

Design, Characterization and Results of ABRF-89SEQ: A Test **Sample** For Evaluating Protein Sequencer Performance in Protein Microchemistry Core Facilities

David W. Speicher¹, Gregory A. Grant², Ronald L. Niece³, Russell W. Blacher⁴,
Audree V. Fowler⁵, and Kenneth R. Williams⁶

¹Wistar Institute
Protein Microchemistry Facility
Philadelphia, PA 19104

⁴Department of Biochemistry and Molecular Biophysics
Washington University School of Medicine
St. Louis, MO 63110

⁶University of Wisconsin Biotechnology Center
Protein/DNA Sequence/Synthesis Facility
Madison, WI 53705

⁵Athena Neurosciences, Inc.
South San Francisco, CA 94080

²Department of Biological Chemistry
UCLA School of Medicine
Los Angeles, CA 90024

³Protein and Nucleic Acid Chemistry Facility
Yale University School of Medicine
New Haven, CT 06510

I. INTRODUCTION

Over the past decade, protein sequence analysis has benefited from a number of substantial advances that have dramatically increased the level of sensitivity that can be achieved. Despite these improvements in performance, results obtained on unknowns are not necessarily on the same level as optimistic claims based on idealized protein standards.

Much of the current protein sequencing capacity is located in core facilities which serve a large number of scientists with diverse backgrounds. A core facility is generally the preferred organizational approach due to the high cost

of equipping and maintaining a facility as well as the considerable professional staff required to run at peak efficiency. In addition, improvements in instrumentation, including advances in computer software, have not diminished the requirement for a high level of expertise and experience. The current complexity and sophistication of available instruments is analogous to an intricate musical instrument - if not carefully tuned, the result is often not very pleasing.

Two of the goals of the sequencing subcommittee of the Association of Biomolecular Resource Facilities (ABRF) are to define the average capability of sequence core facilities and to help improve operational performance. These objectives are expected to benefit both facility operators and users. Last year a purified synthetic peptide, STD-1, was released to 103 core facilities as an "unknown" test sample and the results were evaluated (1). The current sample, ABRF-89SEQ was designed to extend the results of the initial survey, and it was initially distributed to 123 core facilities that are members of the ABRF. The design of this sample and an evaluation based upon the first 50 responses is presented below.

II. MATERIALS AND METHODS

A. Design of ABRF-89SEQ

Since many experimental samples are not homogeneous, a test sample was formulated which included a minor contaminant. This was accomplished by combining a homogeneous 40 residue synthetic peptide (primary or major component) with an unrelated 43 residue synthetic peptide (secondary component or contaminant) in a 5:1 molar ratio. Both sequences (Fig. 1) were designed to ensure that under normal operating circumstances, the presence of the contaminant would not interfere with the sequence determination of the primary component. The lengths of the peptides as well as specific local sequences were designed to challenge both the capabilities of current automated protein sequencing instruments as well as the ability of operators and sequence assignment software to make accurate sequence determinations. While the sample was designed to be challenging, it was also intended to be both realistic and feasible; therefore, unusual amino acid derivatives were avoided. Aliquots sent to core facilities for analysis contained 240 picomoles of the primary component and 48 picomoles of the contaminant. At this level, assignment of most residues in the primary sequence was expected to be possible with a sequencer operating at maximum performance. Sequence assignment of the secondary component was expected to be exceptionally challenging and its assignment was optional.

A number of features were included in both peptides to assist data analysis and to test realistic problems in analysis or interpretation of PTH-derivatives. A single amino acid was spaced at regular intervals (Fig. 1) to facilitate repetitive yield evaluation throughout discrete segments of the sequence. The primary sequence also contained: aspartic acid (residue 1) to

1° - DFWYGAHARYYKMQEQQYQPSRRSVPNTLDHYPIISRRRVC
 2° - LAKCEWFSKAEEEYRPDAKMITDPTAQGINRRSAELVHERRAE

Figure 1. Sequences of major (1°) and minor (2°) components. Repetitive residues are underlined.

