Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography B, 824 (2005) 189–200
JOURNAL OF CHROMATOGRAPHY B
www.elsevier.com/locate/chromb

Optimization of protein identification from digests as analyzed by capillary isoelectric focusing-mass spectrometry

Henricus F. Storms *, Robert van der Heijden, Ubbo R. Tjaden, Jan van der Greef

Leiden University, Amsterdam Center for Drug Research, Division of Analytical Biosciences, P.O. Box 9502, 2300 RA Leiden, The Netherlands

Received 29 April 2005; accepted 3 July 2005
Available online 3 August 2005

Abstract

Capillary isoelectric focusing (CIEF) is a high-resolution separation technique for the analysis of peptides and protein digests. When coupled to ion trap-mass spectrometry (CIEF-MS) the unique separation mechanism is combined with a highly efficient detection system. In an earlier report, we described aspects of separation and interfacing in connection to the analysis of a digest of set of standard proteins. Now, we report on different aspects of the process of protein identification. Sequest software parameters were optimized by using a standard protein digest. These settings were used for the analysis of periplasmic proteins from *Escherichia coli*. Since in CIEF peptides are focused according to their p*I* values, the mobilization time of a particular peptide is dependent on its p*I* value. Based on this relation, the identification of some peptides was facilitated. Furthermore, the Sequest settings that were used could be evaluated. In total, 159 proteins were identified in a single run.

Keywords: Capillary isoelectric focusing; Mass spectrometry; Protein analysis; Data analysis; Sequest

1. Introduction

For the understanding of cellular processes, the study of the proteomes has taken an increasingly important role. The use of the so-called shotgun approach has become increasingly popular in the last decade. In this approach, the protein extract is digested at the beginning of the work flow, and the resulting peptides are then separated by liquid chromatography (LC) or capillary electrophoresis (CE) [1,2]. MS/MS analysis is used for the identification of the individual peptides and subsequently the corresponding proteins.

Earlier, we have reported on the use of capillary isoelectric focusing as a separation technique for the shotgun approach. CIEF-MS of complex peptide mixtures was performed without the use of carrier ampholytes, and with the use of low amounts of carrier ampholytes as mere spacers [3]. CIEF is a high-resolution separation technique that can be applied for amphoteric compounds, such as proteins or peptides. These are separated according to their p*I*-values, in a pH gradient formed under the influence of an electric field [4].

Though normally carrier ampholytes are added to the sample in order to establish a linear pH gradient, in the case of peptides this is not needed for focusing to occur; this process is called autofocusing [5–7]. It should be noted though that the peptides lack the high buffering capacity of the carrier ampholytes, but compared to carrier ampholytes the peptides are fully compatible with mass spectrometry.

The described approach, despite a rather limited resolution, allowed the identification of the eight components of a mixture of standard proteins. The addition of a relatively low concentration of carrier ampholytes (0.2%) resulted in increased separation efficiency, although already ion suppression was observed. Furthermore, the use of higher sample concentrations also resulted in improved separation efficiency.

* Corresponding author.
E-mail address: harriestorms@zonnet.nl (H.F. Storms).

1570-0232/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2005.07.020

In continuation of that research we have studied several aspects of data acquisition, data processing and protein identification aiming to obtain clean data, in which the number of correctly identified proteins is maximized while the need for manual checking for false positives is minimized.

In this study, a 10-protein mix was analyzed using CIEF followed by linear ion trap mass spectrometry, which has both superior sensitivity and a higher scanning speed as compared to the conventional 3D ion trap [8]. For identification, Sequest software was used. Sequest software compares the acquired MS/MS spectra with theoretical MS/MS spectra of (tryptic) peptides; these theoretical spectra are based on their sequences which are taken up in a database.

Several earlier studies have made it clear that the parameters and settings used in this software are very important for distinguishing true hits from false positives. Various improvements have been suggested, like the use of discriminant function analysis for optimization of Sequest parameters [9,10] and the use of a machine-learning algorithm [11].

In general, what these studies have made clear is that several parameters contribute significantly to the discrimination of correct and incorrect peptide assignments. The most important one is the cross correlation value (Xcor), which is a correlation coefficient for the match. A minimum threshold for this parameter is commonly being used. Link et al. [14] used 1.5 for singly charged peptides, and 2.0 for multiply charged peptides. Some later reports have used higher thresholds for the Xcor, and on top of that also used a threshold of 0.1 for the ΔC_N delta correlation value (ΔC_n), which is a measure for the difference in the correlation value between a particular hit and the first following hit [15,16].

Other parameters that can be taken into account are the preliminary score based on the number of ions in the MS/MS spectrum that match with the experimental data (Sp), the ranking of the peptide match in the resulting preliminary scoring list (RSp) and finally the coverage of *y*- and *b*-ions. Anderson et al. [11] have shown that these parameters can also contribute in the discrimination of correct and incorrect peptide assignments. Furthermore, they showed that all parameters, including the Xcor, are dependent on the type of mass spectrometer used, resulting from differences in accuracy and amount of noise in the MS/MS spectra. Taking this into account, it is expected that protein identification will benefit from optimization of all mentioned parameters.

When CIEF is performed with protein digests, ordering according to the p*I* values of the peptides is expected. This means that the mobilization times of the peptides are dependent on their p*I* values. Thus, the use of CIEF provides an extra criterion for the identification of the individual peptides. In this study, it is investigated whether this could be used for data analysis.

Finally, the optimized procedures were applied to the analysis of a biological sample, the periplasmatic proteins from *Escherichia coli* cells.

2. Materials and methods

2.1. Chemicals

DTT and iodoacetamide were purchased from Sigma–Aldrich Chemie (Steinheim, Germany). Myoglobin from horse heart, bovine serum albumin, cytochrome *C* from horse heart, human insulin, human serum albumin, carbonic anhydrase I from human erythrocytes, lactoglobulin B from bovine milk, bovine ribonuclease A, lysozyme C from chicken egg and ovalbumin from chicken egg were all purchased from Sigma–Aldrich Chemie (Steinheim, Germany).

2.2. Standard protein mix

Proteins were digested at a concentration of 1 mg/mL with sequencing grade trypsin (Roche Diagnostics Boehringer Mannheim B.V, Mannheim, Germany) according to Matsudaira [12]. The proteins were solved in 50 mM NH_4HCO_3, treated with DTT (2.25 mM) for the reduction of disulfide bonds and were then carboxymethylated with iodoacetamide (5.0 mM) to prevent reoxidation. Digestion was performed by incubating the proteins for 24 h at 37 °C with trypsin (at a 1:30 enzyme:protein ratio).

2.3. Periplasmatic protein extract

E. coli cells (K12 strain) were grown in LB medium at 37 °C. Periplasmatic proteins were isolated by osmotic shock [13]. Harvested cells (20 mL cell suspension) were centrifuged at 1000 × *g* for 10 min and suspended in 50 mM Tris/HCl, pH 7.5, containing 0.5 M sucrose. This was repeated twice, but the second time the cells were suspended in 50 mM Tris/HCl, pH 7.5, containing 0.5 M sucrose and 1 mM EDTA. After equilibration for 20 min cells were centrifuged again and suspended in 0.20 mL of 50 mM Tris/HCl, pH 7.5, containing 0.5 M sucrose and 1 mM EDTA. Osmotic shock is performed by suspending the cells in 5 mL water at 4 °C. The resulting suspension was centrifuged at 10,000 × *g* and the supernatant was recovered as the periplasmatic fraction. The protein concentration was estimated according to Bradford [14].

