use of different constraints [76].

No.	Constraint used	No. of identified proteins
1.	Tryptic peptide with: $X_{\text{corr}} > 2.0$ (singly charged), $X_{\text{corr}} > 1.5$ (doubly charged), $X_{\text{corr}} > 3.3$ (triple charged), $C_n > 0.08$	851
2.	As 1. and after manual interpretation of MS/MS spectra	542
3.	As 2. and after using of sIEF-based pI constraint	187
4.	As 2. but 1. using without any enzyme fixed	188
5.	As 3. but 1. using without any enzyme fixed	126

**Fig. (4).** Systemomics (biomics) and other new biological-chemical research strategies.

concentration of proteins in the same order of magnitude (Table 2). That strategy marked also in the case of five proteins mixtures differing in the range of the concentration in four orders of magnitude. The sIEF method can be also considered as the additional source of database matching constraint used in the evaluation process of proteomics data. Using pI values calculated for identified peptides validity of the database searching, could be additionally checked on the basis of the occurrence of peptides in the appropriate fractions in the sIEF device.

Another example of the utility of in-solution isoelectric focusing (sIEF) combined with reversed-phase liquid chromatography (RPLC) was proposed [76]. Fractionation of *Saccharomyces cerevisiae* protein digest was performed with that two-dimensional separation approach. sIEF fractionation combined with reversed-phase nanocapillary liquid chromatography enabled the evaluation of yeast proteome. Based on the MS/MS results obtained with ESI-MS/MS instrument, 851 proteins were identified. However, after

bioinformatic analysis of the 2D-separation and mass spectrometry data, reduction in number of proteins to 126 was obtained (Table 3). The whole approach was proposed also as the additional source of database matching constraint used in the evaluation process of proteomics data. Using pI values calculated for identified peptides the validity of the database searching was again checked on the basis of the occurrence of peptides in the appropriate sIEF fractions.

CONCLUSIONS

Taking into account that proteomics seems to be one of the main area of the future life sciences, anyone can not forget about the potential of genomics, transcriptomics, metabolomics and other modern research strategies. Hence, the whole knowledge about the life has a chance to be soon one large area of research called as biomics or systemomics (Fig. 4).

The success of that area will depend strongly on the ability to design and use of new analytical and bioinformatic strategies, which will enable to study in the fast, efficient and accurate manner the huge amount and type of the activity of thousands of biomolecules existing in organisms.

ACKNOWLEDGEMENTS

The author is grateful to the Foundation for Polish Science and to the Polish State Committee for Scientific Research Project 2 P05F 012 27 for the support during the course of this research.

