

Evaluation of Protein Sequencing Core Facilities: Design, Characterization, and Results from a Test Sample (ABRF-9ISEQ)

Dan L. Crimmins¹, Gregory A. Grant², Liane M. Mende-Mueller³, Ronald L. Niece⁴,
Clive Slaughter⁵, David W. Speicher⁶, and K. Ümit Yüksel⁷

1 Howard Hughes Medical Institute Core Protein/Peptide Facility, Washington University School of Medicine; St. Louis, MO 63110

2 Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine; St. Louis, MO 63110

3 Protein/Nucleic Acid Shared Facility, Medical College of Wisconsin; Milwaukee, WI 53226

4 Protein/DNA Sequence/Synthesis Facility, University of Wisconsin Biotechnology Center; Madison, WI 53705

5 Howard Hughes Medical Institute, University of Texas Southwest Medical Center; Dallas, TX 75235

6 Protein Microchemistry Facility, Wistar Institute; Philadelphia, PA 19014

7 Biopolymer Analysis Laboratory, Texas College of Osteopathic Medicine; Fort Worth, TX 76107

I. Introduction

Protein N-terminal sequence analysis has continued to play a major role in the elucidation of complex biological research issues. This sophisticated procedure is currently used to analyze a wide spectrum of samples, including: quality control of recombinant proteins and synthetic peptides; purity, identification, and evolutionary assessment of isolated proteins; analysis of protein fragments for the purpose of producing oligonucleotide probes and generating antibodies via the corresponding synthetic peptide; and, in favorable circumstances, identification of modified residues. Protein chemistry laboratories, as a result, are faced with challenging sample analyses on an ever increasing basis. Clearly, not every protein sequencing facility will possess the same degree of expertise and instrumentation as others. This has tended to promote unrealistic and uninformed expectations from users of these facilities.

A continuing effort to inform investigators about the average capabilities of protein sequencing core laboratories has been mounted by the Association of

Biomolecular Resource Facilities (ABRF) sequencing committee for the past several years. Samples with both specific design goals and well-characterized components were distributed to ABRF-member core facilities world-wide. The analysis of these data for STD-1 [1], ABRF-89SEQ [2], and ABRF90SEQ [3] have provided valuable information to users regarding the performance of the average facility. Furthermore, these studies allow the participating facilities to objectively evaluate their strengths and weaknesses. In continuation of this effort, a test sample designated ABRF-91SEQ was distributed to 198 ABRF-member protein sequencing core facilities and this report describes the results from the analysis of the first 90 sets of data returned by May 15, 1991.

II. Materials and Methods

A. Design of the ABRF-91SEQ Test Sample

The test sample, designated ABRF-91SEQ, was designed to possess similarities to the previous years sample, ABRF-90SEQ, but with several significant differences. The intention being to determine whether or not those differences had a practical effect on the sequencability of the sample. In 1990, the 29 residue synthetic peptide of ABRF-90SEQ was covalently coupled to transferrin, mixed with glucose-6-phosphate isomerase, and blotted onto polyvinylidenedifluoride (PVDF) membrane [3]. Since glucose-6-phosphate isomerase is naturally blocked and transferrin becomes mostly blocked during the first step of the coupling procedure, this produced a sample that would mimic a large molecular weight protein in both solubility properties and background. The present sample, ABRF-91SEQ contained all of the same components except that the peptide was not covalently linked to the transferrin and the sample was supplied in a form to be dissolved for loading. The transferrin was left unblocked in this case but was present in the sample at a level comparable to the level of unblocked transferrin in ABRF-90SEQ.

