

ABRF-00SEQ: SEQUENCE ANALYSIS OF A POST- TRANSLATIONALLY MODIFIED PEPTIDE

ABRF Protein Sequence
Research Group

ABRF Protein Sequence Research Group Members

- Steven Carr, SmithKline Beecham
- John Crabb, Cleveland Clinic Foundation
- Gary Davis, Bayer
- Bryan Dunbar, University of Aberdeen
- David Dupont, PE Biosystems
- Terry Lee, Beckman Research Institute of the City of Hope
- Len Packman, University of Cambridge
- Linda Siconolfi-Baez, Albert Einstein College of Medicine

Len Packman was an *ad hoc* member.

Purpose of Study

- Determine the state of the art in protein sequencing
- Compare analysis strategies
- Educate scientific community
- Permit ABRF members to evaluate their own performance anonymously

Design of ABRF-00SEQ

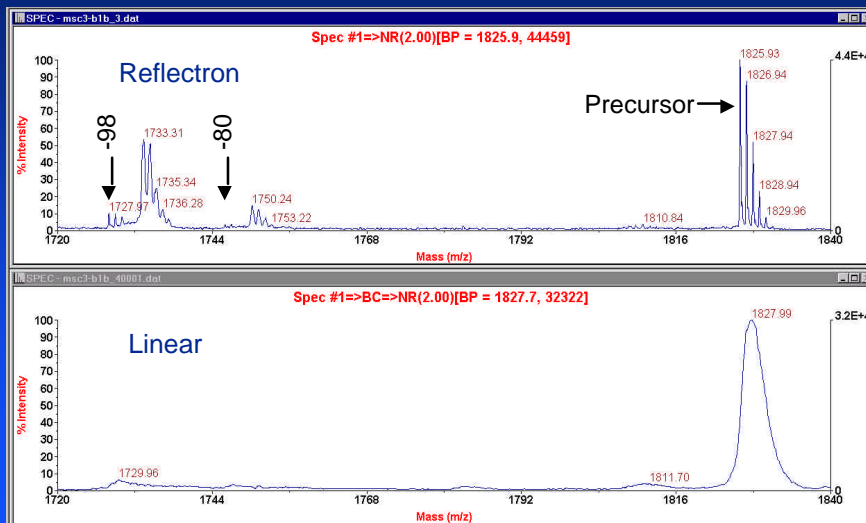
- Distributed 5 pmol of a 17-residue posttranslationally modified peptide
- Synthesized by Fmoc chemistry
- Compatible with Edman, MS/MS, and PSD sequencing
- Representative of a fragment from a tryptic digest

The sample was offered to all ABRF members through a mailing and through the ABRF newsgroup. We received and filled 108 requests for samples. Despite some sample tubes broken in shipment and people being busy during the holiday season at the end of the year, we received 46 sets of data by the deadline (a 43% response).

Participants were not told that the peptide was posttranslationally modified.

The committee thanks Dick Noble of PE Biosystems for the synthesis and purification of the peptide.

MALDI-TOF MS of ABRF-00SEQ



This page is best viewed when printed or enlarged (CTRL++).

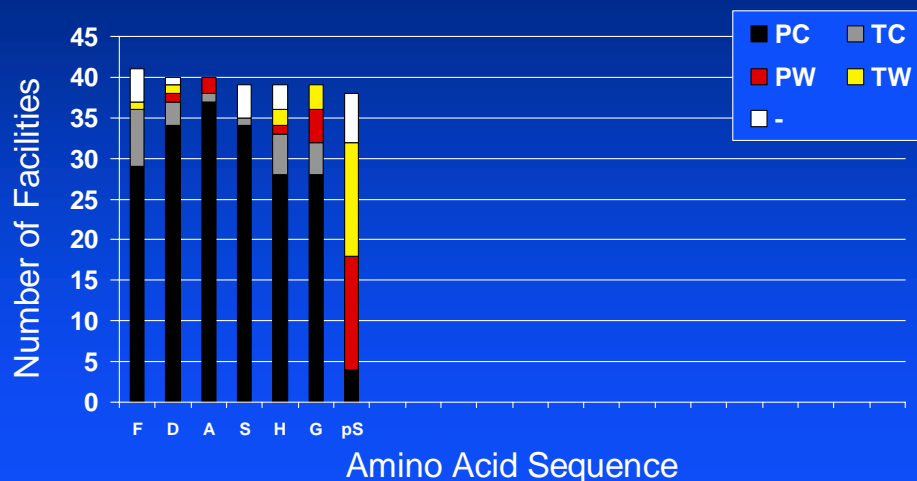
The Research Group performed MALDI-TOF mass spectrometry on ABRF-00SEQ using a Voyager DE-Pro in linear and reflectron modes. A phosphorylated peptide gives a characteristic spectrum from the loss of masses 98 (H_3PO_4) and 80 (HPO_3).