REFERENCES

- [1] International Human Genome Sequencing Consortium *Nature*, **2001**, *409*, 860.
- [2] The Celera Genomics Sequencing Team *Science*, **2001**, *291*, 1304.
- [3] Wilkins, M.R.; Sanchez, J.C.; Gooley, A.A.; Appel, R.D.; Humphrey-Smith, I.; Hochstrasser, D.F.; Williams, K.L. *Biotechnol. Genet. Eng. Rev.*, **1996**, *13*, 19.
- [4] Liebler, D.C. *Introduction to Proteomics*, Humana Press: Totowa, NJ, **2002**.
- [5] Pandey, A.; Mann, M. *Nature*, **2000**, *405*, 837.
- [6] Wehr, T. *LCGC North America*, **2002**, *20*, 954.
- [7] Wehr, T. *LCGC North America*, **2001**, *19*, 702.
- [8] Rabilloud, T. *Proteomics*, **2002**, *2*, 3.
- [9] Righetti, P.G.; Stoyanov, A.V.; Zhukov, M.Y. *The Proteome Revisited. Theory and Practice of All Relevant Electrophoretic Steps*, Elsevier: Amsterdam, **2001**.
- [10] Beranova-Giorgianni, S. *Trends Anal. Chem.*, **2003**, *22*, 273.
- [11] Görg, A.; Obermaier, C.; Boguth, G.; Harder, A.; Scheibe, B.; Wildgruber, R.; Weiss, W. *Electrophoresis*, **2000**, *21*, 1037.
- [12] Caron, M.; Imam-Sghiouar, N.; Poirier, F.; Le Caër, J.-P.; Labas, V. Joubert-Caron, R. *J. Chromatogr. B*, **2002**, *771*, 197.
- [13] Nakanishi, T.; Koyama, R.; Ikeda, T.; Shimizu, A. *J. Chromatogr. B*, **2002**, *776*, 89.
- [14] Noël-Georis, I.; Bernard, A.; Falmagne, P.; Wattiez, R. *J. Chromatogr. B*, **2002**, *771*, 221.
- [15] Sickmann, A.; Dormeyer, W.; Wortelkamp, S.; Woitalla, D.; Kuhn, W.; Meyer, H.E. *J. Chromatogr. B*, **2002**, *771*, 167.
- [16] Hubbard, M.J.; Kon, J.C. *J. Chromatogr. B*, **2002**, *771*, 211.
- [17] Adkins, J.N.; Varnum, S.M.; Auberry, K.J.; Moore, R.J.; Angell, N.H.; Smith, R.D.; Springer, D.I.; Pounds, J.G. *Mol. Cell Proteomics*, **2002**, *1*, 947.
- [18] Anderson, N.L.; Anderson, N.G. *Mol. Cell Proteomics*, **2002**, *1*, 845.
- [19] Tichá, M.; Pacáková, V.; Stulik, K. *J. Chromatogr. B*, **2002**, *771*, 343.
- [20] Link, A.J.; Eng, J.; Schieltz, D.M.; Carmack, E.; Mize, G.J.; Morris, D.R.; Garvik, B.M.; Yates III, J.R. *Nat. Biotechnol.*, **1999**, *17*, 676.
- [21] Washburn, M.P.; Wolters, D.; Yates, III, J.R. *Nat. Biotechnol.*, **2001**, *19*, 242.
- [22] Peng, J.; Elias, J.E.; Thoreen, C.C.; Licklider, L.J.; Gygi, S.P. *J. Proteome Res.*, **2003**, *2*, 43.
- [23] Wagner, K.; Miliotis, T.; Marko-Varga, G.; Bischoff, R.; Unger, K.K. *Anal. Chem.*, **2002**, *74*, 809.
- [24] Opiteck, G.J.; Lewis, K.C.; Jorgenson, J.W.; Anderegg, R.J. *Anal. Chem.*, **1997**, *69*, 1518.
- [25] Davis, M.T.; Beierle, J.; Bures, E.T.; McGinley, M.D.; Mort, J.; Robinson, J.H.; Spahr, C.S.; Yu, W.; Luthy, R.; Patterson, S.D. *J. Chromatogr. B*, **2001**, *752*, 281.
- [26] Opiteck, G.J.; Jorgenson, J.W.; Anderegg, R.J. *Anal. Chem.*, **1997**, *69*, 2283.
- [27] Lewis, K.C.; Opiteck, G.J.; Jorgenson, J.W.; Sheeley, D.M. *J. Am. Soc. Mass Spectrom.*, **1997**, *8*, 495.
- [28] Licklider, L.J.; Thoreen, C.C.; Peng, J.; Gygi, S.P. *Anal. Chem.*, **2002**, *74*, 3076.
- [29] Martin, S.E.; Shabanowitz, J.; Hunt, D.F.; Marto, J.A. *Anal. Chem.*, **2000**, *72*, 4266.
- [30] Gygi, S.P.; Rist, B.; Gerber, S.A.; Turecek, F.; Gelb, M.H.; Aebersold, R. *Nat. Biotechnol.*, **1999**, *17*, 994.
- [31] Gygi, S.P.; Rist, B.; Griffin, T.J.; Eng, J.; Aebersold, R. *J. Proteome Res.*, **2002**, *1*, 47.
- [32] Kellner, R.; Lottspeich, F.; Meyer, H.E. *Microcharacterization of Proteins*, Wiley-VCH: Weinheim, **1999**.
- [33] Yates III, J.R. *J. Mass Spectrom.*, **1998**, *33*, 1.
- [34] Perkins, D.N.; Pappin, D.J.C.; Creasy, D.M.; Cottrell, J.S. *Electrophoresis*, **1999**, *20*, 3551.
- [35] Eng, J.K.; McCormack, A.L.; Yates III, J.R. *J. Am. Soc. Mass Spectrom.*, **1994**, *5*, 976.
- [36] Yates III, J.R.; Eng, J.K.; McCormack, A.L.; Schieltz, D. *Anal. Chem.*, **1995**, *67*, 1426.