The two proteins, human transferrin (75.2kDa) and yeast glucose-&phosphate isomerase (E.C. 5.3.1.9; 2x63 kDa; naturally blocked N-terminus) were added at 0.3 μ g (4 pmol) and 4.93 μ g (38 pmol) respectively, to 0.27 μ g (80 pmol) of a 29-residue synthetic peptide. The synthetic peptide itself (Fig. 1) possessed several features in common with the synthetic peptide portion of ABRF-90SEQ [3]. These include (i) a 29-residue length with identical overall amino acid composition; (ii) sequence identity at positions Trp-6, Leu-10, Arg-11, Pro-14, Ser-25, Phe-27, and Lys-28; (iii) an Asn-8:Asp-9 pair to evaluate possible effects of deamidation on sequence assignment; and (iv) a Leu-3:Leu-7:Leu-10 triplet early in the sequence to assess repetitive yield. Several new features were also incorporated into the synthetic peptide portion of ABRF-91SEQ. Glycine was chosen as the N-terminal residue to test the operator's ability to correctly assign this frequently occurring contaminating residue, particularly since the second amino acid, Arg, of ABRF-91SEQ is typically a "low-yield" residue. An unmodified cysteine was placed early in the sequence, Cys-5, but before any authentic Ser residues, to ascertain if this difficult residue could be tentatively inferred from the potentially diagnostic phenylthiohydantoin (PTH) adduct peaks. Lastly, a second potential deamidation pair, Gln-15:Glu-16, was incorporated into the sequence.

5
10
15
20
25
G-R-L-K-C-W-L-N-D-L-R-S-Y-P-Q-E-M-T-H-V-A-S-I-D-S-V-F-K-V

Figure 1. Amino Acid Sequence of ABRF-91SEQ .

B. Peptide synthesis and characterization

The 29-residue synthetic peptide was synthesized with t-Boc chemistry on an Applied Biosystems Model 430A Synthesizer by the Protein Chemistry Laboratory at the Washington University School of Medicine and HF-cleaved from the resin using established procedures as previously described[2,3]. The crude product was subjected to preparative reversed-phase high-performance liquid chromatography(HPLC) and the purified peptide distributed to several of the authors for N-terminal sequence and compositional analysis. The peptide was judged to be pure by sequence analysis at sample loads of 50-200 pmol with initial yield(%) of 54.7 ± 21.3 (average \pm sample standard deviation, seven independent determinations) with a range of 23 to 85%. Two independent Fast Atom Bombardment Mass Spectrometry determinations were performed on the putatively pure peptide yielding a m/z value of 3393.0 and 3393.4 (calculated m/z = 3394.96) as the major component (-90%). A minor component at 1106.0 (-10%) was observed and is consistent with an N-terminal acetylated α -mer comprising residues. 20-29. Two peaks consistent with this ratio and which eluted/migrated in the expected order for these species were also observed when the sample was analyzed by both analytical reversed-phase and strong-cation exchange HPLC, and free-solution capillary electrophoresis. This N-terminally blocked contaminant should not interfere with the sequence analysis, and was therefore not removed from the intended sequencable target peptide.

C. Preparation and distribution of the test sample

The three components were mixed in the proportions described above, aliquotted into pre-washed [1,2] 1.5 ml polypropylene microfuge tubes, and then vacuum dried. These samples were mailed to 198 core facility members of the ABRF along with a brief questionnaire and a detailed set of instructions for sample solubilization, loading, and data reporting. Data were returned to an independent third party who removed any identifying marks. These anonymous data were then forwarded to the authors for analysis.

III. Results and Discussion

A. Survey Responses of the Participating Facilities

The number of facilities participating in this year's study was about 1.6-fold greater (90 vs. 56) than in last year's study, yet the response rate was similar, 45.5% (90/198) compared to 42.4% (56/132) [3]. A short survey was included in the report with 86 of the 90 respondents completing this part of the questionnaire. There was no significant change in the average number of sequencers/facility (1.56, range 1-6) compared to the previous year [3] although the average age of the sequencers used to analyze the ABRF-91SEQ sample did decline to 4.2 years (range 0.8-9). Most of the facilities (78/86) have Applied Biosystems (ABI) sequencers(470-17,473-6,475-14, 477-41) with the remaining eight responses distributed between Porton (2090-5), Milligen (6625-2), and Beckman (890-1) sequencers. Approximately two-thirds (67.9%) of the sequencing laboratories purchase all of their sequencer/analyzer supplies from the manufacturer of their instrument(s).

