

INTERNAL PROTEIN SEQUENCING OF SDS PAGE-SEPARATED PROTEINS A COLLABORATIVE ABRF STUDY

Ken Williams 1, Ulf Hellman 2, Ryuji Kobayashi 3, William Lane 4, Sheenah Mische 5, and David Speicher 6

IHMI Biopolymer Laboratory/W.M. Keck Foundation Biotechnology Resource Laboratory, Yale University, New Haven, CT 06536; 2Ludwig Institute for Cancer Research, Uppsala, Sweden; 3 Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; 4 Microchemistry Facility, Harvard University, Cambridge, MA 02138; 5 Protein/DNA Technology Center, Rockefeller University, New York, NY 10021; and 6The Wistar Institute, Philadelphia, PA 19104

I. Introduction

Since many eukaryotic proteins have blocked NH₂-termini (1) and SDS polyacrylamide gel electrophoresis (PAGE) appears to be the current method of choice for final purification of proteins destined for amino acid sequencing, internal sequencing of these samples represents an important core laboratory activity that had not yet been addressed in a collaborative Association of Biomolecular Resource Facilities (ABRF) study. The goals of this first such study were five-fold. 1) provide a mechanism for ABRF laboratories to anonymously compare their internal sequencing capabilities with other core laboratories, 2) provide a reasonable sample and well proven protocols to facilitate introduction of this technology into those laboratories that do not yet offer internal sequencing, 3) obtain data that may help determine the relative efficacy of internal sequencing from PVDF blots versus from in-gel samples, 4) determine if there are any significant commonalities among the best in-gel and PVDF digests to help optimize these protocols, and 5) compile data obtained by multiple laboratories on the same "unknown" sample that may help establish realistic expectations for internal sequencing.

II. Materials and Methods

A. *Sample Preparation and Distribution*

The 1996 ABRF internal sequencing samples consisted of three samples: 1) a 28 kD recombinant β -spectrin fragment; 2) the same β -spectrin fragment with an additional, unique, 15-residue tryptic peptide sequence inserted near its

NH₂-terminus resulting in a mass of about 30kD and 3) an external peptide standard (450 pmol) that was provided dry in an eppendorf tube and that had the same amino acid composition as the unique tryptic peptide insert but whose sequence was randomized. In the case of the two protein samples, 70 pmol of each had been subjected to SDS PAGE and were supplied either as Coomassie Blue stained gel slices or as a section of amido black stained PVDF membrane. In the case of the PVDF samples, an oversize piece of PVDF was included so that a section could be used as a digest control, and in the case of the gel samples, a blank section of gel was included for the same purpose in a separate eppendorf tube.

In response to a descriptive letter sent to 258 ABRF Directors, which provided the option of receiving either the PVDF or gel samples, or both sets of samples, 112 laboratories requested a total of 100 PVDF and 90 gel samples.

B. Protocol for the 1996 ABRF Internal Digest Study

Participants were requested to digest the two protein samples and the control with trypsin following either their own procedure or a representative procedure included with the samples. Since neither protein contained cysteine, modification of this amino acid was not required. Participants were then asked to subject the three digests and 22.5 pmol of the external standard to reverse phase HPLC and to forward the resulting chromatograms, along with a 3 page sample data sheet, to the Internal Protein Sequencing Committee. Anonymity of participants was ensured by having the data returned to the Committee via a disinterested third party, who numbered the data sets in order of receipt and removed all identifying marks. For those laboratories that wished to proceed further, it was suggested that the 30kD digest be collected and that the unique, 15 residue tryptic peptide insert be further characterized by mass spectrometry and/or amino acid sequencing. This "target" peptide could be identified by its presence in the 30 kD digest and absence in the 28kD digest, by its above average absorbance due to the presence of aromatic amino acids (see below), and by its elution close to the external standard peptide. A minor complication occurred during the week long period of time required to prepare the 190 samples. Apparent partial proteolysis occurred near the COOH-terminal region of the 30 kD protein which resulted in some cross-contamination of the 30kD fragment into the 28kD gel band. Based on NH₂-terminal sequencing of selected samples, the ratio of the target peptide in the 30 kD versus 28 kD bands on SDS PAGE varied from ~4: 1 to 2: 1 instead of the target peptide being unique to the 30kD sample.