By ultrafiltration (Microcon centrifugal filter devices, MWCO 10 kDa, Millipore, Amsterdam, the Netherlands) the protein extract of *E. coli* cells was washed and the buffer was replaced with digestion buffer (50 mM NH_4HCO_3). The lysate was treated with DTT and iodoacetamide, and digested with sequencing grade trypsin (Roche Diagnostics Boehringer Mannheim B.V., Mannheim, Germany) according to Matsudaira [12], as described for the 10-protein mix.

2.4. CIEF

Fused-silica capillaries (75 μm i.d., 375 μm o.d.) were obtained from Bester (Amstelveen, the Netherlands). In order to prevent adsorption at the capillary walls and to reduce the

electroosmotic flow, the capillary walls were coated with a two-layer siloxanediol-polyacrylamide coating as described by Schmalzing et al. [15]. The length of the capillaries was 105 cm.

For CIEF, a programmable PrinCE capillary electrophoresis system was used (Prince Technologies, Emmen, The Netherlands). An acetic acid solution (0.5%) was used as a weak acid (anolyte) as well as a weak base (catholyte) to establish the pH gradient in the capillary as described by Lamoree et al. [16]. First, the capillary was filled for approximately 85% with sample. During focusing, 25 kV was applied over the capillary for 12 min. Mobilization was obtained by applying appropriate pressure (0–80 mbar) over the capillary while maintaining the voltage. For experiments with carrier ampholytes, Pharmalyte pH 3–10 was used (Amersham, Uppsala, Sweden).

2.5. Mass spectrometry

All mass spectrometric measurements were performed using a LTQ linear ion trap mass spectrometer (Thermo Electron, San Jose, USA) equipped with a custom-made electrospray ionization source in a coaxial sheath-flow configuration. The sheath liquid consisted of 0.5% acetic acid in methanol/water (80/20, v/v) and was delivered at a flow rate of 2 μL/min. The electrospray voltage was held at ground potential during focusing, and set at 3.8 kV during mobilization to establish a spray. The voltage applied at the capillary inlet was 25 kV during focusing and 28.8 kV during mobilization, respectively.

Precursor ion scanning from 250 to 1500 *m*/*z* was followed by a zoomscan of the highest peak in the spectrum, again followed by the generation of an MS/MS spectrum. Masses that had been analyzed for more than three times this way were automatically taken up into an exclusion list for 3 min.

2.6. Protein identification

For the identification of peptides, tandem mass spectrometry was performed. Bioworks software, version 3.1, from Thermo Electron (San Jose, USA) was used for automatic sequencing and database search. For all protein searches, one missed cleavage was allowed. For the analysis of the 10-protein mix, a Swiss-Prot based database was made containing all known sequences from human, horse, bovine and chicken proteins, based on the general Swiss-Prot protein sequence database. For the analysis of the *E. coli* extract, a database was made containing all known sequences from *E. coli* K12, based on a general nonredundant protein database.

3. Results and discussion

For the optimization of the identification process, data acquisition (e.g. scanning speed of the mass spectrometer) and data processing (Sequest search parameters) were studied using a mixture of known proteins analyzed by CIEF-MS/MS.

3.1. CIEF-MS/MS of a standard set of proteins

In our previous study on CIEF-MS/MS of protein digests we suggested that the sequence coverage of the proteins would be higher if the scan speed of the ion trap mass spectrometer was higher. We tested this statement by using a fast scanning linear ion trap. The linear ion trap used in this study (LTQ from Thermo Electron) is able to obtain approximately six MS/MS spectrum every 2 s in data dependent mode, compared to one MS/MS spectrum every 2 s for the conventional ion trap (Deca XP Plus from Thermo Electron) we used in our previous study [3].

Fig. 1 displays a typical electropherogram resulting from the analysis of digest of a mixture of 10 standard proteins at a concentration of 0.3 mg of each protein/mL. Carrier ampholytes were added to a concentration of 0.20%. At this concentration of proteins, we found that the viscosity gets relatively high after focusing and high pressures (10–50 mbar) had to be applied at the inlet for mobilization.

3.2. Optimizing search parameters

One of the limitations of using computer software for automatic sequencing and database search is the possibility of false-positive hits. This can be solved by setting strict thresholds for the identifications, but that will also lead to less true positive identifications. Thus, an optimum has to be found.

Using data sets of three sequential runs of the standard protein mixture the Sequest parameters were optimized to decrease the number of false positives and increase the number of true positives. As discussed above, it was chosen to use all five Sequest parameters available: Xcor, ΔC_N, Sp, RSp and the coverage of *y*- and *b*-ions. In our experience, when using all five parameters, it is best to consider a given peptide as a good hit when in the first place the Xcor value matches the criterion that is set and in the second place when of the remaining four parameters at least three match the criteria that are set.

The optimized threshold values are shown in Table 1. When using all five criteria, we found that lower values for the Xcor can be used than when using only Xcor plus ΔC_N. This is illustrated in Fig. 2, which displays the results for one of the three runs. The Xcor values used for the 'five criteria'-method are 1.5, 1.8 and 2.3 for singly charged, doubly charged

Table 1
Optimized Sequest score used in 'five-criteria' method

Xcor (charge)	1.5 (1+), 1.8 (2+), 2.3 (3+)
ΔC_N	0.1
Coverage *b*/*y* ions (%)	45
Sp	550
RSp	4