The overwhelming majority of facilities (84/86) use on-line HPLC PTH analysis, mainly on ABI 120 instruments(77/85). Most perform the chromatography on a C18 column from ABI/Brownlee (72/85) with a 2.1 mm internal diameter column being most popular. The amount of sample injected onto the HPLC fell into three

ranked groups of 50 μL (36/81), 100 μL (25/81), and 80 μL (7/81) [overall average of 54.8 \pm 19.0 percent injected]. Not surprisingly, since the number of participating facilities increased so did the total number of samples analyzed in these laboratories (total 22,264; average 174, range 4-800/instrument; average 282, range 4-3000/facility). Both of these averages are down about 17% from 1990 [3]. Outside service to other investigators accounts for the bulk of the workload (66%) at the protein sequencing core facilities. The average site performs 25.6 (range 1-250) standard **protein/peptide** analyses per year per facility to routinely validate and trouble-shoot instrument performance. This translates to an average of one standard analysis per eleven sequencing runs.

Sequencing yield in each cycle from the ABRF-9ISEQ sample was quantified after correcting for the fraction of injected sample (71/81) in units of picomoles (80/81) and in conjunction with an average loading of 49.8 pmoles of PTH amino acid standard but without the use of an internal standard (66/80). The sample was apparently loaded and analyzed as requested using standard cycles (74/86), on a glass-fiber filter (74/86) that had been treated with polybrene and pie-cycle (63/86). Only 3.7% of the facilities monitor absorbance at 313 nm, 6.2% utilize serine cycles, 56.3% do not utilize proline cycles although 32.5% do so 5 percent of the time, and 29.4% incorporate OPA blocking in their sequencing strategy. Virtually all laboratories sequence samples that have been electroblotted from one-dimensional gels (94%, 78/83) and almost one-half (46.7%, 35/75) analyze samples electroblotted from two-dimensional gels. The widespread success of the sequencing, and the relative ease of preparing these immobilized samples [4] is a likely factor in the near doubling over 1990 [3] of the facilities that offer electroblotting as a service (49% vs. 26.4%). Interestingly, this increase has occurred despite little change in the number of facilities that performed electroblotting this year (63.9%) compared to those who responded to the ABRF-9OSEQ questionnaire (56%) [3].

B. Sequence Results for ABRF-9ISEQ

A total of 1906 sequencing cycles were reported for the analysis of ABRF-91 SEQ (Fig. 2) with an average of 22 cycles/report (range 1-30). There were 1421 (74.6%) positive and 200 (10.5%) tentative assignments, with 285 (15%) unassigned cycles ("holes"). Accuracy was 83.0% for positive assignments and 55.0% for tentative assignments. Thus, 1291 (67.7%) of the 1906 cycles were correctly identified. These values were all within a few percent of those reported for ABRF-9OSEQ [3]. During this study, 9 (10%) instances of instrument malfunction were reported as compared to 8 (14.5%) for ABRF-9OSEQ [3]. These included single instances of R2 delivery malfunction, computer failure, vacuum failure, and injection failure that terminated the run. There were also two undefined malfunctions and two injector malfunctions that did not result in run termination, and one fraction collector failure, on an instrument that did not employ on-line PTH-amino acid injection, that terminated the run.

Leu-3 was the most correctly identified residue overall (76 correct and 2 wrong positive; 3 correct and 2 wrong tentative; 0 holes), followed closely by Lys-4. Within the first 22 cycles, which was the average number reported, Cys-5 was the most difficult to identify as expected (6 positive and 5 tentative correct out of a total of 83 attempts). Position 5 was most often misidentified as threonine (29 positive and 9 tentative assignments) which was probably due to threonine being at position 5 in transferrin, although transferrin was present at only 5% of the synthetic **peptide** present in the sample. No other apparent misassignments related to the transferrin sequence were made although several responses correctly identified this as a minor secondary sequence.

Results of the sequence assignments for ABRF-9ISEQ.

