# ABRF-PRG08: Qualitative Proteomics Study - Identifying Differences in Primary Structure


N.A. Yates[1], A. Chien[2], D.B. Friedman[3], J. Krijgsveld[4], W.S. Lane[5], K.S. Lilley[6], M.J. MacCoss[7], N.E. Sherman[8], C.W. Turck[9], S.T. Weintraub[10], H.E. Witkowska[11]

[1]Merck Research Laboratories, Rahway NJ; [2]Stanford Univ., Stanford CA; [3]Vanderbilt Univ., Nashville TN; [4]Utrecht Univ., Utrecht, The Netherlands; [5]Harvard Univ., Cambridge MA; [6]Univ. of Cambridge, Cambridge, UK; [7]Univ. of Washington, Seattle WA; [8]Univ. of Virginia, Charlottesville VA; [9]Max Planck Institute of Psychiatry, Munich, Germany; [10]Univ. of Texas Health Science Center at San Antonio TX; [11]Univ. of California, San Francisco CA


## Introduction

Advances in protein characterization are central to furthering our understanding of life's biological processes. For 2008, the ABRF Proteomics Research Group (PRG) sponsored a qualitative proteomics study that investigated methods for the characterization of three protein fragments that were related in sequence but different in length. The study samples were designed to be amendable to a wide variety of analysis methods. Laboratories that specialize in top-down, bottom-up, gel-based, and gel-free methods of analysis were encouraged to participate in the study and to use any method of choice for their analysis. Results were collected via an on-line survey. In addition, participants were asked to submit a detailed description of their experimental methods and major findings. The results show that many laboratories were able to successfully identify and differentiate the three protein fragments. Results from unsuccessful analyses highlight key issues that should be considered when analyzing protein isoforms.

## Objectives

The 2008 PRG study was designed to compare different strategies used for qualitative characterization of proteins with maximum protein sequence coverage, and to enable participants to evaluate their proficiency in these techniques. The results of this study provide a survey of current methodologies as well as a vehicle for sharing further information regarding the experimental protocols, for the education and advancement of the proteomics community.

Participants were asked to:

- Analyze two study samples (labeled "A" and "B") that consisted of preparations of related human proteins that had been expressed in *Escherichia coli* and affinity purified
- Identify the major protein(s) in each sample
- Determine any differences in protein composition between A and B
- Report the sequence coverage obtained for the major protein(s) in each sample

## Sample Preparation

The extracellular region of the receptor for advanced glycation end products (RAGE), designated as sRAGE for soluble RAGE (consisting of V, C1, and C2; see structural representation in Figure 1), and two truncated variants (VC1 and V, tandem and single domain fragments, respectively) were overexpressed in *Escherichia coli* strain OrigamiB-(DE3) (Novagen) as described in Dattilo *et al., Biochemistry* 46:6957-6970, 2007. The sequences of these proteins are shown in Figure 1. Initial solutions (pH 7) were supplied as follows: sRAGE, 1.5 mg/ml in 20 mM HEPES, 200 mM NaCl; VC1, 1.0 mg/ml in 20 mM HEPES, 75 mM NaCl, 0.1 mM EDTA; V, 0.7 mg/ml, in 20 mM HEPES, 150 mM NaCl. Stock solutions for study samples A and B were then prepared in 25 mM ammonium bicarbonate, and 50-µl aliquots were transferred to polypropylene tubes. The samples were dried in a vacuum centrifuge (Speedvac, < 30 °C) and packaged for shipping. Samples for North American locations were stored at room temperature for 2 days prior to shipping in order to approximate conditions that might be experienced by samples sent to international participants.

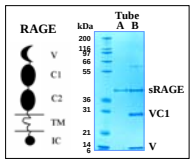
| | | Quantity | |
|---|---|---|---|
| Protein | MW (Da) | Sample A | Sample B |
| sRAGE | 33,169 | 5 µg (151 pmole) | 5 µg (151 pmole) |
| VC1 | 24,585 | | 3 µg (122 pmole) |
| V | 12,524 | | 2 µg (160 pmole) |

PRG members used SDS PAGE and mass spectrometry to analyze the sRAGE and truncated variant samples in order to confirm the composition, stability at room temperature, and amenability to different methods of analysis.