- [37] Tabb, D.L.; McDonald, W.H.; Yates III, J.R. *J. Proteome Res.*, **2002**, *1*, 2.
- [38] Anderson, D.C.; Li, W.; Payan, D.G.; Noble, W.S. *J. Proteome Res.*, **2003**, *2*, 137.
- [39] Baczek, T.; Bucinski, A.; Ivanov, A.R.; Kaliszán, R. *Anal. Chem.*, **2004**, *76*, 1726.
- [40] Zupan, J.; Gasteiger, J. *Anal. Chim. Acta*, **1991**, *248*, 1.
- [41] Zupan, J.; Gasteiger, J. *Neural Networks for Chemists. An Introduction*, VCH: Weinheim, **1993**.
- [42] Petritis, K.; Kangas, L.J.; Ferguson, P.L.; Anderson, G.A.; Pasatolic, L.; Lipton, M.S.; Auberry, K.J.; Strittmatter, E.F.; Shen, Y.; Zhao, R.; Smith, R.D. *Anal. Chem.*, **2003**, *75*, 1039.
- [43] Kaliszán, R. *Structure and Retention in Chromatography, A Chemometric Approach*, Harwood Academic: Amsterdam, **1997**.
- [44] Schneider, G.; Wrede, P. *Prog. Biophys. Mol. Biol.*, **1998**, *70*, 175.
- [45] Isu, Y.; Nagashima, U.; Hosoya, H.; Aoyama, T. *J. Chem. Software*, **1994**, *2*, 76.
- [46] Andrea, T.A.; Kalayeh, H. *J. Med. Chem.*, **1991**, *34*, 2824.
- [47] So, S.-S.; Richards, W.G. *J. Med. Chem.*, **1992**, *35*, 3201.
- [48] Ajay, A. *J. Med. Chem.*, **1993**, *36*, 3565.
- [49] Brickley, M.R.; Shepherd, J.P.; Armstrong, R.A. *J. Dent.*, **1998**, *26*, 305.
- [50] Snow, P.B.; Rodvold, D.M.; Brandt, J.M. *Urology*, **1999**, *54*, 787.
- [51] Wei, J.T.; Tewari, A. *Urology*, **1999**, *54*, 945.
- [52] Krongrad, A.; Lai, S. *Urology*, **1999**, *54*, 949.
- [53] Lee, C.W.; Park, J.-A. *Inform. Manage.*, **2001**, *38*, 231.
- [54] Jalali-Heravi, M.; Parastar, F. *J. Chromatogr. A*, **2000**, *903*, 145.
- [55] Loukas, Y.L. *J. Chromatogr. A*, **2000**, *904*, 119.
- [56] Jimenez, O.; Marina, M.L. *J. Chromatogr. A*, **1997**, *780*, 149.
- [57] Bucinski, A.; Baczek, T. *Pol. J. Food Nutr. Sci.*, **2002**, *11*, 47.
- [58] Kaliszán, R.; Baczek, T.; Bucinski, A.; Buszewski, B.; Sztupecka, M. *J. Sep. Sci.*, **2003**, *26*, 271.
- [59] Palmblad, M.; Ramström, M.; Markides, K.E.; Håkansson, P.; Bergquist, J. *Anal. Chem.*, **2002**, *74*, 5826.
- [60] Palmblad, M.; Ramström, M.; Bailey, C.G.; McCutchen-Maloney, S.L.; Bergquist, J.; Zeller, L.C. *J. Chromatogr. B*, **2004**, *803*, 131.
- [61] Meek, J.L. *Proc. Natl. Acad. Sci. USA*, **1980**, *77*, 1632.
- [62] Browne, C.A.; Bennett, H.P.J.; Solomon, S. *Anal. Biochem.*, **1982**, *124*, 201.
- [63] Casal, V.; Martin-Alvarez, P.J.; Herraiz, T. *Anal. Chim. Acta*, **1996**, *326*, 77.
- [64] Guo, D.; Mant, C.T.; Taneja, A.K.; Parker, J.M.R.; Hodges, R.S. *J. Chromatogr.*, **1986**, *359*, 499.
- [65] Guo, D.; Mant, C.T.; Taneja, A.K.; Hodges, R.S. *J. Chromatogr.*, **1986**, *359*, 519.
- [66] Mant, C.T.; Zhou, N.E.; Hodges, R.S. *J. Chromatogr.*, **1989**, *476*, 363.
- [67] Houghton, R.A.; DeGraw, S.T. *J. Chromatogr.*, **1987**, *386*, 223.
- [68] Zhou, N.E.; Mant, C.T.; Hodges, R.S. *Pept. Res.*, **1990**, *3*, 8.
- [69] Kaliszán, R.; Baczek, T.; Cimochovska, A.; Juszczak, P.; Wisniewska, K.; Grzonka, Z. Prediction of HPLC retention of peptides with the use of quantitative structure-retention relationships. *Proteomics*, in press.
- [70] Kaliszán, R. *Quantitative Structure-Chromatographic Retention Relationships*, Wiley: New York, **1987**.
- [71] Joubert, R.; Strub, J.-M.; Zugmeyer, S.; Kobi, D.; Carte, N.; van Dorsselaer, A.; Boucherie, H.; Jaquet-Gutfreund, L. *Electrophoresis*, **2001**, *22*, 2969.
- [72] Perrot, M.; Saggiocco, F.; Mini, T.; Monribot, C.; Schneider, U.; Shevchenko, A.; Mann, M.; Jenö, P.; Boucherie, H. *Electrophoresis*, **1999**, *20*, 2280.
- [73] Poutanen, M.; Salusjarvi, L.; Ruohonen, L.; Penttilä, M.; Kalkkinen, N. *Rapid Commun. Mass Spectrom.*, **2001**, *15*, 1685.