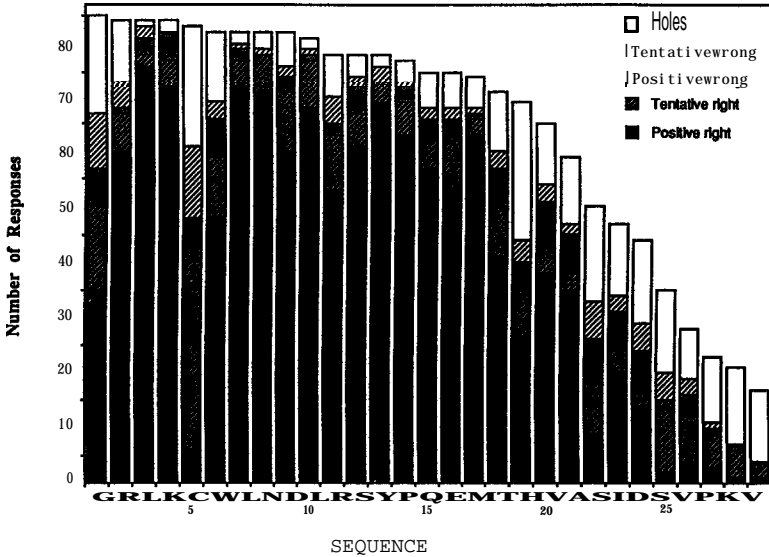


Figure 2. Results of Sequence Assignments of ABRF-9ISEQ.

It was assumed that the data reported on the returned sequence data sheets were operator assigned sequence unless otherwise specified by the respondent. Scoring of residue assignments, exemplified by Cys-5, were as follows: C or C, (T) = positive correct (C) or (C), (T) or (C,T) = tentative correct; T or T,C or T, (C) = positive wrong; and (T) or (T), (C) or (T,C) = tentative wrong. An "X", "?", "-", or greater than two residues being reported at a cycle was considered a "hole".

The single tryptophan residue, which was present at position 6 in both ABRF-90SEQ and ABRF-9ISEQ, was identified with much greater success this year with 48 positive and 7 tentative correct (55/81,67.9%) as opposed to 13 positive and 4 tentative correct last year (17/55,30.9%) [3]. The reason for this greater than two-fold increase in Trp identification is not apparent, but may be attributable to adsorption to PVDF or an intrinsic recovery problem of Trp from PVDF.

Of the 3 additional residues that were present at identical positions within the first 22 residues in ABRF-90SEQ and ABRF-9ISEQ (Leu-10, Arg-11, and Pro-14) only **Arg-11** showed a significant difference in identification (70.5% positive and tentative correct out of 78 responses this year as compared to 35.4% out of 48 responses for ABRF-90SEQ [3]).

While the first residue is often difficult to identify because there is no previous cycle for comparison, the glycine at position 1 in ABRF-9ISEQ was the least correctly identified N-terminus of all previous ABRF samples[1-3]. This may be due to the fact that glycine is often seen as a substantial contaminant at cycle 1 in experimental samples and may be dismissed as a valid sequence residue. Indeed, this was in fact the reason why glycine was placed at this position in ABRF-9ISEQ and the results indicate that the coincidence of glycine contamination and its presence as the N-terminal residue present serious problems in assignment.

Table I. Sequencing results for ABRF-91SEQ (a)

Instrument		Repetitive Yield (d)	Number of Positive correct	Accuracy of Positive Calls (e)
Model (b)	# (c)			
470/120	17 [12]	87±/4.9 (79-96)	15±/6.5 (1-23)	83±/31.0 (4-100)
475/120	14 [10]	87±/5.5 (77-95)	11±/8.0 (0-22)	65±39.4 (0-100)
477/120	41 [37]	81±/10.0 (49-96)	14±/7.0 (1-27)	77±/28.1 (17-100)
473/120	6 [5]	75±/11.0 (55-81)	17±/6.2 (9-24)	87±/11.3 (75-100)
2090	5 [4]	86±/5.9 (77-90)	12±7.8 (0-21)	78±/43.5 (0-100)
6625	2 [1]	87	10 (1-19)	65 (50-79)
890	1 [0]	.	6	35
All Instruments	86 [69]	83±/9.0 (49-96)	14±/6.7 (0-27)	83±/21.8 (0-100)

(a) Values are expressed as mean ± standard deviation, with ranges in parentheses.

(b) 2090E is a Porton sequencer with a Hewlett Packard 1090 HPLC; 6625 is a Milligen 6600 sequencer with a Waters 625E HPLC; 890 is a Beckman 890M-2 sequencer with a Beckman 344 HPLC and all other instruments are from Applied Biosystems.

(c) Number of instruments for which data was reported. Numbers in brackets are those returns used for repetitive yield calculations.

(d) Expressed as percent. Based on background corrected yield of Leu-3 and 7.

(e) Expressed as percent. Based on correct calls for positively identified residues only. Cycles that were left blank or had tentative assignments were not scored.