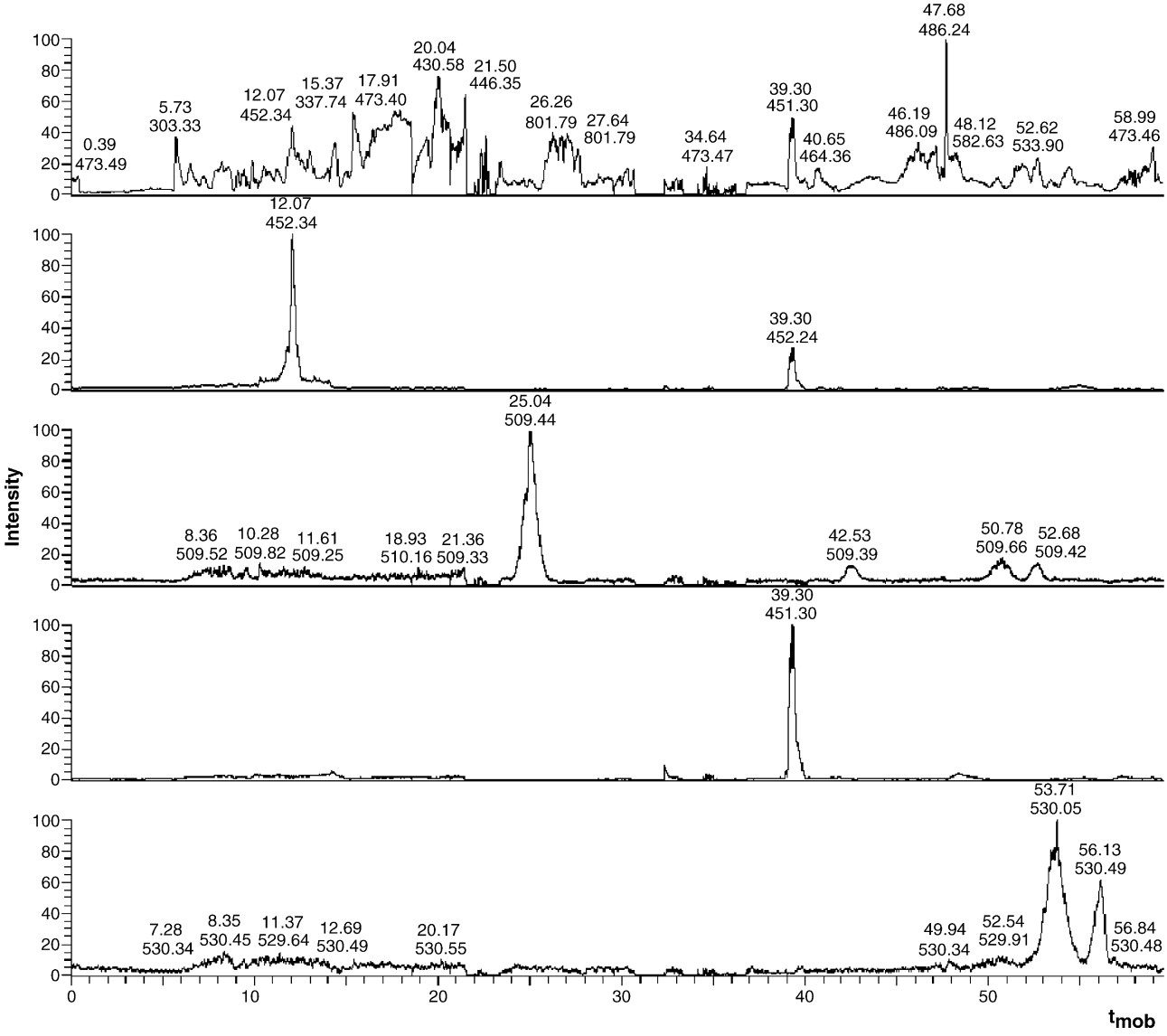

Fig. 1. Electropherogram of a digest of a 10-protein mixture (horse myoglobin, horse cytochrome *C*, human carbonic anhydrase I, bovine albumin, bovine *P*-lactoglobulin B, bovine ribonuclease I, human insulin, human albumin, chicken ovalbumin, chicken lysozyme) at a concentration of 0.3 mg/mL per protein. The samples were focused for 14 min using 25 kV. Mobilization was achieved by applying 50 mbar of pressure until the first peaks were seen. Then ($t = 0$ min), 20 mbar was applied for the first 25 min, and 50 mbar for the remaining time. The electropherogram on top shows the total ion current. Below, four masses have been selected to show their peak shape.

and triply charged peptides respectively (see Table 1). For the 'two criteria' method, the Xcor values were set at 1.8, 2.1 and 2.6 for singly charged, doubly charged and triply charged peptides, respectively, while a threshold of 0.1 for ΔC_N was maintained.

All false positive protein identifications were based on single peptide sequences. Because of their more ambiguous nature it is common to increase the criteria for proteins identified with only a single peptide hit. For those, the Xcor thresholds were increased to 1.8, 2.1 and 2.6 for singly charged, doubly charged and triply charged peptides, respectively, while the remaining thresholds were kept the same. This way, only three incorrectly assigned peptide sequences were left (Table 2).

Table 2 shows a list of all identified peptide sequences, including the incorrectly assigned ones. When a peptide was identified more than once, only the match with the highest Xcor value is displayed, while between brackets the number of times it was identified is given.

One of the proteins that was identified needs special attention. Ovomucoid was identified in all three runs. Though this protein was not added to the sample, we found that this protein is a common impurity of commercial ovalbumin, both being egg white proteins [17]. Therefore, it was not considered to be a false positive. The purity of the ovalbumin as stated by the manufacturer was 98%.

The three spectra of the incorrectly assigned peptides contained very much noise, which lead to the assignment of the

Table 2
Filtered Sequest results for 10-protein digest

Protein	Sequence	z	Xcor	ΔC_N	Sp	Rsp	Coverage
Carbonic anhydrase I	K.VLDALQAIK.T (11)	2	3.44	0.48	836.2	1	15/16
	K.HDTSLKPISVSYNPATAK.E (3)	3	3.71	0.52	683.1	1	27/68
	K.YSAELHVAHWNSAK.Y (2)	2	4.10	0.54	1409.8	1	21/26
	K.HDTSLKPISVSYNPATAK.E (2)	3	3.97	0.66	1953.1	1	32/68
	R.SLLSNVEGDNAVPM∼QHNNRPTQPLK.G (2)	3	4.26	0.47	588.7	1	29/96
	K.GGPFSDSYR.L (3)	1	1.78	0.38	245.8	1	11/16
	K.YSSLAEAASK.A (2)	1	1.61	0.36	405.1	1	11/18
	K.LYPIANGNNQSPVDIK.T (5)	2	4.73	0.48	962.8	1	24/30
	K.ESISVSSEQLAQFR.S (4)	2	2.78	0.40	1134.6	1	18/26
Albumin (bovine)	K.KQTALVELLK.H (4)	2	2.10	0.38	336.0	1	12/18
	R.RHPEYAVSVLLR.L (3)	2	2.35	0.44	676.9	1	16/22
	K.LVNELTEFAK.T (4)	2	2.45	0.43	860.3	1	15/18
	R.FKDLGEEHFK.G (2)	2	2.39	0.38	1226.3	1	16/18
	K.LKECCDKPLLEK.S	2	1.91	0.07	504.0	1	14/22
	K.QTALVELLK.H	1	1.65	0.12	98.4	2	11/16
	K.NYQEAK.D	1	1.52	0.32	182.4	4	8/10
	K.IETMR.E (2)	1	1.60	0.24	191.7	1	6/8
	K.HLVDEPQNLIK.Q	2	2.79	0.56	1024.2	1	16/20
	K.LGEYGFQNALIVR.Y (3)	2	2.96	0.46	1953.0	1	19/24
	K.AEFVEVTK.L (2)	1	2.17	0.36	703.9	2	9/14
	K.DLGEEHFK.G (2)	2	1.84	0.41	859.4	1	12/14
Albumin (human)	K.KVPQVSTPTLVEVSR.N (4)	2	3.33	0.60	540.1	1	17/28
	R.RHPDYSVVLLLR.L	2	2.01	0.18	239.1	1	14/22
	K.VPQVSTPTLVEVSR.N (2)	2	2.50	0.42	911.6	1	18/26
	K.FQNALLVR.Y (3)	2	2.36	0.29	1111.5	1	14/14
	K.KQTALVELVK.H	2	2.36	0.38	623.7	1	13/18
	K.HPEAK.R	2	1.71	0.18	735.7	1	8/8
	R.RPC*FSALEVDETYVPK.E	2	3.34	0.53	571.8	1	18/30
	R.FKDLGEENFK.A (3)	2	2.42	0.40	1086.3	1	16/18
	K.EC*C*EKPLLEK.S	2	1.86	0.24	149.6	1	11/18
	K.VFDEFKPLVEEPQNLIK.Q (2)	2	5.23	0.52	777.3	1	21/32
	K.AAFTEC*C*QAADK.A	2	1.83	0.42	522.7	1	14/22
	K.QNC*ELFEQLGEYK.F	2	3.08	0.43	484.1	1	17/24
Cytochrome *C*	K.TGPNLHGLFGR.K (5)	2	2.89	0.55	519.7	1	14/20
	K.MIFAGIK.K (2)	1	1.78	0.24	825.1	1	10/12
	K.TGPNLHGLFGR.K	2	3.43	0.52	766.1	1	16/20
	K.KTEREDLIAYLK.K (2)	3	2.58	0.26	639.5	2	23/44
	K.GITWKEETLMEYLENPKK.Y	3	3.48	0.52	762.4	1	23/68
	K.TEREDLIAYLK.K	2	2.57	0.23	375.6	2	10/20
	K.EETLMEYLENPKK.Y (2)	2	3.67	0.45	681.4	1	18/24
	K.IFVQKCAQCHTVEK.G (2)	2	2.49	0.17	446.3	1	12/26
	K.EETLMEYLENPK.K (2)	2	2.25	0.13	412.5	1	13/22
Lactoglubin B	K.TKIPAVFK.I	1	1.57	0.23	203.7	4	7/14
	K.IPAVFK.I (3)	1	1.62	0.16	266.1	1	8/10
	K.VLVLDTDYKK.Y	2	1.88	0.42	610.7	1	13/18
	R.LSFNPTQLEEQC*HI.– (2)	2	3.29	0.51	690.8	1	16/26
	K.VLVLDTDYK.K	1	1.62	0.27	994.6	1	13/16
	R.VYVEELKPTPEGDLEILLQK.W	3	4.90	0.61	1304.2	1	33/76
	K.IDALNENK.V	1	2.23	0.25	321.2	1	10/14
	R.TPEVDDEALEKFDK.A (2)	2	3.69	0.52	1135.1	1	19/26
Lysozyme C	R.GYSLGNWVC*AAK.F	2	1.91	0.36	268.4	1	13/22
	R.HGLDNYR.G (2)	2	2.35	0.38	992.5	1	12/12
	R.C*ELAAAMK.R	1	2.05	0.42	302.8	1	10/14
	K.FESNFNTQATNR.N	2	2.79	0.42	1415.7	1	20/22
	K.GTDVQAWIR.G (2)	2	2.98	0.30	1400.8	1	15/16
Ovalbumin	K.HIATNAVLFFGR.C	2	3.26	0.67	1007.9	1	20/22
	K.ISQAVHAAHAEINEAGR.E (2)	3	4.11	0.55	636.5	1	25/64
	R.YPILPEYLQC*VK.E	2	3.53	0.42	570.1	1	18/22
	R.GGLEPINFQTAADQAR.E (2)	2	3.60	0.52	1742.1	1	22/30