On the reflectron spectrum, the peptide is seen at m/z 1825.93 with its typical isotopic envelope, and a small peak can be seen at m/z 1727.97 from the loss of 98 due to in-source decay. Nearby, at slightly higher m/z are much larger peaks formed by post-source dephosphorylation of the peptide. These peaks show a characteristic broadness compared to the sharp resolution of the precursor ion because the reflectron is not properly focused for the energy of these ions, and they show an apparent mass difference of less than 98 from the precursor. The same phenomenon is also observed for the -80 peaks, but they are smaller and require a higher than normal laser power to be seen. In PSD mode, and with ESI-MS, the -98 and -80 ions would be seen only at the proper masses. In linear mode, all ions from post-source decay fly at the same velocity and are seen as a single peak at the expected mass for dephosphorylation. The dephosphorylation is much more evident in the reflectron mode than in the linear mode.

The dephosphorylation is observable by ESI-MS as well as by MALDI-TOF, but it is more obvious and harder to overlook by MS/MS. Of the 17 participants that determined the mass of the peptide using either MALDI-TOF in reflectron mode or ESI-MS, only 8 noted the presence of phosphorylation and 5 of these had used MS/MS.

Results of ABRF-00SEQ

No one called entire sequence correctly
7 of 46 facilities reported no positive correct residues
(pS = phosphoserine)



These are the results for the first 7 residues of the 17-residue peptide. No one called the entire sequence correctly. PC is positive correct, TC is tentative correct, PW is positive wrong, TW is tentative wrong, and - is a gap in the sequence. pS is phosphoserine.

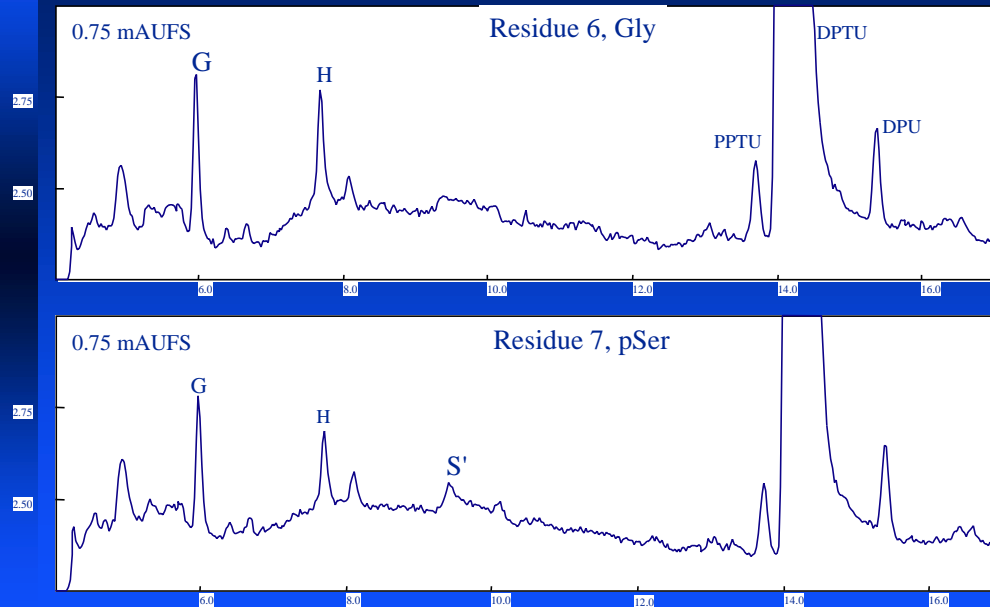
Five facilities called no sequence. Two of these facilities had complete Edman-sequencing failures, one due to an A/D box failure while another had a bad buffer B. One, using MS/MS only, got no signal using nanospray, while another got no Edman sequence. Finally, one received the sample in a broken tube and was only able to do a mass determination. In addition, there was one other participant that had an instrument problem of an injection failure in one cycle, while another had intermittent peak shifts in the chromatogram.