Potential deamidation, as judged by identification of the Asn-Asp pair at positions 8 and 9, did not seem to be a problem. The Gln-Glu pair at positions 15 and 16 was consistent with this observation.

The average response reported identifications for 22 cycles including tentative assignments with an average of 14 positive correct assignments (see Table I). In comparison, eight responses (477-6, 473-1, and 470-1) identified at least 22 residues positively correct. Two of the longest and better responses reported 25 positive correct out of 28 calls, with 2 positive incorrect and 1 hole while another response reported 24 positive correct calls out of 25 attempts.

Analysis of the data for repetitive yield and accuracy were grouped by instrument type and are presented in Table I. Statistically there does not appear to be any significant difference in average repetitive yield, average number of positive correct, and average accuracy of positive calls between instruments.

Table II. Comparison of ABRF-91SEQ to Previous Test Samples (a)

sample(b)	Description	Amount (pmol)	Repetitive Yield %	Number of Correct Res.	Accuracy % (c)
STD-1(1988)	single soluble peptide, 40 aa	100	88k3.9 (76-93)	24+/-7.1 (9-37)	95+/-7.0 (65-100)
ABRF-89SEQ	soluble peptides (5:1) 40 & 43 aa	240	89k2.4 (77-99)	30+/-6.5 (16-38)	95±6.6 (80-100)
ABRF-90SEQ	PVDF blot mimic 100 kDa protein, 29 aa	30	89+/-3.7 (82-95)	13+/-6.4 (1-27)	82+/-19.1 (36-100)
ABRF-91SEQ	soluble peptide, 29 aa, 100 kDa protein	80	83+/-9.0 (49-96)	14+/-6.7 (0-27)	83+/-21.8 (0-100)

(a) Values are expressed as mean \pm standard deviation, with ranges in parentheses.

(b) Data for STD-1 ABRF-89SEQ and ABRF-90SEQ are from references [1] [2] and [3], respectively.

(c) Percentage of positive sequence assignments that were correct

In previous ABRF sequencing studies, a comparison of the accuracy of automated, software assisted calls with manual, operator assignments indicated that the experienced sequencer operator is more accurate [1-3]. This year a similar analysis resulted in 54.0+/-29.0 percent accuracy for those respondents that reported software derived sequence assignments compared to a value of 83.0+/-21.8 percent accuracy (Table I) obtained from all responses to ABRF-91SEQ. These numbers and their differences are virtually identical to those reported for ABRF-90SEQ [3] (58.0 and 82.0 percent accuracy, respectively) and once again this demonstrates that more sophisticated versions of the software will likely be required to increase the accuracy of the automated calls. It should be pointed out however, that most versions of the current software, by default, make positive sequence assignments at most of the cycles. It is not surprising then, that these positive cycle assignments, on average, will statistically contain more positive wrong calls compared to an operator who can provide tentative cycle assignments.

Table II compares the results of ABRF-91SEQ to previous samples. Although the average repetitive yield for ABRF-91SEQ appears to be lower than in previous years, the range is also broader and there is some question as to whether this difference is significant. More importantly, perhaps, the table clearly indicates that the average number of positive correct and average percent accuracy are essentially identical for ABRF-91SEQ and ABRF-90SEQ. As discussed in the section on sample design, ABRF-91SEQ was designed as a soluble sample for comparison to ABRF-90SEQ to evaluate any differences in these parameters due to presenting the sample on PVDF.

The actual amount of sequencable peptide in both samples was approximately equivalent. Since it was not feasible to directly quantitate the amount of peptide present on the PVDF blot for ABRF-90SEQ, the 30 pmol so indicated in the table reflected the average initial yield produced by the samples. Initial experiments by the sequencing committee indicated that for an 80 pmol load of the peptide in ABRF-91SEQ, the average initial yield was 54.7+/-21.3% (range 23-85%) corresponding to an average initial yield of 44 pmol. This was verified by the data from the responses to ABRF-

9ISEQ which indicated an average initial yield of 36 ± 18.6 pmol (range 3-89 pmol). Thus, the average available sample for sequencing was similar for both ABRF-9OSEQ and ABRF-9ISEQ. Therefore, the values for these two samples in Table II are directly comparable and indicate that the number of identifiable residues and the accuracy of those identifications did not appear to be a function of immobilization onto PVDF or cross-linking to a carrier protein. In our view, this is not meant to indicate that there is no advantage to immobilizing samples on to PVDF, but rather, once equivalent amounts of sample are delivered to the sequencer, the analysis proceeds equally well in both cases. Indeed, many reports and individual experiences indicate that PVDF is an excellent medium for capturing and transferring small quantities of protein [4].