## Proteins in the Study

| | | Tube A | Tube B |
|---|---|---|---|
| **sRAGE 33,169 Da pI 7.8** | GSHM**AQNITAR**IGEPLVLKCKGAPKKPPQRLEWKLNTGRTEAWKVLSPQGGG PWDSVARVLPNGSLFLPAVGIQDEGIFRCQAMNRNGKETKSNYRVRVYQIPG KPEIVDSASELTAGVPNKVGTCVSEGSYPAGTLSWHLDGKPLVPNEKGVSVK EQTRRHPETGLFTLQSELMVTPARGGDPRPTFSCSFSPGLPRHRALRTAPIQ PRVWEPVPLEEVQLVVEPEGGAVAPGGTVTLTCEVPAQPSPQIHWMKDGVPL PLPPSPVLILPEIGPQDQGTYSCVATHSSHGPQESR**AVSISIIEPGEEG** | 5 µg | 5 µg |
| **domain VC1 24,585 Da pI 9.5** | GSHM**AQNITAR**IGEPLVLKCKGAPKKPPQRLEWKLNTGRTEAWKVLSPQGGG PWDSVARVLPNGSLFLPAVGIQDEGIFRCQAMNRNGKETKSNYRVRVYQIPG KPEIVDSASELTAGVPNKVGTCVSEGSYPAGTLSWHLDGKPLVPNEKGVSVK EQTRRHPETGLFTLQSELMVTPARGGDPRPTFSCSFSPGLPRHRALRTAPIQ PR**VWEPVPLEEVQLVVE** | | 3 µg |
| **domain V 12,524 Da pI 9.8** | GSHM**AQNITAR**IGEPLVLKCKGAPKKPPQRLEWKLNTGRTEAWKVLSPQGGGP WDSVARVLPNGSLFLPAVGIQDEGIFRCQAMNRNGKETKSNYRVR**VYQIPGKP EIVDSASE** | | 2 µg |

Figure 1. RAGE is composed of a large extracellular ligand-binding region with three Ig-like domains (one V-type and two C-type, based on sequence homology), a short transmembrane helix, and a highly acidic intracellular domain that is essential for RAGE signaling. The *C*-terminal tryptic peptide for each of the fragments used in this study is indicated in red, and the *N*-terminal tryptic peptide in blue. Sequences in the His-tag remaining after thrombin cleavage are shown in light blue. The image to the right illustrates 1D gel electrophoresis of the RAGE fragments included in the study samples. Proteins (~1.25 µg of each) were separated on a 10% NuPAGE Bis-Tris gel (Invitrogen) using a MOPS buffer system and stained with Colloidal Coomassie Blue (Invitrogen).




## Results and Conclusions

***Results*** - Samples were requested by 114 laboratories; 57 participants (ABRF members, 24; non-members, 33) submitted datasets, corresponding to a 50% return rate, as summarized in Figure 2.

Eight out of 57 participants reported the correct termini for sRAGE, VC1 and V, attesting to the feasibility of the study. Numerous participants identified the full-length form of sRAGE present in samples A and B but were not able to distinguish the truncated forms that were present in sample B.

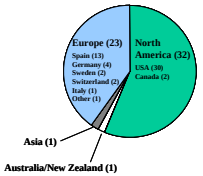
A summary of the survey results submitted by the participants is shown in Table 1. Routine analytical approaches (*e.g.*, 1D and 2D SDS PAGE, MALDI-TOF/MS, and LC-MS/MS) made up the majority of methods used. Several groups used multiple enzymes to increase sequence coverage. None of the participants reported the use of electron capture dissociation (ECD) or electron transfer dissociation (ETD) to sequence the intact proteins. A summary of the approaches used by the participants is shown in Figure 2. An example of one participant's description of their methods and findings is shown in Figure 3 along with several example responses to a question in the electronic survey.

Issues that were noted by some participants included the presence of contaminants (*e.g.*, keratin), sample degradation, and the need for more material. Interestingly, in a few result submissions, post translational modifications were reported that are not known to occur in *E. coli*. expression systems.