C. Design of the Unique Peptide Insert

The NH₂-terminus sequence of the recombinant 30kD fragment was

NH₂ -G-S-P-K-N-Y-E-V-H-T-W-D-V-E-L-S-Q-F-K-G-S-V...

The primary concerns in choosing the sequence of the unique, peptide insert were that a tryptic digestion of the 30kD sample would release the target

peptide (underlined above) in good yield and that it should not co-elute with other major peaks in the 30kD chromatogram. Hence, the peptide was preceded and followed by lysine, proline was avoided after the lysines, and no acidic residues were included near either intended tryptic cleavage site. The 5 residue length was chosen to be within the range commonly seen for tryptic peptides. To ensure the target peptide was a major absorbance peak, one tryptophan and one tyrosine were included, and to avoid the necessity of reduction/cysteine modification, cysteine was not included. To avoid co-elution with other tryptic peptides derived from the 28 kD protein, the amino acid composition of the peptide was chosen so that it would elute near 30% acetonitrile based on published retention coefficients and a constant parameter that is a function of the particular column and HPLC system being used (2).

A synthetic peptide analogue of the unique insert actually eluted at about 28% CH_3CN both from a Vydac C-18 column on the system being tested (data not shown) and from a Zorbax C-18 column on an HPLC system located in a different laboratory (Fig. 1). In the latter instance the peptide insert eluted close to a minor peak eluting at about 54 min in the 28kD chromatogram. As noted previously, the external peptide standard had the same amino acid composition but a different sequence from that of the unique peptide insert. The external standard had the following sequence:



Somewhat surprisingly, the external standard usually eluted at an CH_3CN concentration that was from 5-8% less than that of the target peptide.

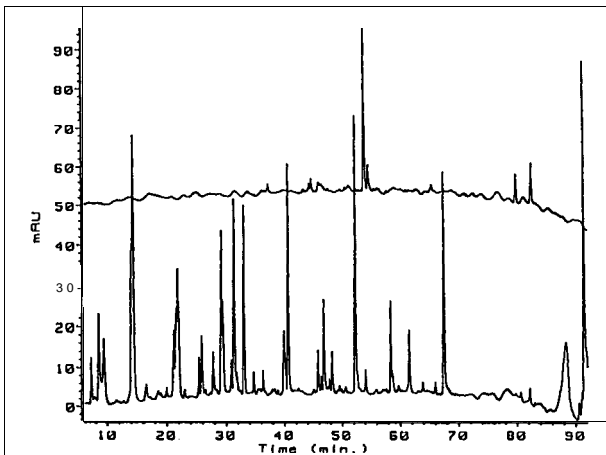


Figure 1. Reverse phase HPLC separation of the external peptide standard (50 pmol, top chromatogram) and an *in situ* PVDF digest of the 28 kD recombinant (bottom chromatogram). Following SDS PAGE of a mixture of 50 pmol of each protein and blotting onto PVDF, the 28 kD protein was digested with trypsin (3) and subjected to reverse phase HPLC on a Zorbax C18 column (1 x 150 mm) eluted at 371 G at a flow rate of 75 $\mu\text{l}/\text{min}$. The column was equilibrated with 95% buffer A (0.06% TFA) and 5% buffer B (0.055% TFA in CH_3CN) and was then brought to 33% and 60% buffer B with linear gradients extending to 63 and 95 min respectively.

D. Data Analysis

By reference to the external peptide standard it was possible to correct for differences in flow rates, path cell lengths, and other HPLC variables and to thus subject the chromatographic profiles to semi-quantitative analysis. Hence, the relative peak height for each 30kD chromatogram was calculated from the sum of the measured peak heights of the most intense peaks (avoiding obvious artifact peaks at the beginning and end of the profiles) relative to that of the external standard peak height. The number of peaks in 30 kD chromatograms was defined as the number of peaks with >20% the peak height of the external peptide standard. Similarly, the number of background peaks was defined as the number of peaks in the blank digest with >20% the peak height of the external standard. A composite, relative chromatography score was calculated by adding together the relative peak heights and the number of 30 kDa peaks and then subtracting the number of background peaks from this sum. In each of these three categories, the ratio of the individual score to that of the best score was calculated prior to calculating a composite score. Hence, the composite scores can range between 1.0 (worst) and 2.0 (best). A qualitative assessment of chromatographic reproducibility was based on overlaying the 30 kD and 28 kD chromatograms to determine if it was reasonably possible to identify co-eluting peaks in these two chromatograms. The sequencing yield for the target peptide was based on the reported yield of valine at position 4 (Val4) in the sequence.