Table 2 (*Continued*)

Protein	Sequence	z	Xcor	ΔC_N	Sp	Rsp	Coverage
Myoglobin	K.HGTVVLTALGGILKK.K	2	2.59	0.38	307.7	1	14/28
	K.HGTVVLTALGGILK.K	2	2.86	0.57	662.2	1	19/26
	K.YLEFISDAIIHVLHSK.H	2	4.17	0.58	1317.0	1	20/30
	K.ALELFRNDIAAK.Y	2	2.93	0.36	812.6	1	16/22
	K.VEADIAGHGQEVLIR.L	3	2.81	0.60	596.8	1	28/56
	–.GLSDGEWQQVLNVWGK.V	2	3.92	0.51	1767.9	1	20/30
Insulin	R.GFFYTPK.T (6)	1	1.65	0.21	173.8	2	8/12
Ribonuclease	K.C*KPVNTFVHESLADVKAVC*SQK.K (2)	3	4.33	0.48	821.8	2	25/84
	K.TTQANK.H	2	1.56	0.17	161.1	1	8/10
Ovomucoid	R.HDGGC*R.K	2	1.96	0.23	613.8	1	10/10
	K.VEQGASVDKR.H (2)	2	2.42	0.47	1232.3	1	17/18
Brain-specific homeobox/POU domain protein (human)	K.NM∼CKLKPLLNK.W	2	2.13	0.18	700.1	1	13/20
Adenylate cyclase. type I (bovine)	R.DDMEKVKLDNK.R	2	2.86	0.20	650.2	1	14/20
Sortilin-related receptor (chicken)	R.C*DDDNDC*RDWSDEANC*TMFR.T	3	2.66	0.15	508.5	1	22/76

(*) Carboxymethylation; (∼) oxidation.

y and b ions to relatively small peaks. So, manual examination could exclude these peptide hits. However, it would be profitable when these incorrectly assigned sequences could be excluded by using another criterion. To this end, the correlation between the mobilization time and the p*I* values was examined.

3.3. *pI as an additional data for peptide identification*

By CIEF, ordering of the peptides according to their p*I* value is expected. Fig. 3 displays the p*I* values of all identified peptides of the 10-protein mix (see Table 2), plotted against their mobilization times. The p*I* values used are theoretical p*I* values as determined by Sequest software.

Since not all peptides appear as sharp peaks, the times are based on the acquisition time of the MS/MS spectra. Because of the peak widths, in several cases the same peptide is identified more than once, at subsequent times. These multiple identifications are also taken up in the plot.

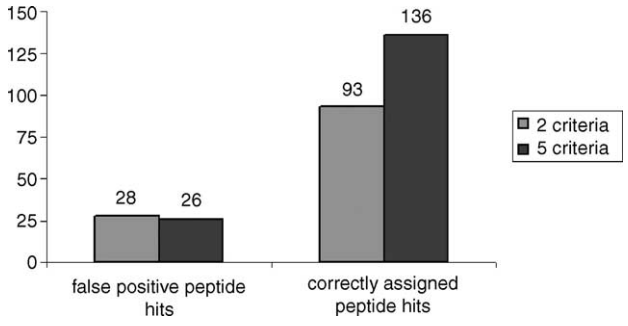

Fig. 2. Comparison data analysis. The same datasets of a known 10-protein mix were analyzed by Sequest software in two ways. First conventional thresholds for Xcor and ΔC_n as they are commonly used in literature were applied, then thresholds for all five available criteria: Xcor were set: ΔC_N, Sp, RSp and the coverage of *b*/*y* ions. The amount of incorrectly and correctly identified pep tides was compared.

Some scattering is seen, but there is a clear correlation. Based on in silico trypsin digestion of the 10 proteins, only a few peptides are obtained with a p*I* around 7, which results in the steeper slope in the plot around this value.

The scattering is likely a result from two effects. In the first place, the theoretical p*I* values will not fully correspond to the real values. Furthermore, as was shown in our previous study, the pH gradient is easily disrupted because of the limited buffering capacity of the peptides. During the experiment, the pH inside the capillary will decrease because of the

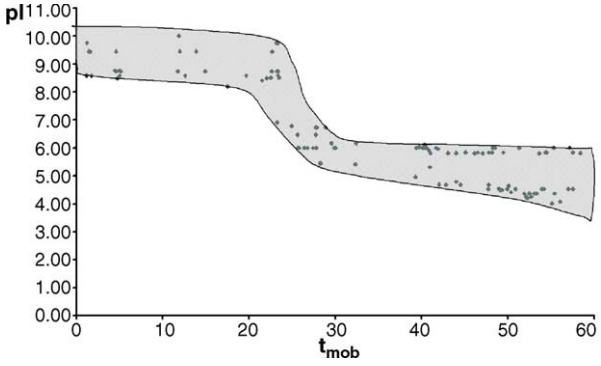

Fig. 3. p*I* values of the peptides identified for the 10-protein mix. Theoretical p*I* values as given by Bio works software were used. For the mobilization times of the peptides, the times of acquisition of their MS/MS spectra haves been taken. Making use of all points plotted, a restricted p*I* range can be drawn, which is displayed as the grey area.