evaluate resolution from early eluting dithiothreitol (DTT) or ammonia derivatives; tryptophan (residue 3) to evaluate interference from diphenylurea (DPU); histidine was positioned between alanines to evaluate resolution of this pair (residues 6-8); arginine (residue 9) was placed between alanine and tyrosine to evaluate resolution of these amino acid derivatives; several histidines and arginines were also included later in the sequence to evaluate shifts in resolution of these derivatives with time; proline (residues 20, 26, and 33) to increase lag or carryover which complicates sequence assignment especially with repetitive sequences; clusters of serine and arginine (residues 21-24, 36-38) to provide sequence assignment challenges while the arginines late in the sequence were expected to minimize washout. The cysteine at the carboxyl-terminal was included for potential directed chemical coupling and **it was not expected that this residue could be assigned**. The secondary component shared many of the design features included **in the primary component. It also included several high yield residues early in the sequence to clearly establish the presence and level of this peptide, and several low yield residues early in the sequence that might be difficult to assign at a low level including cysteine, serine and tryptophan.**

Both peptides were also designed to include a number of chemical and enzymatic cleavage sites that could broaden the potential utility of these samples as potential standards for other protein chemistry methods. Each peptide has a single methionine for cyanogen bromide cleavage and the secondary component contains an asp-pro bond. The primary component has a single lysine for endoproteinase lysine-C cleavage, a single glutamic acid for V8 protease cleavage, and a single internal aspartic acid for endoproteinase asp-N cleavage. The secondary component has two internal aspartic acids for testing endoproteinase asp-N with one of the sites potentially constrained since it is preceded by a proline.

B. Synthesis and characterization of the major and minor components

The major component was synthesized in the Protein Chemistry Laboratory at the Washington University School of Medicine, St. Louis and the minor component was synthesized in the Yale University School of Medicine, Protein and Nucleic Acid Chemistry Facility. Both peptides were synthesized on Applied Biosystems Model 430A peptide synthesizers using t-Boc chemistry and PAM resins. All residues in the major component and most residues in the minor component were double coupled. After standard HF cleavage from the resin, extraction, and lyophilization, preparative HPLC was performed on Vydac C-18 columns (22 X 250 mm) with a linear gradient of acetonitrile in 0.1% TFA. Fractions were analyzed by analytical HPLC and amino acid analysis indicated the expected composition.

Both peptides were further characterized by FAB positive ion mass spectrometry. The major component contained a strong peak at an m/z of 4706.1 which is in close agreement with the predicted protonated average molecular weight of 4706.4. A secondary peak at 4728.2 in the same spectrum corresponded to the sodium adduct of the peptide. The minor component contained a strong peak at an m/z of 5033.1 which compared well with the predicted protonated average molecular weight of 5034.6. The accuracy of the sequences was verified by sequence analysis of several nmoles of the individual peptides by at least two different laboratories.

C. Preparation and distribution of the sample

The major and minor components were mixed in a 5:1 molar ratio and replicate aliquots containing 240 pmoles of the major component and 48 pmoles of the minor component were transferred to 500 μ l polypropylene microfuge tubes prewashed with 0.1% TFA, 50% acetonitrile and dried under vacuum. Samples were initially mailed as an "unknown" to 123 core facility members of ABRF. Detailed instructions for the solubilization and loading of the sample onto automated sequencers and for reporting the data were included. To guarantee the confidentiality of the resulting data, the responses were returned to third parties who removed postmarks and other identifiers prior to forwarding the data to the authors for analysis.

III. RESULTS AND DISCUSSION

A. Description of responding core facilities and sequencing throughput

The sequence report included a short survey concerning the instruments used, as well as the types of samples and sequence related services offered by respondent core facilities. This summary is based on 50 responses which included 49 completed surveys with 48 reporting sequence data.