III. Results

As shown in Table I, 76% of the 39 laboratories that participated in this study routinely carry out *in situ* PVDF and/or in-gel digests, and trypsin (78%) or endoproteinase Lys-C (43%) are the two most frequently used enzymes. The most commonly cited protocols that were routinely used included those by

Table I. Summary of responses to selected sample submission questions

Question	n	Response
Routinely perform in-gel or PVDF digestions?	38	76%
Routinely use peptide mass database algorithms for protein identification?	39	26%
Perform mass analysis of HPLC isolated peptides prior to sequence analysis?	39	39%
Routinely provide database search as a service in your laboratory?	35	86%
What percentage of the proteins you receive for sequence analysis are N-terminally blocked?	24	25% (2-80) a
What percentage of the proteins you receive for sequence analysis ultimately prove to have already been sequenced as evidenced by database searches?	28	60% (10-95) a

^aMedian value is given followed by the range.

Fernandez *et al.* (3), 3 1%, for PVDF, and those by Rosenfeld *et al.* (4), 15%, and Hellman *et al.* (5), 10% for in-gel digests. The most commonly used HPLC columns were C 18 (58%), and the most commonly used column dimensions were 2 to 2.1 mm (67%) with lengths between 150-250 mm (62%). Although respondents indicated an average of 60% of proteins submitted for internal sequencing ultimately prove to have already been sequenced, only 6% of the participants routinely use peptide mass database algorithms for protein identification (Table I). Since <40% of the participants routinely perform mass analysis of HPLC isolated peptides prior to sequence analysis (Table I), this suggests the relatively low fraction of facilities that routinely use peptide mass searching primarily reflects lack of routine access to necessary equipment. In view of data suggesting that about 80% of soluble proteins from Ascites cells are N-a-acetylated (1), it is somewhat surprising that participants in this study estimate that only 25% of proteins received for sequencing are blocked (Table I). However, the very large range (from 2-80%) in responses to this question suggests that either some laboratories may receive a high proportion of proteins from prokaryotic sources where the occurrence of blocked N-termini is very low (6) or there is considerable error in this estimate.

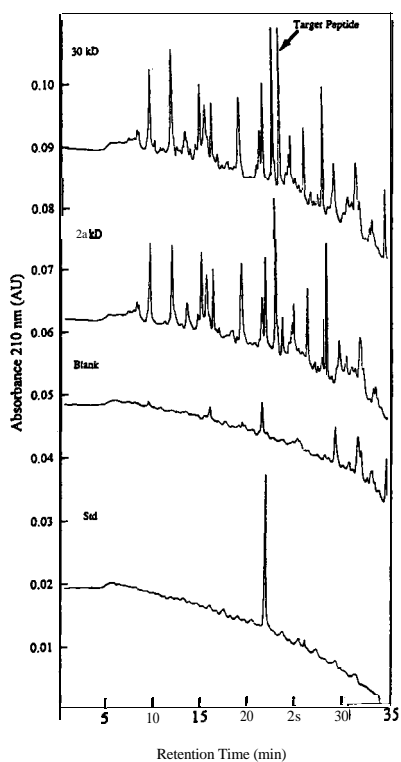


Figure 2. An above average *in situ* PVDF digest (laboratory #1). HPLC separation was at room temperature on a 1 x 250 mm Aquapore RP300 column at 150 μ l/min. Initial conditions at 100% buffer A (0.1% TFA) were followed by linear gradients to 55% and 85% buffer B (0.08% TFA in 70% CH₃CN) to 30 and then to 40 min respectively.