***Conclusions*** - These results suggest that definitive determination of protein sequence and identification of unexpected sites of truncation can be accomplished by MS-based techniques. In many cases, the combination of two complementary approaches (*e.g.*, 1D SDS PAGE to resolve protein components followed by LC-MS/MS for sequence information) gave a higher success rate than use of a single experimental approach.

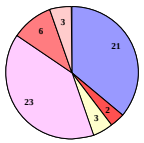
**Please visit www.abrf.org/ResearchGroups/Proteomics/Studies/ABRF_PRG08_Results.html for more in-depth examination of the methods used and the experimental results obtained.**

### Participants



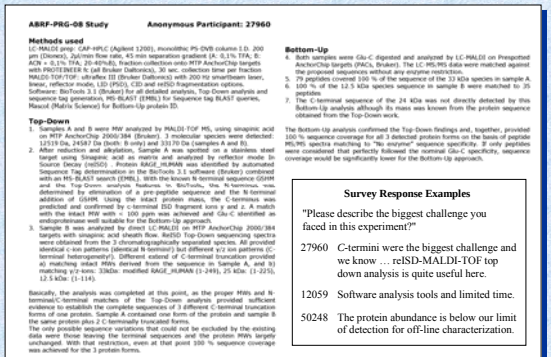

**ABRF: member, 24; non-member, 33**
**Type of lab: core, 16; core + research, 32; non-core, 8; other, 1**
**Affiliation: acad., 36; gov., 7; manuf., 7; biotech/pharma, 1; other, 1**

### Methods Used




- Bottom-up only
- Top-down only
- Unspecified top-down
- Bottom-up 1D gel
- Bottom-up 2D gel
- Bottom-up 1D and 2D gel

Figure 2. Geographic distribution of participants and summary of methods used



**ABRF-PRG-08 Study** **Anonymous Participant: 27960**

**Methods used**

LC-MALDI prep: CAP-HPLC (Agilent 1200), monolithic PS-DVB column I.D. 200 µm (Dionex), 2µl/min flow rate, 45 min separation gradient (A: 0.1% TFA; B: ACN + 0.1% TFA; 20-60%B), fraction collection onto MTP AnchorChip targets with PROTEINEER fc (all Bruker Daltonics), 30 sec. collection time per fraction
MALDI-TOF/TOF: ultraflex III (Bruker Daltonics) with 200 Hz smartbeam laser, linear, reflector mode, LID (PSD), CID and reISD-fragmentation options.
Software: BioTools 3.1 (Bruker) for all detailed analysis, Top-Down analysis and sequence-tag generation, MS-BLAST (EMBL) for Sequence tag BLAST queries, Mascot (Matrix Science) for Bottom-Up protein ID.

**Top-Down**

1. Samples A and B were MW analyzed by MALDI-TOF MS, using sinapinic acid on MTP AnchorChip 2000/384 (Bruker). 3 molecular species were detected: 12519 Da, 24587 Da (both: B only) and 33170 Da (samples A and B).
2. After reduction and alkylation, Sample A was spotted on a stainless steel target using Sinapinic acid as matrix and analyzed by reflector mode (In Source Decay (reISD)). Protein RAGE_HUMAN was identified by automated Sequence Tag determination in the BioTools 3.1 software (Bruker) combined with an MS-BLAST search (EMBL). With the known N-terminal sequence GSHM and the Top-Down analysis features in BioTools, the N-terminus was determined by elimination of a pre-peptide sequence and the N-terminal addition of GSHM. Using the intact protein mass, the C-terminus was predicted and confirmed by c-terminal ISD fragment ions y and z. A match with the intact MW with < 100 ppm was achieved and Glu-C identified as endoproteinase well suitable for the Bottom-Up approach.
3. Sample B was analyzed by direct LC-MALDI on MTP AnchorChip 2000/384 targets with sinapinic acid sheath flow. ReISD Top-Down sequencing spectra were obtained from the 2 chromatographically separated species. All provided identical c-ion patterns (identical N-termini!) but different y/z-ion patterns (C-terminal heterogeneity!). Different extend of C-terminal truncation pointed a) matching intact MWs derived from the sequence in Sample A, and b) matching y/z-ions: 33kDa: modified RAGE_HUMAN (1-249), 25 kDa: (1-225), 12.5 kDa: (1-114).