Low PC in cycle 1 has been common in recent studies, and is likely due to a high background in the first cycle.

Phosphoserine, which is at residue-7, is not typically observed in Edman sequencing. The four facilities that did correctly identify the phosphoserine did so by MS/MS. There was a strong tendency to try to assign one of the 20 common amino acids at the phosphoserine cycle, hence the large number of positive and tentative wrong calls.

It was our expectation that with 5 pmol of peptide, many facilities would determine the entire 17-residue sequence except for this cycle, but with an accurate mass determination, they could predict the presence of phosphoserine here. As it turned out, most facilities had sufficient ambiguities in their sequence that they could not make this assignment.

Chromatogram Comparison

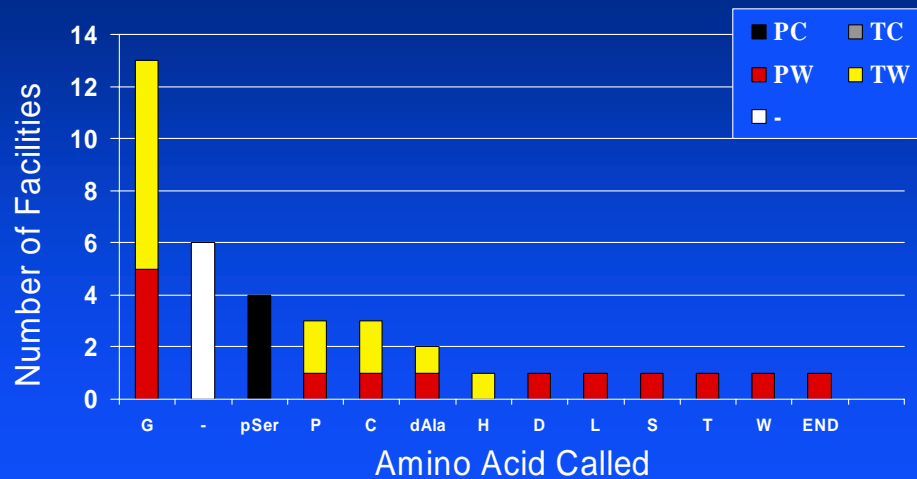


These are the chromatograms from cycles 6 and 7 of ABRF-00SEQ run on a PE Biosystems 49x-HT. The S' is PTH-dehydroalanine, which is produced in Edman sequencing by the dephosphorylation of phosphoserine. This peak may also be seen at cycles containing cysteine.

Note the large amount of glycine lag from the previous cycle. This is thought to be caused by poor coupling near the phosphoserine.

Calls for cycle 7 (pSer)

(pSer = phosphoserine; dAla = dehydroalanine)



pSer is phosphoserine. dAla is dehydroalanine.

This slide shows the amino acid calls at residue 7, the phosphoserine. The most common error was calling glycine, which was present in the previous cycle. This is not surprising, because the presence of phosphoserine can cause significant lag in that region of Edman sequencing.

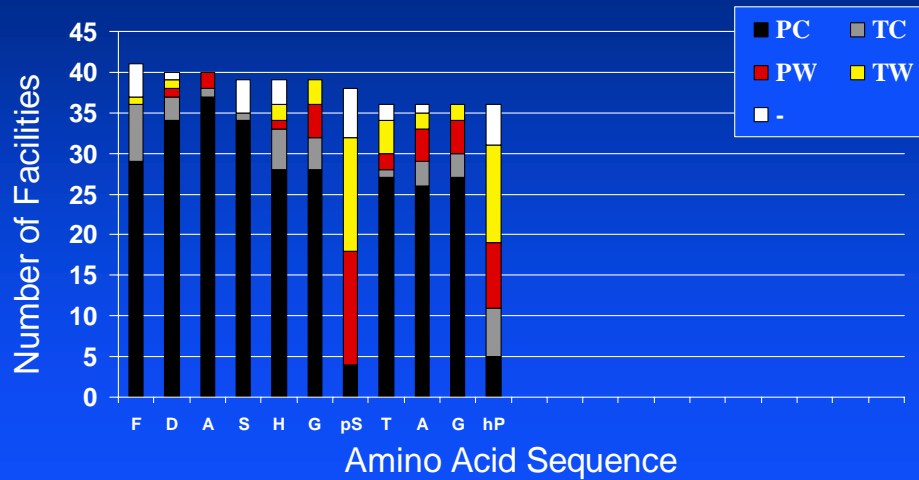
In low level analysis, a gap in the sequence can also occur at an unmodified cysteine residues. In previous studies, we often observed that cysteine was called at such gaps, but surprisingly, only three participants called cysteine here for residue 7.