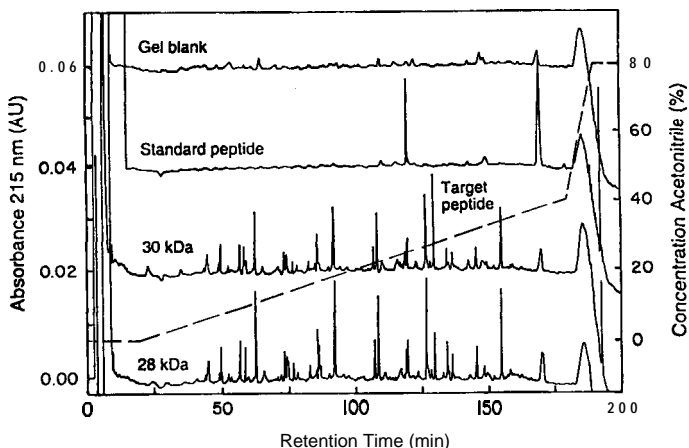


Figure 3. An above average *in situ* gel digest (laboratory #7). HPLC separation (80% of digest was injected) was at 21 °C on a 2.1 x 100 mm Pharmacia μ RPC (C2/C 18) column at 100 μ l/min. The column, equilibrated at 100% buffer A (0.065% TFA) was eluted isocratically for 20 min followed by linear gradients to 40% and 80% buffer B (0.05% TFA in CH_3CN) to 180 and 190 min respectively.

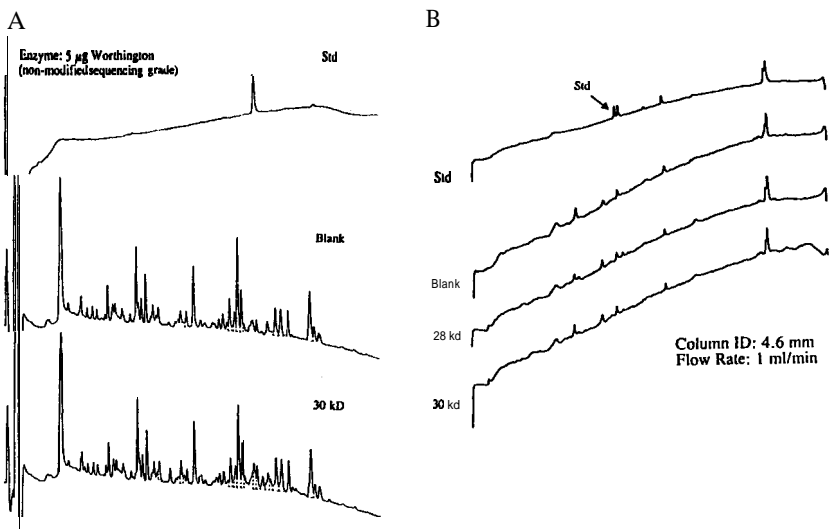


Figure 4. Examples of submitted HPLC chromatograms with either high background (A) or without significant 30 kDa absorbance peaks (B). (A) This *in situ* PVDF data set (laboratory #27) was chromatographed at 30 °C on a Vydac C18 column (2.1 x 250 mm) eluted at 500 μ l/min and monitored at 200 nm. The column equilibrated with 100% buffer A (0.1% TFA) was brought to 70% and then to 100% buffer B (0.075% TFA in 70% CH_3CN) with linear gradients extending to 60 and 70 min respectively. (B) This *in situ* PVDF data set (laboratory #16) was obtained at room temperature on a Vydac C 18 column (4.6 x 150 mm) eluted at 1 ml/min and monitored at 215 nm. The column equilibrated with 100% buffer A (0.1% TFA) was then brought to 70% buffer B (0.085% TFA in CH_3CN) at 90 min.