Table 3
p*I* values of the false positives of the 10-protein digest

Peptide	t_{mob}	p*I* value	Restricted p*I* range (Fig. 3)
K.NM∼CKLKPLLNK.W	40.8	9.7	4.4–6.2
R.DDMEKVKLDNK.R	32.5	4.7	5.0–6.8
R.C*DDDNDC*RDWSDEANC*TMFR.T	47.1	3.8	4.2–6.1

(*) Carboxymethylation; (∼) oxidation.

Table 4
Filtered Sequest results for digest of periplasmatic proteins of *E. coli*

Protein	GI number (NCBI database)	Coverage (%)	Different peptides (*n*)
1. Alkaline phosphatase	16128368	71.5	27
2. Protein chain elongation factor EF-Tu	15803852	66.5	25
3. GTP-binding protein chain elongation factor EF-Tu	15803853	52.7	27
4. *sn*-Glycerol 3-phosphate transport system	15803962	65.5	21
5. Malate dehydrogenase	15803770	69.9	16
6. Enolase	15832893	43.1	16
7. Isocitrate dehydrogenase	16129099	45.9	17
8. PstS	16131596	32.1	9
9. A64948 adhesin homolog yebL precursor	7449165	49.4	10
10. Protein chain elongation factor EF-Ts	15799852	63.7	17
11. OppA (periplasmatic oligopeptide-binding protein precursor)	16129204	36.5	16
12. Phosphoglycerate kinase	16130827	55.0	13
13. Alkyl hydroperoxide reductase, C22 subunit	15800320	67.4	14
14. Periplasmic binding protein Component of Pn transporter	16131931	45.0	13
15. Dipeptide transport protein	16131416	32.2	14
16. Phosphoglyceromutase 1	15800464	26.0	7
17. Tetrahydropteroyltriglutamate methyltransferase	16131678	21.0	11
18. Adenylate kinase activity	15800203	42.5	11
19. Fructose-bisphosphate aldolase class II	15833050	36.5	10
20. Aspartate aminotransferase	16128895	22.0	10
21. Hypothetical protein b1973	16129919	22.7	5
22. Chaperone Hsp70	15799694	17.2	8
23. Acyl carrier protein	15801211	33.3	4
24. Threonine synthase	16127998	33.6	10
25. BLEC leucine/isoleucine/valine-binding protein precursor	7428840	36.8	10
26. Glyceraldehyde-3-phosphate dehydrogenase A	15802193	29.3	9
27. Osmotically inducible protein Y	16132194	27.4	4
28. PTS system, glucose-specific IIA component	16130343	41.4	5
29. Universal stress protein; broad regulatory fu	15804030	35.4	3
30. Serine hydroxymethyltransferase	16130476	25.9	11
31. Putative receptor (hypothetical protein b1452)	16129411	26.6	7
32. Hypothetical protein (quorum-sensing protein)	16130599	32.8	4
33. Hypothetical protein b3509	16131381	47.3	5
34. Thiol peroxidase	15801846	18.4	2
35. SAICAR synthetase	15802999	35.0	6
36. Triosephosphate isomerase	15804508	30.6	5
37. Transaldolase B	15799688	21.8	7
38. 3-Oxoacyl-[acyl-carrier-protein] synthase I	16130258	16.3	6
39. Thioredoxin 1	15804371	30.7	4
40. trp Repressor binding protein	16128970	17.2	4
41. Glucosephosphate isomerase	15804618	9.3	4
42. Beta-hydroxydecanoyl thioester dehydrase	15800813	28.5	4
43. Malonyl-CoA-[acyl-carrier-protein] transacylase	16129055	21.4	5
44. GroES (10 kD chaperone)	15804734	42.3	3
45. PTS system protein HPr	15802948	70.6	5
46. LivK (periplasmatic binding protein)	16131330	6.1	2
47. Aspartate-semialdehyde dehydrogenase	15803942	21.0	4
48. Succinyl-CoA synthetase, beta subunit	15800432	13.1	6
49. Trigger factor	15800166	16.1	6
50. Isocitrate dehydrogenase (NADP+)	4062742	38.9	2
51. Succinyl-CoA synthetase, alpha subunit	15800433	17.0	4
52. Tryptophan synthase alpha K	16129221	11.6	3
53. Protein disulfide isomerase I	15804445	13.5	2
54. Phosphoribosylglycinamide formyltransferase 2	16129802	14.0	4
55. 2,3,4,5-Tetrahydropyridine-2-carboxylate *N*-succinyltransferase	16128159	15.3	4
56. Enoyl-[acyl-carrier-protein] reductase (NADH)	15801888	11.1	2
57. UDP-Glucose 4-epimerase (galactowaldenase)	120920	16.0	3
58. 5-Enolpyruvylshikimate-3-phosphate synthetase	15800769	8.2	3
59. Putative GTP-binding factor	16131711	4.4	3
60. Methionine adenosyltransferase 1	15803481	5.2	2
61. Bacterioferritin comigratory protein	15803003	21.8	2
62. 5,10-Methylenetetrahydrofolate reductase	16131779	19.3	3
63. Hypothetical protein	16128314		2

Table 4 (*Continued*)

Protein	GI number (NCBI database)	Coverage (%)	Different peptides (*n*)
64. orf, hypothetical protein 24922	15804054	31.8	3
65. Galactose-binding transport protein	16130088	14.8	3
66. Hypothetical protein b1967	16129913	9.9	2
67. 3-Phosphoserine aminotransferase	16128874	15.5	4
68. Transketolase 1 isozyme	16130836	8.3	3
69. D-3-Phosphoglycerate dehydrogenase	15803448	6.6	2
70. Galactokinase	15830039	7.9	2
71. Aconitate hydrase B	16128111	9.1	3
72. Ketol-acid reductoisomerase	16131632	9.2	3
73. Lysine, arginine, arnithine-binding periplasmatic protein	16130245	8.5	2
74. High-affinity glycine betaine/proline transferase	15832796	16.4	3
75. Mannitol-1-phosphate dehydrogenase	16131471	6.5	2
76. Survival protein	15799738	6.3	2
77. Asparagine tRNA synthetase	16128897	5.8	2
78. NAD synthetase	16129694	8.4	2
79. 3-Isopropylmalate isomerase (dehydratase) subunit	16128066	8.2	2
80. Oxygen-insensitive NAD(P)H nitroreductase	16128561	15.2	3
81. Histidine-binding periplasmic protein of high	15802856	23.1	4
82. 30S ribosomal subunit protein S1	15800772	4.7	4
83. Ribosome releasing factor	15799854	11.4	2
84. orf, hypothetical protein z4376	15803566	20.0	3
85. Hypothetical protein	4062579	4.9	2
86. Aminoacyl-histidine dipeptidase	16128223	5.8	4
87. Homoserine kinase	16127997	10.0	2
88. Stringent starvation protein A	15803763	9.7	2
89. Superoxide dismutase manganese	33347807	9.7	2
90. Transcription elongation factor	15803721	15.7	2
91. 2,3-Dihydro-2,3-dihydroxybenzoate synthetase	15800310	8.8	2
92. Arginine 3rd transport sustem 897y897	16128828		
93. 50S ribosomal subunit protein L9 [Escherichia	15804792	15.4	2
94. FKBP-type 22 kD peptidyl-prolyl *cis–trans* isomerase	15834439		2
95. Peptidyl-prolyl *cis–trans* isomerase B	16128509	13.4	2
96. Transaldolase A	15802986	8.2	2
97. L-Histidinal:NAD+ oxidoreductase	16129961	5.3	2
98. Adenine phosphoribosyltransferase	16128453	17.5	2
99. Homolog of Salmonella UTP–glucose-1-P uridyltransferase	15802521	6.4	2
100. Hypothetical protein b1178	16129141	15.8	2
101. Heat shock protein	15832709	3.0	2