One site that used MS/MS only, and one using Edman with PSD, called dehydroalanine, which is derived from the dephosphorylation of phosphoserine. One site incorrectly called this the end of the sequence.

In all, the participants called 11 different amino acids for residue 7.

Results of ABRF-00SEQ

(pS = phosphoserine; hP = hydroxyproline)

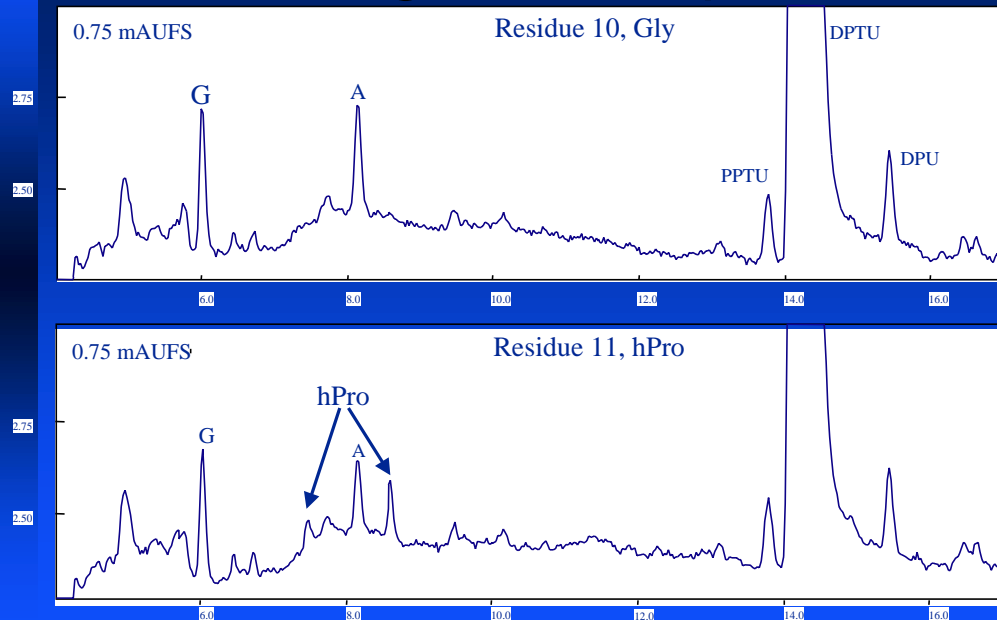


pS is phosphoserine. hP is hydroxyproline.

As we continue through the sequence up to residue 11, hydroxyproline, we find once again facilities had trouble identifying the post-translationally modified amino acid. Only 5 participants positively identified the hydroxyproline.

As with the phosphoserine, there was a strong tendency to assign one of the common amino acids, so there were many positive or tentative wrong calls.