Although the 39 participating laboratories submitted 27 PVDF and 30 in-gel data sets, only 20 PVDF and 22 in-gel data sets could be quantified based either on the sequencing yields reported for the target peptide and/or on the submitted chromatograms. In the remaining cases, the target peptide was not sequenced, or the chromatograms could not be quantified due to off-scale peaks, or the absence of a chromatogram corresponding to the external peptide standard and/or the blank digest. To identify factors that might account for the wide range in results obtained in this study (see Figure 3-4), several potentially important parameters are summarized in Table II for the 30% of the PVDF data sets that identified the most residues in the target peptide (i.e., the 6 above average data sets) versus the 30% of data sets that had the lowest composite chromatography scores (i.e., the 6 below average data sets). The most significant difference between these two data sets would seem to be the apparent lack of a blocking detergent in 3 of the 6 below average data sets (Table II). In the absence of a detergent such as Triton X-100 it is likely the added protease would be lost due to adsorption onto the PVDF membrane thus accounting for the failure of these digests. Other factors that probably contributed to the relative success of the above average data sets were the more favorable HPLC conditions that were used (smaller column ID and lower flow rates), their slightly higher level of routine experience in carrying out in situ PVDF digests, and particularly important, the fact that these laboratories routinely digest an approximately 3-fold lower range of protein than that for the below average data sets (Table II). In contrast to these parameters, since there was a 10-fold range in the total PVDF wash volume among the above average data sets it is unlikely the generally larger wash volumes used by these laboratories contributed significantly to their success.

Table II. Comparative data from above and below average PVDF digests*

Description	6 Above Average Sets			6 Below Average Sets		
	n	Median	Range	n	Median	Range
Relative peak height	4	4.6	3.3-11	6	1.1	0.12-2.2
Number of 30 kD peaks	4	14	10-15	6	2.5	0-8
Number of background peaks	4	6.0	3.0-21	6	7.0	2.0-3.5
Composite chromatography score	4	1.2	1.0-1.3	6	0.11	-0.9-0.6
Number of residues sequenced	6	14	13-15	6	0	0
Val4 sequencing yield in target peptide (pmol)	5	3.0	1.8-5.7	6	-	
Total PVDF wash volume (ml)	6	0.5	0.2-2.0	6	0.2	0.05-0.6
Triton X-100 was used (%)	6	100		6	50	-
Column ID (mm)	6	1.6	0.5-2.1	6	2.1	0.8-2.1
Column flow rate (µl/min.)	6	120	20-200	6	200	17-500
Routinely perform PVDF digests (%)	6	83		6	67	
Quantity of protein routinely digested						
Minimum (pmol)	3	20	10-25	6	59	20-270
Maximum (pmol)	3	70	50-80	3	200	130-500

*See Materials and Methods for details concerning calculation of relative peak height and other data summarized above.

Table III summarizes a similar comparison of above and below average in-gel data sets. In this instance there is stronger positive correlation between several factors that probably contributed to the success of the laboratories that submitted the above average datasets. These include their use of smaller ID columns, lower flow rates and their significantly higher level of experience in carrying out in-gel digests at a level (i.e., 10-50 pmol) at or below that at which this study was carried out (i.e., in this study 70 pmol each of the 28 and 30 kD samples were subjected to SDS PAGE). Although the absence of Tween 20 or other detergent in the digest buffer and the use of trypsin from a particular vendor seemed to correlate with the above average data sets, the small sample size and the use of only a single protein requires that additional studies be carried out to determine if this correlation is significant.

Success in this study required both that the sample be digested and then fractionated via HPLC, hence, problems may occur during either or both of these procedures. Since a blocking detergent was apparently not included in the PVDF digest shown in Fig. 4A, this below average chromatogram may have resulted from a failed digest. The high background in this chromatogram may be due to use of a 10-fold higher amount of trypsin (5 μ g as opposed to the 0.2-0.5 μ g range used by the other laboratories that submitted PVDF digests) that had not been modified to minimize autolysis. In contrast, while there does not appear to be any obvious reasons (based on the data provided in the accompanying sample sheets) why the digest shown in Fig. 4B failed, the conditions under which this HPLC chromatogram was carried out were not optimum (flow rate of 1 ml/min on a 4.6 mm ID column). In comparison, the median flow rates used for the above average PVDF and in-gel data sets were

Table III. Comparative data from above and below average in-gel digests[†]

Description	7 Above Average Sets			7 Below Average Sets		
	n	Median	Range	n	Median	Range
Relative peak height	3	4.1	3.2-4.6	7	1.6	0-2.1
Number of 30 kD peaks	3	16	12-17	7	4.0	0-7.0
Number of background peaks	3	0	0-7.0	7	12	4.0-16
Composite chromatography score	3	1.6	1.0-1.6	7	0.13	-0.34-0.55
Number of residues sequenced	7	11	7-15	7	0	0
Val4 sequencing yield in target peptide (pmol)	7	3.3	1.2-4.2	7	-	-
Detergent was used (%)	7	14		7	71	-
Promega trypsin used for digest (%)	7	100		7	29	
Column ID (mm)	7	1.0	0.5-2.1	7	2.1	2.0-2.1
Column flow rate (μ l/min)	7	50	20-200	7	200	150-500
Routinely perform gel digests (%)	7	100		7	33	-
Quantity of protein routinely digested						
Minimum (pmol)	5	20	10-50	5	100	50-10000
Maximum (pmol)	6	100	20-200	2	570	130-1000

[†]See Materials and Methods for details concerning calculation of relative peak height and other data summarized above.