progressive migration of acetate ions towards the anode and H^+ towards the cathode when voltage is applied. The peptides will obtain a charge, resulting in electrophoretic mobility, and deviations from the isoelectric ordering. To what extent these effects can be seen, is dependent on the concentration of the digest itself and the concentration of the carrier ampholytes, if added.

For every mobilization time, the pH range of the peptides that elute is restricted to a range of about two pH units. This is marked by the area in Fig. 3. Though this range is chosen in a rather subjective way, it meets the demands of this research, which is to assess the possibility of using the p*I* value as an extra criterion.

This range can now be used to evaluate ambiguous peptide hits. To give an example of this, Table 3 shows the retention times and the p*I* values of the false positives that were taken up in Table 2. The false positive identifications are not within the pH range as indicated in Fig. 3. Using this information, these spectra could be excluded beforehand without studying the MS/MS spectra.

3.4. Carrier ampholyte free CIEF-MS/MS of a digest of E. coli periplasmatic proteins

To investigate whether the described procedure is also applicable for biological samples, a digest of *E. coli* periplasmatic proteins was analyzed. As reported before, a higher concentration of the sample leads to a better separation efficiency in carrier ampholyte-free CIEF. Therefore the sample was washed and concentrated by ultrafiltration. A concentration of 10 mg/ml was digested by trypsin. Before analysis, the sample was diluted 1:1 with water. At this concentration, no carrier ampholytes are needed. The electropherogram is shown in Fig. 4.

For the data analysis by Sequest, the five optimized parameters were used. The question might arise whether the threshold values are transferable to a complete different dataset from another species. Especially the size of the database might have its influence on the optimum values. Therefore, extra caution was taken. For all proteins that were identified with less than three peptides, more stringent criteria were

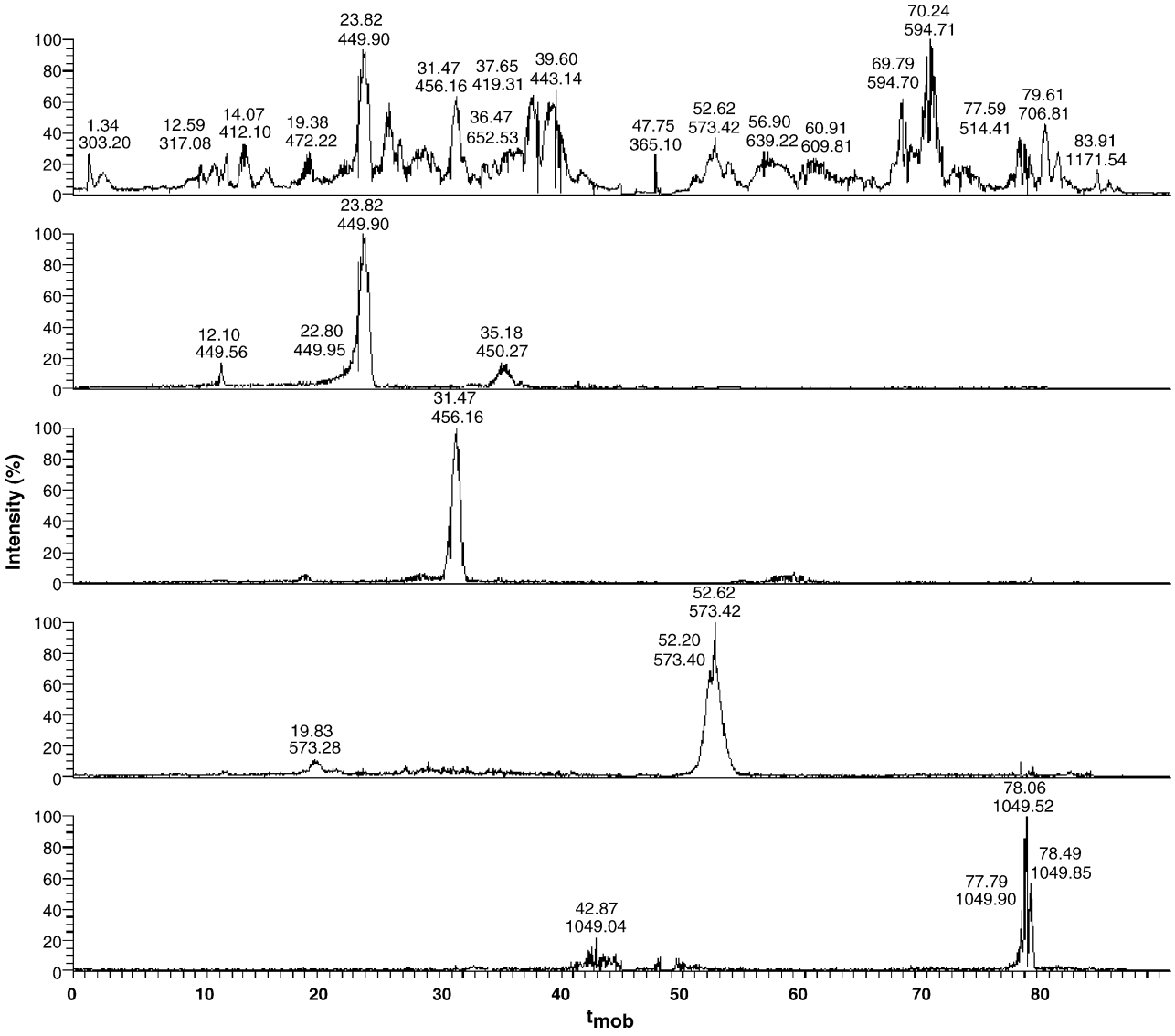

Fig. 4. Electropherogram of a digest of an *E. coli* periplasmatic protein extract. The sample was focused for 12 min using 25 kV. Mobilization was achieved by applying 70 mbar of pressure until the first peaks were seen. Then ($t = 0$ min), 20 mbar was applied for the first 25 min, 50 mbar until $t = 60$ min and then 80 mbar until the end. The electropherogram on top shows the total ion current. Below, four masses have been selected to show their peak shape.

chosen. At least one of the peptides had to comply with the higher criterion for the Xcor: 1.8 for single charged, 2.1 for double charged and 2.5 for triple charged, respectively. On top of that, all MS/MS spectra for these proteins were manually evaluated. For all proteins identified with less than five peptides, at least one spectrum was manually evaluated, randomly chosen.

Table 4 a list of the proteins that were identified. When only proteins that are identified with at least two peptides are taken into account, already 101 are identified, some with sequence coverage of over 80%.