- [74] Salusjarvi, L.; Poutanen, M.; Pitkanen, J.-P.; Koivistoinen, H.; Aristidou, A.; Kalkkinen, N.; Ruohonen, L.; Penttila, M. *Yeast*, **2003**, *20*, 295.
- [75] Baczek, T. *J. Pharm. Biomed. Anal.*, **2004**, *34*, 851.
- [76] Baczek, T. *J. Pharm. Biomed. Anal.*, **2004**, *35*, 895.
- [77] Lubman, D.M.; Kachman, M.T.; Wang, H.; Gong, S.; Yan, F.; Hamler, R.L.; O'Neil, K.A.; Zhu, K.; Buchanan, N.S.; Barder, T.J. *J. Chromatogr. B*, **2002**, *782*, 183.
- [78] Herbert, B.; Righetti, P.G. *Electrophoresis*, **2000**, *21*, 3639.
- [79] Zuo, X.; Hembach, P.; Echan, L.; Speicher, D.W. *J. Chromatogr.*, **2002**, *782*, 253.
- [80] Ros, A.; Faupel, M.; Mees, H.; Van Oostrum, J.; Ferrigno, R.; Reymond, F.; Michel, P.; Rossier, J.S.; Girault, H.H. *Proteomics*, **2002**, *2*, 151.
- [81] Cretich, M.; Pirri, G.; Carrea, G.; Chiari, M. *Electrophoresis*, **2003**, *24*, 577.
- [82] Kang, X.; Frey, D.D. *Anal. Chem.*, **2002**, *74*, 1038.
- [83] Shen, Y.; Xiang, F.; Veenstra, T.D.; Fung, E.N.; Smith, R.D. *Anal. Chem.*, **1999**, *71*, 5348.
- [84] Shen, Y.; Berger, S.J.; Anderson, G.A.; Smith, R.D. *Anal. Chem.*, **2000**, *72*, 2154.
- [85] Tan, A.; Pashkova, A.; Zang, L.; Foret, F.; Karger, B.L. *Electrophoresis*, **2002**, *23*, 3599.
- [86] Kaliszan, R.; Wiczling, P.; Markuszewski, M.J. *Anal. Chem.*, **2004**, *76*, 749.

Received: 15 September, 2004

Accepted: 19 October, 2004