The types of instruments used in the 49 facilities which completed the questionnaire were predominantly the three models of Applied Biosystems automated sequencers (see Table I). These 49 facilities contain a total of 76 sequencers (1.6 sequencers/facility, range 1 - 4), and a total of approximately 14,858 experimental samples are analyzed per year (303 sequences/facility; 196 sequences/instrument). Apparently the majority of facilities are using their sequencers near maximum capacity since these values represent only experimental samples; in addition these facilities devote approximately 7% (range 1 - 30%) of their sequencing effort to standards and approximately 6% (range 0 - 25%) to methods development. Responses to the range of amounts typically loaded on the sequencer were very heterogeneous. The most typical responses were approximately 10 to 200 or 300 picomoles. Of the 28 facilities indicating that a major mission of their facility included sequencing at maximum sensitivity, the lower limit specified was an average of 13 picomoles. The smallest specified lower limit was 1 picomole and the largest lower limit in this group was 50 picomoles.

Over the past two years, sequence analysis of samples electroblotted to polyvinylidene difluoride (PVDF) has quickly become a major application. Of the 47 facilities capable of using PVDF membranes, 45 (96%) use this media and more than 30% of their sequences (range 1 - 80%) are run in this format on average. Only a few laboratories use alternative blotting media and these are almost exclusively directed toward methods development. Also, 60% of these facilities perform electroblotting in their laboratory, but only 10% of the facilities offer it as a service.

B. Sequence results for the major component

The results of last year's survey (I) included two sequencer failures out of a total of 54 data sets (4%) which resulted in a partial loss of sequence information. The current data, based on 48 respondents reporting sequence

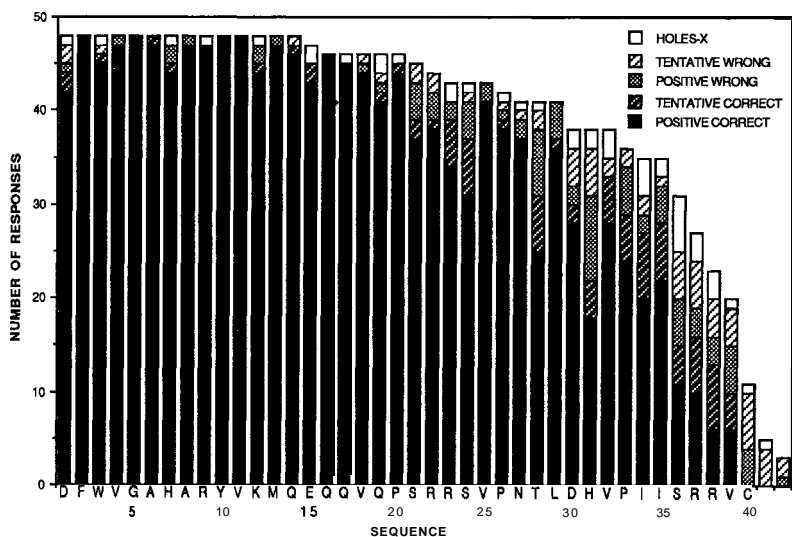


Figure 2. Distribution of sequence assignments at each cycle for the major component. The sequence of the major component is listed across the bottom with the amino-terminal on the left. If two residues were listed for a cycle it was scored as tentative (correct if one of the two was correct), while assignment of more than two residues was scored as a hole. X's and blanks were scored as holes if any assignments were made in later cycles.

data, indicates a higher rate of failures. Two facilities encountered power failures resulting in premature sequence termination; one restarted but subsequent data may have been compromised. In another case, an apparent operator-software interface problem resulted in premature termination of the sequence. Another "software bug?" was blamed for a lost injection, but a manual injection prevented any loss of data. Three other instruments encountered problems with injection onto the PTH analyzer; two cases involved a single cycle and the third case involved 10 cycles which were manually injected. Although a relatively large number of instruments encountered some problem (7/48 = 15%), the actual non-recovered data loss was quite modest since only three facilities lost more than 2 residues (the premature sequence terminations). Overall, the largest data loss occurred due to power failures suggesting that laboratories which frequently load 100% of difficult to replace samples would benefit from power backup units for the sequencer system.

A graphic representation organized by sequence position (Fig. 2) summarizes the results of the 48 facilities which reported sequence data. As expected, the number of total assignments and the percent correct calls decrease in latter parts of the sequence. Also, recognized difficult residues such as serine and threonine indicated lower accuracy levels.