Concerning the identified proteins, it might appear strange that elongation factor Tu would be present in the periplasm, elongation factors exclusively functioning within the nucleus. However, some earlier papers have already reported on their release in the periplasm as a result from osmotic shock treatment [18,19]. Thus, its presence in considerable quantities is expected.

On top of the proteins that were identified with multiple peptides, for 61 proteins only one peptide was identified. To assess the reliability of these hits, we made use of the theoretical p*I* values.

Fig. 5 displays the p*I* values of the peptides identified for the first 25 proteins in Table 4, plotted against their mobilization times, similarly as for the 10-protein mix in Fig. 3. Again, because of the peak widths, in several cases the same peptide is identified more than once, at subsequent times. These multiple identifications are also taken up in the plot. An exception was made for higher abundant peptides, since these were sometimes still identified at times removed from the main body of the peak shapes. For these, only the times were chosen corresponding to their peak areas.

Table 5
p*I* values of the 'single peptide' hits

Protein	Sequence	t_{mob}	p*I*
1. Elongation factor P (gi 15834382)	K.VPLFVQIGEVIK.V	68.2	5.97
2. Hypothetical protein (gi 15831005)	R.LQLLHDEGR.L	57.8	5.32
3. Tryptophan synthase beta (gi 16129222)	K.TNQVLGQALLAK.R	32.2	8.41
4. 50S ribosomal subunit protein L7/L12	K.DLVESAPAALK.E	84.9	4.37
5. orf, Hypothetical protein Z01770 (gi 15799843)	K.IGIIGAMEEEVTLLR.D	81.9	4.25
6. orf, Hypothetical protein Z2676 (gi 15802068)	R.QIAENPILLYMK.G	71.6	6.0
7. Cysteine synthase A, *O*-acetylserine sulfhydrolase A	K.ALGANLVLTEGAK.G	69.0	6.1
8. 2-Deoxyribose-5-phosphate aldolase	R.FGASSLLASLLK.A	34.7	8.8
9. Hypothetical protein b1019 (gi 16130800)	R.LPLTLMTLDDWALATITGADSEK.Y	86.2	3.8
10. Arginine 3rd transport system periplasmic binding protein	R.IDGVFGDTAVVTEWLK.D	85.3	4.0
11. orf, Hypothetical protein z0529 (gi 15800156)	K.VLSESDFQVNQLLDILR.A	85.0	4.0
12. Hypothetical protein b0329 (gi 16128314)	K.AEFEKVESQYEK.I	79.1	4.5
13. FKBP-type peptidyl-prolyl *cis–trans* isomerase	K.DVFMGVDELQVGMR.F	86.1	4.0
14. Adenosine 5′-phosphosulfate kinase	K.STVAGALEEALHK.L	53.1	5.4
15. Putative dehydrogenase (gi 16128769)	R.AFGQVAAHEAMALGIEK.A	55.0	5.4
16. Hypothetical protein b1171 (gi 16129134)	K.DPQMLLITAIDDTMR.A	86.9	3.9
17. Imidazole glycerol phosphate synthase holoenzyme	R.DPDVVLLADK.L	86.1	3.9
18. Spermidine synthase (putrescine aminopropyltransferase)	K.HVLIIGGGDGAMLR.E	55.4	6.7
19. Cold shock protein (gi 15800338)	K.GFGFITPEDGSK.D	82.2	4.4
20. FKBP-type peptidyl-prolyl *cis–trans* isomerase	K.LDKDQLIAGVQDAFADK.S	84.4	4.1
21. Putative alpha helix protein	R.KLENLTDIER.Q	63.7	4.7
22. 2-Dehydro-3-deoxyphosphooctulonate aldolase	K.KPQFVSPGQMGNIVDKFK.E	19.8	9.7
23. Glutaredoxin 2	K.RSPAIEEWLR.K	51.5	6.1
24. Glutaredoxin 3	K.GVSFQELPIDGNAAK.R	82.3	4.4
25. Outer membrane protein 3a (II*;G;d)	K.DVVTQPQA.-	88.8	3.8
26. PhoU (gi 16131592)	R.HTIQMLHDVLDAFAR.M	52.8	6.0
27. PEP-protein phosphotransferase system enzyme	R.TMDIGGDKELPYMNFPK.E	79.7	4.6
28. Gluconate-6-phosphate dehydrogenase (decarboxylating)	K.QQIGVVGMAVMGR.N	37.5	9.8
29. Trehalase	R.HFVNVNFTLPK.E	33.0	8.8
30. Heat shock protein (gi 15832709)	K.RLDMLNEELSDKER.Q	78.3	4.5
31. Isoleucine tRNA synthetase	R.LGVLGDWSHPYLTMDFK.T	77.8	5.2
32. Hypothetical protein b1990 (gi 16129931)	R.LYAIHGTNANFGIGLR.V	29.1	8.8
33. Hypothetical protein b1780 (gi 33347584)	R.KLDELDLIVVDHPQVK.A	77.1	4.7
34. Aspartokinase III	K.VLHPATLLPAVR.S	27.0	9.7
35. 2-Isopropylmalate synthase	K.AIVGSGAFAHSSGIHQDGVLK.N	33.8	7.0
36. orf, Hypothetical protein Z5276 (gi 15804356)	R.HGYAFNELDLGKR.E	46.2	6.8
37. Peptide chain release factor RF-3	R.TFAIISHPDAGK.T	39.0	6.4
38. Hypothetical protein b1034 (gi 16128997)	K.HQVALEINNSSFLHSR.K	37.3	6.9
39. Putative yhbH sigma 54 modulator	K.HEDMYTAINELINKLER.Q	75.7	4.4
40. Uracil phosphoribosyltransferase	R.AGLGMMDGVLENVPSAR.I	82.6	4.4
41. Glutamate-1-semialdehyde aminotransferase	R.AFTGVGGTPLFIEK.A	68.6	6.1
42. Putative arylsulfatase (gi 16128838)	R.NLALHVDGAR.I	38.8	6.7
43. Phosphotransferase system enzyme IIA	K.THLHTLSLVAK.R	27.0	8.5
44. Purine-nucleoside phosphorylase	R.VGSC*GAVLPHVK.L	24.4	8.2
45. 3-Isopropylmalate dehydrogenase	K.ANVLQSSILWR.E	29.3	9.8
46. Tyrosine aminotransferase	R.HAIAPLLFGADHPVLK.Q	36.8	6.9
47. Putative GTP-binding protein (gi 15830962)	K.C*GIVGLPNVGK.S	38.9	8.22
48. orf, Hypothetical protein Z0588 (gi 15800200)	R.RVEIDPSLLEDDKEMLEDLVAAAFNDAAR.R	89.2	4.0
49. orf, Hypothetical protein Z5710 (gi 15804700)	K.DANGNLLADGDSVTIIK.D	86.4	3.9
50. B64858 probable ATPase ycfB	K.IAEDLGLVTAK.K	83.1	4.4
51. Mannose-6-phosphate isomerase	R.LSELFASLLNMQGEEK.S	82.3	4.3
52. Phosphoheptose isomerase	R.NELNEAAETLANFLK.D	83.7	4.3
53. Hypothetical protein, CP4-57 prophage (gi 16130543)	K.NIILQFGPNK.F	30.1	8.8
54. Phosphoglycerate mutase III, cofactor-independent	R.AFFANPVLTGAVDK.A	77.3	5.9
55. Glutamine tRNA synthetase	K.GVIHWVSAAHALPVEIR.L	33.3	6.9
56. Cold shock protein (gi 15802236)	K.GFGFITPADGSK.D	79.0	5.8
57. Menaquinone biosynthesis, unknown (gi 15804520)	R.ASFGGQIITVK.C	29.7	8.8
58. OsmC, osmotically inducible protein (gi 16129441)	K.AEITLDYQLK.S	82.6	4.4
59. Chaperonin GroEL	K.DTTTIIDGVGEEAAIQGR.V	86.8	3.9
60. Guanine-hypoxanthine phosphoribosyltransferase	K.YIVTWDMLQIHAR.K	40.3	6.7
61. Putative aminotransferase (gi 15803057)	K.VDLMSFSGHK.I	35.5	6.7