In general, considering the results from the four years that the ABRF has been distributing samples, it appears that the main factor in determining the number of residues that can be sequenced and the accuracy of that sequencing, is the total amount of sample delivered to the sequencer. Therefore, the results of **ABRF-91** SEQ are consistent with and further support the conclusions reached previously [5].

It must be pointed out that the data presented here represent the average for all facilities that reported data. This includes returns not only from experienced laboratories but also from laboratories that are just starting out and are still learning the techniques and art of sequencing. Moreover, these results are based on a one time analysis of a sample with little background information provided. Also, because it is an outside test sample, some operators might have been a little more daring in attempting to make positive calls than with actual experimental samples. As such, the results of this study should be viewed conservatively. The average results presented here, particularly the % correct positive calls, should not be regarded as representing an acceptable level of performance for a facility on a daily basis. On the contrary, facility personnel should strive to achieve the performance level of the best laboratories reported. There were 19 responses with 100 % accuracy, and 26 responses with 90-99.9 % accuracy. An experienced investigator will be very careful in assigning positive calls and will not hesitate to categorize a call as tentative if there is any doubt. Certainly, an accuracy of 90 % or better is desirable and realistic since 45 out of the 90 respondents achieved this level.

Acknowledgements

This work was partially supported by NSF grant DIR 9003100 to John Crabb (W. Alton Jones Cell Science Center) on behalf of the ABRF. We are grateful to the anonymous ABRF member facilities that participated in this study without whose interest and effort this project would not have been possible. The assistance of Melanie Budd (University of Wisconsin, Biotechnology Center) in coordinating the data return and providing the anonymity of the respondents is appreciated. Several members of the authors' laboratories contributed to various aspects of this project, including: Mark Frazier (Washington University School of Medicine, Protein Chemistry Laboratory) for synthesizing the peptide; Richard Thoma for amino acid analysis of the final test sample and David McCourt (Washington University School of Medicine, Howard Hughes Medical Institute) for initial sequence analysis; Brady Stoner and Rodney Prell (Medical College of Wisconsin) for preparation of the test sample; Jean Shelton and Peggy Vroman (Texas College of Osteopathic Medicine) for data entry; Carolyn Moomaw (University of Texas Southwest Medical Center, Howard Hughes Medical Institute) for sequence analysis of the test sample; and Lorraine Whiteley and Pat Parvin (Washington University School of Medicine, Howard Hughes Medical Institute) for expert secretarial assistance. Their efforts are gratefully acknowledged.

References

1. Niece, R.L., Williams, K.R., Wadsworth, C.L., Elliot, J., Stone, K.L., McMurray, **W.J.**, Fowler, A., Atherton, D., Kutny, R., & Smith, A.J. (1989) *in* Techniques in Protein Chemistry, T.E. Hugli, ed., pp 89-101 Academic Press San Diego.
2. Speicher, D.W., Grant, G.A., Niece, R.L., Blacher, R.W., Fowler, A.V., & Williams, K.R. (1990) *in* Current Research in Protein Chemistry, J.J. Villafranca, ed., pp 159-166 Academic Press, San Diego.
3. Yüksel, K.Ü., Grant, G.A., Mende-Mueller, L.M., Niece, R.L., Williams, K.R. & Speicher, D.W. (1991) *in* Techniques in Protein Chemistry II, J.J. Villafranca, ed., pp 151-162 Academic Press, San Diego.
4. LeGendre, N. (1990) Biotechniques 9 (6), 788-805.
5. Niece, R.L., Ericsson, L.H., Fowler, A.V., Smith, A.J., Speicher, D.W., Crabb, J.W., & Williams, K.R. (1991) *in* Methods of Protein Sequence Analysis, H. Jomvall, J.-O. Hoog, A.-M. Gustavsson, eds., pp 133-141 Birkhauser Verlag, Basel.