The data are also grouped by instrument type in Table I. No significant differences in repetitive yield, number of correct residues, or percent accuracy were observed between the three Applied Biosystems models with on-line Model 120 HPLC's. Due to the small number of responds from off-line 470 and Beckman 890M sequencers, no firm conclusions can be drawn for these instruments. The larger amount in this year's sample, 240 pmol vs. 100 pmol for last year's sample (I), is probably the major reason why the average number of correct calls of all Applied Biosystems models has increased from 24 to 30.

TABLE I. Sequencing results for the major component of ABRF-89SEQ

<u>Model a</u>	<u>No. b</u>	<u>Repet. Yield'</u>	<u>No. Correct'</u>	<u>Accuracy (%) d</u>
477/120	16(14)	90± 1.9 (87-93)	32±4.2 (22-37)	96±5.5 (82-100)
475/120	10(10)	89±2.2 (86-92)	30±4.5 (19-37)	94±6.6 (80-100)
470/120	18(18)	88±4.3 (77-96)	29±6.5 (16-38)	94±5.3 (84-100)
470	2(2)	90± 1.1 (88-91)	25±0.0 (25)	96±0.0 (96)
890M	2(1)	96±2.8 (93-99)	13	93

Values are expressed as averages ± s.d. with ranges in parentheses.

a 890M manufactured by Beckman, others are Applied Biosystems. A Model 475 is a 470/120 updated with a 900 data analysis system. Several instruments with this configuration, but listed by respondents as 470's, were reclassified as 475's.

'Number of instruments for which data was reported and used for repetitive yield calculation, () number of data sets used for calculation of correct residues and percent accuracy - did not include three instruments where sequence was prematurely terminated as described above.

'Reported as percent. Based on background corrected yield of PTH-val at 4 and 25, or yield at 18 or 11 substituted for 25 in shorter sequences.

'Based on positively identified cycles only, cycles that were left blank or that had multiple or tentative assignments were not scored.

The overall accuracy of positive sequence calls for all Applied Biosystems instruments is 94.4% similar to the previous average of 94.6%. This -similar accuracy is especially encouraging since the current sample contained a significant contaminant and more prolines which might be expected to complicate sequence calls.

C. Comparison of automated and manual sequence calls.

The value of automated sequence assignments in contrast to manual assignments has been somewhat controversial. Last year's survey which was based on results of only five facilities and a total of 97 data system calls, indicated that automated calls were less accurate than manual calls (1).

Table II compares automated and manual calls from the 22 facilities in the current study which reported 900A data system calls as well as their manual assignments. These results show that trained operators call sequences longer and with more accuracy than the automated data systems available here. Above a picomole ratio of 20 (a software defined parameter relating the assigned amino acid quantity to its background level), the data system approaches the accuracy of a positive manual call (90.2% for data system vs. 94.6% for manual calls). However, at this stringent confidence level the data system would only correctly assign an average of 24.4 correct residues vs. 31.2 residues for manual calls by these facilities. It should also be noted that the accuracy of manual assignments for these 22 facilities is essentially the same as the average (94.4%) for all facilities with Applied Biosystems instruments.

Trained operators also have an excellent ability to discriminate between strong and weak calls. As indicated in Table II, the average accuracy of a manual positive call is 94.6% correct vs. 56.9% correct for tentative calls for the same 22 facilities which provided automated calls. Similar results are obtained when the accuracy of positive and tentative manual calls of all 48 data sets are averaged - positive calls = 94.5% and tentative calls = 56.1%. These similar accuracy averages of different data sets also strongly suggest that these sampling populations (n = 19) are large enough to eliminate sampling bias.

TABLE II. Comparison of automated and manual calls for 22 facilities

A. Automatic Calls[‡]

<u>Pmole ratio</u>	<u>Correct</u>	<u>Wrong</u>	<u>% Correct</u>
0-5	12	37	24.5
5-10	30	25	54.5
10-20	53	35	60.2
<u>>20</u>	536	58	<u>90.2</u>
Total	631	155	80.3
Avg/facility	28.7	7.0	

B. Manual Calls[‡]

Positive	686	39	94.6
<u>Tentative</u>	41	31	<u>56.9</u>
Total	727	70	91.2
Avg/facility	33.0	3.2	

[‡] Automatic calls made to the last manual call on the same data set.