120 and 50 μ l/min respectively on either 1.0 or 2.1 mm columns, with the latter conditions providing as much as a 20-fold possible increase in detection sensitivity (assuming similar flow cell path lengths, monitoring wavelengths and proportional peak volumes) compared to the chromatograms shown in Fig. 4B.

Finally, one of the goals of this study was to compare the relative effectiveness of *in situ* PVDF versus *in-gel* digests. Since equal numbers of both types of digests were submitted (27 PVDF and 30 *in-gel*), there does not appear to be a clear consensus in terms of the best way to proceed. Indeed, this supposition is supported by the data in Table IV where, with the possible exception of a small increase in the number of background peaks observed in the *in-gel* samples, no significant difference was seen in the quality of either the HPLC chromatograms or the accompanying sequencing data for the PVDF versus *in-gel* digests. With regards to sequencing yield, it is interesting to note that from the amount of protein loaded onto the gel (about 50 pmol after correcting for the 20-33% of 30 kD band shifted into the 28kD region - see Section IIB) and the overall median initial sequencing yield of 3.2 pmol (based on the 2.3 pmol median yield of Val4 corrected to cycle 1 using a repetitive yield of 90%), the median overall recovery is about 6.4% (range= 3% to 16%). Assuming an average coupling yield of 50%, this corresponds to an actual average recovery of about 12.8%, which is a figure that must be kept in mind in terms of deciding the amount of protein that must be submitted for internal sequencing to ensure a reasonable probability of success. Overall, approximately 63% of the chromatograms appeared to be reproducible and the median mass determined by 11 laboratories for the target peptide was 1895.60, which compares with the predicted (average) mass of 1895.06. The median mass error was $\pm 0.028\%$.

Table IV. Comparative data from PVDF and *in-gel* digests*

Description	PVDF			Gel		
	n	Median	Range	n	Median	Range
Relative peak height	20	2.4	0-14	18	2.6	0-5.8
Number of 30kD peaks	20	8.0	0-15	18	8.5	0-17
Number of background peaks	21	4.0	0-35	19	7.0	0-22
Number of residues sequenced	9	13	6-15	7	13	7-15
Val4 sequencing yield in target peptide (pmol)	8	2.5	1.0-5.7	6	2.3	1.2-4.2

*See Materials and Methods for details concerning calculation of relative peak height and other data summarized above.

IV. Conclusions

Based on results obtained on the recombinant 30kD protein, the submitted data sets argue persuasively that there is no significant difference in the overall effectiveness of PVDF and in-gel approaches to internal sequencing. Rather, the choice between these two approaches would seem to rest largely with personal preference and perhaps to some extent with other factors specific to the protein being studied (such as difficulties in obtaining near quantitative blotting efficiency or, particularly in the case of low molecular weight proteins, unusually high losses during washing of SDS polyacrylamide gel slices prior to in situ digestion). Clearly, Figures 1-3 and Tables II and III demonstrate that excellent results were obtained by several laboratories using either of these approaches. In this regard the PVDF and in-gel data sets submitted by laboratory #3 were particularly noteworthy in that in both instances all 15 residues in the target peptide were correctly identified. Possible reasons for less than optimal results range from apparent methodological errors and sub-optimal HPLC conditions discussed above to the potentially more interesting finding that the presence of a detergent may contribute to a less than optimum in-gel digest. In terms of the overall success rate in this study, 1% of the laboratories that participated (i.e., 20 out of a total of 39) either obtained 6 or more residues of sequence from the unique target peptide or sufficient other internal sequence to identify the parent protein as being derived from spectrin. In one instance, this identification was also made based on peptide mass