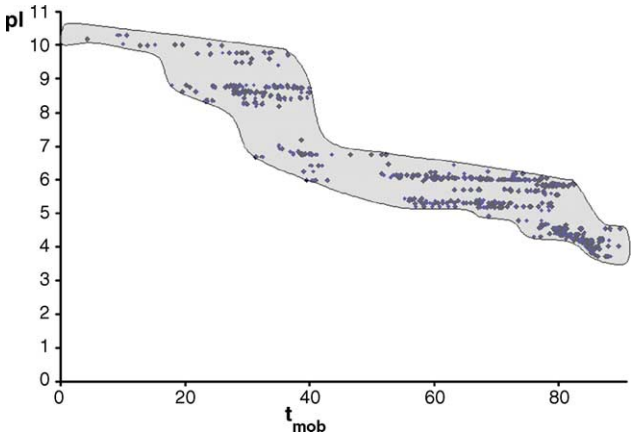

Fig. 5. p*I* values of peptides identified for the *E. coli* periplasmatic protein digest. Theoretical p*I* values as given by Bioworks software were used. For the retention times of the peptides, the times of acquisition of their MS/MS spectra haves been taken. Making use of all points plotted, a restricted p*I* range can be drawn, which is displayed as the grey area.

For every mobilization time, the pH range of the peptides that elute is restricted to a range of about two p*I* units. This is marked by the area in Fig. 5. The plot was used to evaluate those protein hits that were based on a single peptide match. All these peptides, with their mobilization times and their p*I* values, are presented in Table 5.

Except for one, all peptides fall within the tolerated pH range. This adds to the credibility of these hits. Though 100% reliability cannot be obtained for the individual hits, the fact that hardly any peptides are dismissed on the basis of their p*I* value does confirm the overall reliability of the used criteria.

Interestingly, one of the peptides of the first 25 proteins fell well out of the restricted p*I* range. The sequence of this peptide was RSDIEIVAINDLL from the protein glyceraldehyde 3-phosphate dehydrogenase, with a mobilization time of 68.5 min and a p*I* value of 4.0. The tolerated p*I* range in this part of the plot is 4.3–6.3. However, the same MS/MS spectrum was also assigned to the sequence HMGWTEAADLIVK from the protein isocitrate dehydrogenase, even with a higher Xcor value. Since the p*I* value of this peptide was 5.3, this one is more likely to be the correct identification. This shows again that the p*I* values can be very useful for identification purposes.

4. Conclusion

The experiments described above have shown that CIEF coupled to MS is suitable for complex protein mixtures. CIEF automatically results in concentrating of the analytes, which helps to increase the number of proteins identified.

In a mixture of known proteins, ovomucoid was identified as one of the constituents. Since this protein was only present as an impurity, its concentration was considerable lower than the concentration of the other constituents. This suggests that the described method has a considerable dynamic range concerning the concentration.

The analysis of the crude extract of periplasmatic proteins from *E. coli*, suggests the same. A vast amount of proteins was identified, 159 in total, which is a very high amount for a system with a single mode of separation. As compared to the literature, such a high number of identified proteins easily matches the number of identified proteins when 1D-LC is used, while the amount of sample used is less.

We expect that, by performing a prefractionation of the proteins in the sample, even more can be identified. This would be equivalent to the use of 2D-LC, in which the peptides are fractionated. Especially when high concentrations of the fractions can be obtained, identification of a substantial part of the proteome is feasible. The described method might also be applicable with lower concentrations, by adjusting the amount of carrier ampholytes as was described in our earlier study.

The potential of this method was furthermore shown by the fact that a good correlation could be seen between the mobilization time and the theoretical p*I* values of the peptides.

By assessing the theoretical p*I* values in a more accurate mater, some improvement might be possible. Furthermore, the addition of carrier ampholytes might result in improved correlation. This will also result in ion suppression, so an optimum has to be found. Already, the plot obtained could be used to evaluate the ambiguous peptide hits, and some peptides could be excluded. Furthermore, when two possible peptide hits were given by the software of almost equal probability, the p*I* value could function as an extra criterion. It would be possible to integrate such a feature in the software used for the analysis of the MS/MS spectra, especially when software is used that uses probability scores.

Acknowledgement

We thank L. de Vrind-de Jong for providing us with the periplamstic protein extract of *E. coli* cells.

References

[1] A.J. Link, J. Eng, D.M. Schieltz, E. Carmack, G.J. Mize, D.R. Morris, B.M. Garvik, J.R. Yates, Nat. Biotechnol. 17 (1999) 676.
[2] G.J. Opiteck, K.C. Lewis, J.W. Jorgensons, R.J. Anderegg, Anal. Chem. 69 (1997) 1518.
[3] H.F. Storms, R. van der Heijden, U.R. Tjaden, J. van der Greef, Electrophoresis 25 (2004) 3455.
[4] S. Hjerten, M.D. Zhu, J. Chromatogr. 346 (1985) 265.
[5] O. Sova, J. Chromatogr. 320 (1985) 15.
[6] J. Pospíchal, E. Glovinová, J. Chromatogr. A 918 (2001) 195.
[7] M. Yata, K. Sato, K. Ohtsuki, M. Kawabata, J. Agric. Food. Chem. 44 (1996) 76.
[8] J.W. Hager, Rapid Commun. Mass Spectrom. 16 (2002) 512.
[9] S. Purvine, A.F. Picone, E. Kolker, OMICS J. Integr. Biol. 8 (2004) 79.

[10] A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold, Anal. Chem. 74 (2002) 5383.
[11] D.C. Anderson, L. Weiqun, D.G. Payan, W.S. Noble, J. Proteome Res. 2 (2003) 137.
[12] P. Matsudaira, A Practical Guide to Protein and Peptide Purification for Microsequencing, second ed., Academic Press, San Diego, 1993, p. 48.
[13] Methods in Enzymology, vol. 22, Academic Press, New York, 1971.
[14] M.M. Bradford, Anal. Biochem. 72 (1976) 248.
[15] D. Schmalzing, C.A. Piggee, F. Foret, E. Carrilho, B.L. Karger, J. Chromatogr. A 652 (1993) 149.
[16] M.H. Lamoree, U.R. Tjaden, J. van der Greef, J. Chromatogr. A 777 (1993) 31.
[17] R.K. Bush, S.L. Taylor, J.A. Nordlee, Allergy Proc. 10 (1989) 261.
[18] C. Berrier, A. Garrigues, G. Richarme, A. Ghazi, J. Bacteriol. 182 (2000) 248.
[19] G.R. Jacobson, J.P. Rosenbusch, Nature 261 (1976) 23.