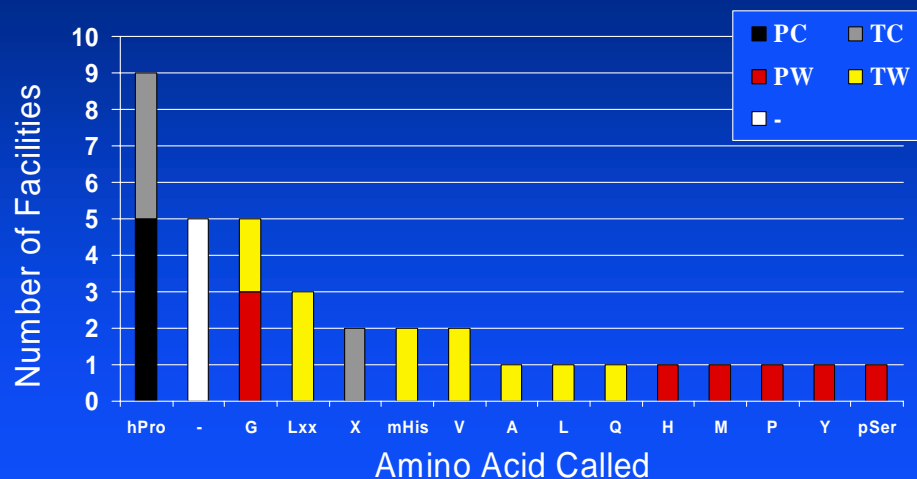
Chromatogram Comparison



Hydroxyproline is easily seen by Edman sequencing as two peaks around alanine, one before it and one after. Again, you see the large amount of lag from glycine in the previous cycle and alanine from the cycle before that.

Calls for cycle 11 (hPro)

(hPro = hydroxyproline; mHis = methylhistidine; pSer = phosphoserine)



hPro is hydroxyproline. mHis is methylhistidine. pSer is phosphoserine.

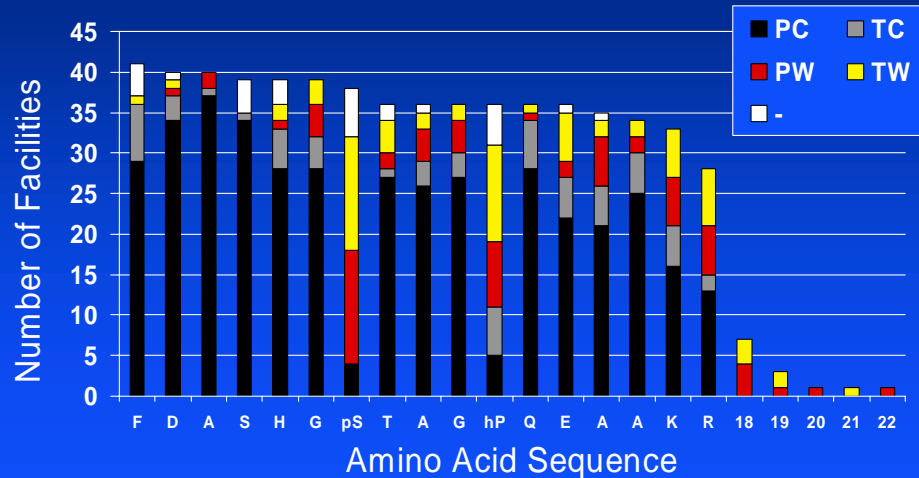
This slide shows the amino acid calls at residue 11, the hydroxyproline. There were 5 participants that positively identified the hydroxyproline, 4 more identified it tentatively, and 2 recognized the presence of an unusual amino acid, but did not identify it (the call of X, meaning an unidentified amino acid, was assigned in the chart as tentative correct). The most common error was calling glycine, which was present in the previous cycle. This is not surprising, because of the significant lag.

Three sites that used MS/MS called Lxx (Ile/Leu) at this cycle. Lxx has the same nominal mass of 113 as does hydroxyproline, so this is not surprising. The mass difference between Lxx and hPro is 0.036 Da, the same as the difference between Lys and Gln, which can be resolved on high resolution instruments such as the Q-Tof.

In all, the participants called 13 different amino acids for residue 11, including two facilities that tentatively identified methylhistidine.

Results of ABRF-00SEQ

Overall accuracy = 85.9% for positive calls
14/46 were 100% positive correct
(pS = phosphoserine; hP = hydroxyproline)



Here is the complete sequence of ABRF-00SEQ. The Q-E pair was inserted to challenge the mass spectrometrists by the 1 Da difference between these amino acids, and because glutamine has the same nominal mass as lysine (128 Da). The A-A pair provided some challenge for Edman sequencing due to the significant lag. We said that this peptide mimicked a tryptic fragment. Trypsin cleaves at lysine or arginine residues, but incomplete cleavages are common, so, as a final challenge, we added arginine at the end of the sequence after the lysine.

The significant positive-wrong (red) at Ala, Lys, and Arg (residues 14, 16, and 17, respectively) were mostly the result of severe sequencing lag, i.e. four participants positively called Lys in cycle 17 rather than in 16.