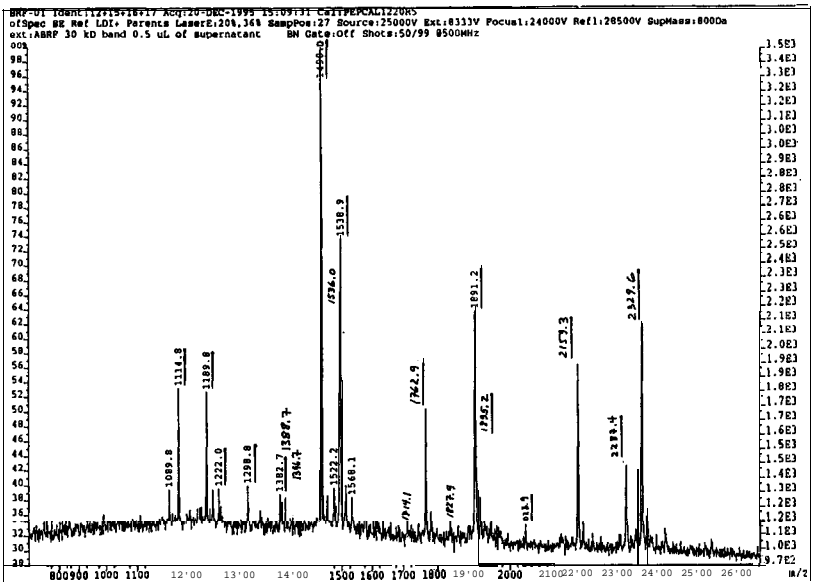


Figure 5. MALDI-MS peptide mass map of 1.7% of the 30 kD in-gel digest as submitted by laboratory 2 1. The resulting data base search matched the underlined masses to the p-chain of human spectrin.

searching of 1.7% of the digest (Fig.5) Since the cover letter that accompanied the samples clearly stated that going beyond the requested digest and analytical HPLC was optional, the actual fraction of participating laboratories that **could** have succeeded with this sample was surely above the 5 1% figure. In this regard, it is also interesting to note again that 24% of the participants in this study (Table I) do not routinely carry out in-gel on-PVDF membrane digests. In fact, in some instances this study apparently represented the laboratory's first attempt at carrying out either of these digests- thus fulfilling one of the intended goals of this first ABRF collaborative research study devoted to internal sequencing of SDS PAGE- separated proteins.

Acknowledgements

This work was partially supported by DOE grant number DE-FGO2-9SER61839 to John Crabb (W. Alton Jones Cell Science Center) on behalf of the ABRF. We especially thank the 39 laboratories that made the substantial commitment necessary to participate in this study. The assistance of Robert Tanis (Harvard Medical School) in coordinating data return and ensuring the anonymity of the participating laboratories is appreciated. We also thank Sandra Harper (Wistar Institute) for constructing the expression vector for the 30 kD protein as well as for expressing and purifying the two proteins used in the study. Several members of the authors' laboratories also contributed to the preparation and evaluation of the samples used in this study, especially: Kathy Stone (Yale University), Nora E. Poppito (Cold Spring Harbor Laboratory), Renee A. Robinson (Harvard University), Joseph Fernandez (Rockefeller University), and David Reim (Wistar Institute).

References

1. Brown, J.L. and Roberts, W.K. (1976) *J. Biol. Chem.* **251**,1009-1014.
2. Guo, D., Mant, C.T., Taneja, A.K., Parker, J.R., and Hodges, R.S. (1986) *J. Chrom.* **359**, 499-517.
3. Fernandez, J., DeMott, M., Atherton, D., and Mische, S.M. (1992) *Anal. Biochem.* **201**, 255-264.
4. Rosenfeld, J., Capdevielle, J., Guillemot, J.C., and Ferrara, P. (1992) *Anal. Biochem.* **203**, 173-179.
5. Hellman, U., Wemstedt, C., Gonen, J., and Heldin, C.-H. (1995) *Anal. Biochem.* **224**, 451-455.
6. Driessen, H.P.C., de Jong, W. W., Tesser, G.I., and Bloemendal, H. (1985) *In* Critical Reviews in Biochemistry (G.D. Fasman, ed.) 281-325.