[‡] Manual calls from the same 22 core facilities as automatic calls.

D. Evaluation of the secondary sequence (minor contaminant)

Although assignment of residues in the secondary sequence was optional, 39 facilities reported data for this component. An average of 11.1 correct positive calls were made (range 1 - 24). However, the average obscures the fact that a number of facilities obtained impressive data for this difficult, low level sequence. Five facilities assigned more than 20 residues correctly as positive calls and three groups made correct assignments (including tentative calls) as far as residue 35. The accuracy of positive calls (88%) was slightly lower than for the primary sequence, however many facilities probably regard an entire secondary sequence to be tentative and may not have applied their normal stringency to assignment of the secondary sequence.

E. Evaluation of common problems in manual sequence assignments

The overall results in Table II are quite good. Of the 42 data sets from facilities equipped with an on-line HPLC detection system, 84% had a better than 90% accuracy for positive manual assignments and 83% correctly called 25 or more residues. Also, when all facilities which reported data are considered, 42% (20/48) made no errors in positive manual assignments. These figures are especially impressive in view of the complexity of the sample and the presence of a minor component. Despite this 20% contaminant, not a single error in manual positive assignments arose from this minor component, indicating that all facilities are quite good at distinguishing minor sequences even when the major sequence contains a difficult residue. In contrast, the automatic data software was much worse since at least 23 assignment errors in the primary sequence were due to the secondary sequence.

The largest number of manual positive assignment errors appeared to be related, at least partially, to lag or carryover (33/83 = 40%). In a number of cases the entire sequence assignment got out of phase resulting in multiple positive errors as well as additional errors in tentative assignments. However this is not an isolated problem since at least 14 facilities (29%) had this problem. The second most frequent problem (17/83) involved incorrect assignment of one low yield residue (serine, threonine, histidine, arginine) for another low yield residue. Another substantial problem (7/83) was misidentification of histidine/alanine or tyrosine/arginine.

IV. CONCLUSIONS

Based on this study, most core facilities will have no trouble making accurate sequence assignments in the presence of a 20% molar contaminant. A laboratory equipped with an Applied Biosystems sequencer and starting with 240 picomoles of the primary component will correctly assign at least 30 residues on average. An average of about 95% of the positive calls and about 57% of the tentative calls made by an experienced operator will be correct. In addition, about 40% of all operators will make no errors in positively assigned residues. Computerized data calling can not currently be expected to perform as well as an experienced operator, at least in the average facility. Automated sequence calls result in far more errors with more holes over a comparable length of sequence.

ACKNOWLEDGEMENTS

This work was partially supported by NSF grant DIR 8903251 to K. Williams on behalf of ABRF.

The cooperation of all the anonymous laboratories that graciously contributed their time and resources to analyze the samples and provide the data requested are gratefully acknowledged. The assistance of the Wisconsin Survey Research Laboratory as well as Dr. Clayton Buck and Ms. MarieLennon, Wistar Institute in providing the anonymity of the respondents is appreciated.

The authors are especially grateful to: Dr. Walter McMurray, Yale University Comprehensive Cancer Center, for performing the mass spectral analysis of these peptides; Drs. James Elliott and William Roberts, Yale University School of Medicine, for preparing the peptide used as the minor component; Mark Frazier and Ella Jones, Washington University School of Medicine Protein Chemistry Laboratory, for preparation of the primary component; as well as Kevin Beam and Clement Purcell, Wistar Institute for testing, aliquoting and packaging the standard samples.

REFERENCES

1. Niece, R.L., K.R. Williams, C.L. Wadsworth, J. Elliott, K.L. Stone, W.J. McMurray, A. Fowler, D. Atherton, R. Kutny, and A.J. Smith (1989) In "Techniques in Protein Chemistry" (T. Hugli, ed.), Academic Press, pp. 89-101.