A few facilities called sequence beyond residue 17. All but one of these had not determined the mass of the peptide, and the one that did determine a mass had errors in the sequence and called only one additional residue.

The overall positive accuracy was 85.9%, slightly lower than what we've seen in recent years. A large part of this error occurred at the post-translationally modified amino acids, phosphoserine and hydroxyproline.

Facilities using MS/MS

MS/MS
Edman
PSD

```

SEQ: F D A S H G pSer T      A G hPro Q E A A K R
ME #2: F D A S H G pSer T      A G hPro Q E A A q R
M #13: f d A S H G dAla pThr A G P      g s E A A K R
MEP#15: F D A S H G pSer T      A G Lxx Q E A A K R
M #22: f D A S H G pSer T      a g Lxx q E A A k R
M #24: F D A S H G pSer T      A G Lxx Q E N R R
MEP#36: f D A S H G p      t      a L -      Q E A A K
M #46: No data.
  
```

Legend on left side, M=MS/MS, E=Edman, P=PSD. Errors are shown in magenta. Lowercase are tentative calls. SEQ, shown in green, is the correct sequence for ABRF-00SEQ.

Comments:

#2: Q ~ K mass.

#13: dAla-pThr ~ pSer-Thr mass, P-G-S = hPro-Q mass, end offset. hPro tends to cleave well on N-terminal side and see little fragmentation from cleavage on C-terminal side.

#15 - 24: Lxx ~ hPro mass.

#24: N-R ~ A-A-K mass.

#36: Automatic MS/MS sequence calling program got 3 correct from N-terminus and 8 correct from C-terminus (except called Lxx for hPro), but investigators appeared not to trust the program, perhaps because they tried to correlate the MS results with poor Edman data. Skipped C-term Arg.

#46: No signal by nanospray.

Best Edman Results (15 or 16 correct)

```
SEQ: F D A S H G pSer T A G hPro Q E A A K R
+ #3: F D A S H G c   T A G X   Q E A A K R
■ #4: F D A S H G G   T A G hPro Q E A A K R
■ #7: F D A S H G C   T A G hPro Q E A A K R
+ #10: F D A S H G P  T a g Y   q e A A K R
■ #11: F D A S H G g  T A G hPro Q E S A K r
■ #14: * D A S H G -  T A G hpro Q E A A K R
```

+ indicates mass determined

The + indicates that an accurate mass was determined. Lower case indicates a tentative call.

Comments:

Facility 10 recognized that the mass by MS did not equal proposed sequence.

Facility 14's sequence began DASH..., so it was offset one residue by the committee.

Most calls good except for pSer. All got C-terminal Arg.

Best Edman Results (15 or 16 correct)

	<u>Instrument</u>	<u>Edman %</u>	<u>Pmol Phe₁</u>
+	#3: 49x-cLC	40%	2.0
■	#4: 49x	100%	2.9
■	#7: G1005A	100%	5.1
+	#10: 49x	40%	1.6
■	#11: 49x	100%	3.4
■	#14: 49x	95%	-

+ indicates mass determined by MS

The + indicates that an accurate mass was determined.

Note variety of instruments and percent loaded.

Facility 14's sequence was offset one residue by the committee so there is no value for Phe in cycle 1.

ABRF-92SEQ

- 500 pmol
- 37-residues
- Phosphoserine at residue 4
- Hydroxyproline at residue 11
- No analysis by MS

Our sample was in many ways similar to ABRF-92SEQ from eight years ago, except that ours was distributed at 1/100th the amount, 5pmol vs. 500 pmol.

For ABRF-92SEQ, only 8% (6/74) of the participants positively identified the pSer vs. 9% (4/46) for ABRF-00SEQ. For ABRF-92SEQ, the pSer was not identified by MS, but rather by:

- Chemical method of Meyers et. al.
- Ratio of serine to dehydroalanine vs, dehydroalanine only for Cys
- No method given for the determination.

For ABRF-92SEQ 46% (34/74) correctly identified hydroxyproline vs. 20% (9/46) for ABRF-00SEQ. An additional 35% (26/74) recognized the presence of an unidentified amino acid for ABRF-92SEQ vs. 4% (2/46) for ABRF-00SEQ. The pair of peaks around Ala in the chromatogram would be much more evident in the 500 pmol sample, so that may explain the lower frequency of identification of hydroxyproline in ABRF-00SEQ.

Comparison of ABRF-SEQs

	ABRF-98SEQ	ABRF-99SEQ		ABRF-00SEQ
		Peptide	Protein	
Pmol distributed	2.8	5	10	5
Length (residues)	17	15	432	17
Avg. # cycles assigned	10.6	13.8	12.4	13.2
Avg. # correct (PC&TC)	8.3	12.6	11.5	10.1
Accuracy of positive calls	90.6%	95.4%	98.7%	85.9%
Accuracy of tentative calls	45.3%	62.2%	58.1%	46.6%

Note low accuracy for ABRF-00SEQ's positive and tentative calls relative to ABRF-99SEQ, which also had 5 pmol of peptide.

More cycles were assigned for the 5.0 pmol ABRF-00SEQ than the 2.8 pmol ABRF-98SEQ, and were similar for the 5 pmol ABRF-99SEQ. The average # correct was 2.5 lower for ABRF-00SEQ compared to ABRF-99SEQ, but that difference was mostly due to miscalls at the two posttranslationally modified residues in ABRF-00SEQ.

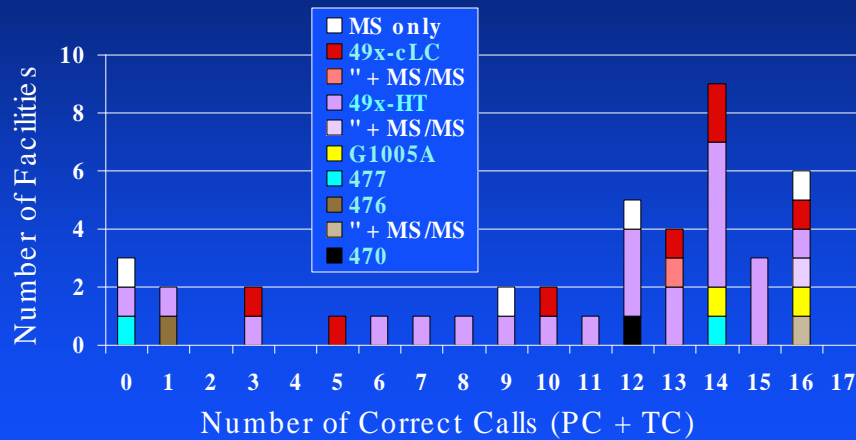
ABRF-00SEQ Positive Calls

- 39 of 46 labs (85%) made positive calls
- 25 of the 39 (64%) made 1-8 positive wrong calls compared to only 34% for ABRF-99SEQ peptide
- 11 of the 17 (64%) that determined the peptide's mass by MS made positive incorrect calls
- 14 of the 22 (63%) that did not determine the peptide's mass by MS made positive incorrect calls

The large number of positive wrong calls was mostly due to a substantial number of errors at the two posttranslationally modified residues.

Mass analysis did not help to minimize positive wrong calls. Most participants had sufficient ambiguity in their sequence that they could not match a calculated mass to the mass determined by MS.

Number of Correct Calls



Correct = PC+TC. PC = positive correct. TC = tentative correct.

Note legend entries are listed in order of the year of Edman instrument introduction.

The majority of instruments were 49x's and cLC's, and there was a wide variation in the number of correct calls using either instrument.

The figure includes three sites that used Edman sequencing and MS/MS.

Instrument Performance for ABRF-00SEQ

Edman Sequencer

Manufacturer	Model	Avg. # Cycles Correct	Positive Accuracy	Avg. % Loaded	n
PE Bio	49x-cLC	11.0	81	54	8
PE Bio	49x-HT	11.0	87	85	24
HP	G1005	*	*	90	2
PE Bio	477	*	*	95	2
PE Bio	476	*	*	90	2
PE Bio	470	*	*	80	1

* Too few instruments for accurate value

n = number of instruments. Avg. number of cycles correct includes positive and tentative calls. Positive accuracy = $(\sum PC) / (\sum PC + \sum PW)$. PC = positive correct, PW = positive wrong.

An instrument failure occurred on one PE Bio 49x-HT, and there was a bad buffer B one instrument, however the participant did not specify what model instrument was used.

Instrument Performance for ABRF-00SEQ

Mass Spectrometer

Manufacturer	Model	n
Bruker	BIFLEX/REFLEX	4
HP	G2025	1
Kratos	MALDI 4	1
Micromass	TofSpec	1
PE Biosystems	Voyager line	14
Finnigan	LCQ	2
Manitoba/Sciex	MALDI-QqTOF	1
Micromass	Q-Tof	4
Sciex	API 3+/API365	2

There were too few of any one model of instrument to accurately compare them.

The average error in mass (~ 0.1 Da) was similar for all instrument types except for three measurements using linear MALDI-TOF MS without delayed extraction where the average error was > 2 Da (linear Voyager, Kratos, and HP).

Survey Results

- 28 of the 46 respondents (61%) participated in the study last year (45 respondents)
- 35 of 46 labs used the recommended solvents
 - ◆ Median volume was 20 μl (range 5 - 200 μl)
 - ◆ Median of 90% was used for Edman sequencing
 - ◆ Median of 10% for mass determination
 - ◆ Median of 20% for MS/MS sequencing

The Research Group did stability studies and recommended how to dissolve sample.

Survey Results for Edman Sequencing

- Average experience per individual was 8.8 years (range 3 mo. - 31 years)
- Glass fiber filters were used by 85%
 - ◆ 5% used biphasic columns
 - ◆ 5% used Porton disks
 - ◆ 5% used PVDF (No positive correct)
- Routine amount sequenced:
 - ◆ 11% > 10 pmol
 - ◆ 76% 1-10 pmol
 - ◆ 14% 0.1 - 1 pmol

No sequence obtained from two labs that used PVDF.

Survey Results for MS

- Average experience per individual was 6.7 years (range 1 - 32 years)
- α -cyano-4-hydroxycinnamic acid matrix used by 85%
 - ◆ 10% used dihydroxybenzoic acid
 - ◆ 5% used sinapinic acid
- Routine amount analyzed:
 - ◆ 17% > 10 pmol
 - ◆ 61% 1-10 pmol
 - ◆ 22% 0.1 - 1 pmol

Excellent mass determinations were made with all three matrices.

Conclusions

- Sequence analysis of the posttranslationally modified ABRF-00SEQ (5 pmol) proved difficult.
 - ◆ The overall accuracy for positive calls was low (85.9%).
 - ◆ No participant called the entire sequence correctly.
- Four participants positively assigned 16 of the 17 residues in ABRF-00SEQ. Two used Edman only, and two used Edman with MS sequencing.

Including tentative calls, one participant that used only MS/MS, correctly identified 16 of the 17 residues.

Two years ago, only two labs attempted analysis of the 2.8 pmol ABRF-98SEQ sample by MS-based sequencing alone. No positive correct calls were made in either case.

Conclusions

- Eight participants noted the presence of a phosphorylated residue in ABRF-00SEQ. The loss of masses 98 and 80 in mass spectra can facilitate the identification of phosphopeptides.
- Only four participants identified phosphoserine at residue 7. All four used MS/MS.
- About 24% of the participants correctly identified hydroxyproline (hPro) at residue 11. hPro was misidentified as Leu/Ile 3x by MS/MS, in part because these residues have the same nominal mass as hPro (113 Da).

A number of participants did not recognize the presence of phosphorylation in their mass spectra. We need to educate our members.

hPro could be distinguished from Lxx by its 0.036 Da mass difference.

Conclusions

- This year, 52% (24/46) of the participants used mass spectrometry compared to 40% last year. Only two facilities that attempted to determine the molecular weight of ABRF-00SEQ were unsuccessful.
- Sequence calls past the end of the peptide were minimized by facilities that determined the peptide's mass.

Seven facilities made overcalls. Six of these did not determine the peptide's mass. The one facility that used MS and made an overcall of one residue, did so because of errors in the sequence from analysis of MS/MS data that resulted in insertion of an amino acid.

Conclusions

- The average error in mass, ~ 0.1 Da, was similar for all types of mass spectrometers except for three measurements using linear MALDI-TOF MS without delayed extraction where the average error was > 2 Da.

Recommendations

- Get an accurate mass of the peptide
- Recognize phosphorylation in MS spectrum
- Watch for unusual peaks in the Edman chromatogram
- Consider that the first Lys or Arg may not be the end of a tryptic peptide
- Track the increasing percent lag
- Use Edman and MS/MS for difficult samples or identification of posttranslational modifications