Basically, the analysis was completed at this point, as the proper MWs and N-terminal/C-terminal matches of the Top-Down analysis provided sufficient evidence to establish the complete sequences of 3 different C-terminal truncation forms of one protein. Sample A contained one form of the protein and sample B the same protein plus 2 C-terminally truncated forms.
The only possible sequence variations that could not be excluded by the existing data were those leaving the terminal sequences and the protein MWs largely unchanged. With that restriction, even at that point 100 % sequence coverage was achieved for the 3 protein forms.

**Bottom-Up**

4. Both samples were Glu-C digested and analyzed by LC-MALDI on Prespotted AnchorChip targets (PACs, Bruker). The LC-MS/MS data were matched against the proposed sequences without any enzyme restriction.
5. 79 peptides covered 100 % of the sequence of the 33 kDa species in sample A.
6. 100 % of the 12.5 kDa species sequence in sample B were matched to 35 peptides.
7. The C-terminal sequence of the 24 kDa was not directly detected by this Bottom-Up analysis although its mass was known from the protein sequence obtained from the Top-Down work.

The Bottom-Up analysis confirmed the Top-Down findings and, together, provided 100 % sequence coverage for all 3 detected protein forms on the basis of peptide MS/MS spectra matching to "No enzyme" sequence specificity. If only peptides were considered that perfectly followed the nominal Glu-C specificity, sequence coverage would be significantly lower for the Bottom-Up approach.

**Survey Response Examples**

"Please describe the biggest challenge you faced in this experiment?"

27960 C-termini were the biggest challenge and we know … reISD-MALDI-TOF top down analysis is quite useful here.

12059 Software analysis tools and limited time.

50248 The protein abundance is below our limit of detection for off-line characterization.

Figure 3. Example of a participant's description of their methods and findings and sample responses to one of the survey questions.

## Table 1

| | | | | | | Sample A Protein 1 | | Sample B Protein 1 | | Sample B Protein 2 | | Sample B Protein 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Gel | E | T | B | %Cov | N-term | C-term | N-term1 | C-term1 | N-term2 | C-term2 | N-term3 | C-term3 | Addl Incrrct |
| 27960v | | | T | B | 100 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 0 |
| 67775 | 1 | E | T | B | 93 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 0 |
| 21094 | 1 | | | B | 91 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 0 |
| 23300 | | | | B | 76 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 0 |
| 14850v | 1 | | | B | 75 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 0 |
| 57021 | 1 | | | B | 65 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 0 |
| 91239 | 1 | | | B | 40 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 0 |
| 12358 | 1 | | | B | 62 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 2 |
| 15987 | 1 | | | B | 76 | 10 | 10 | 10 | 10 | 10 | 10 | | | 0 |
| 47886 | 1 | | | B | 50 | 10 | 10 | 10 | 10 | 10 | 4 | | | 0 |
| 12303 | | | | B | 61 | 10 | 10 | 10 | 10 | 10 | 0 | | | 0 |
| 53104 | | | | B | 98 | 10 | 10 | 10 | 10 | | | | | 0 |
| 12707v | | | | B | 71 | 10 | 10 | 10 | 10 | | | | | 0 |
| 88007v | | | T | B | 92 | 10 | 10 | 10 | 10 | | | | | 3 |
| 14146v | 1 | | T | B | 91 | 10 | 9 | 10 | 9 | | | | | 0 |
| 19327 | 1 | | | B | 61 | 10 | 9 | 10 | 9 | | | | | 0 |
| 46013 | 1,2 | | | B | 79 | 10 | 0 | 10 | 0 | 10 | 0 | 10 | 0 | 0 |
| 93041 | | | | B | 92 | 10 | 0 | 10 | 0 | | | | | 0 |
| 14005 | | | | B | 44 | 10 | 0 | 10 | 0 | | | | | 0 |
| 46011 | | | T | B | 42 | 10 | 0 | 10 | 0 | | | | | 0 |
| 30366 | 1 | | | B | 0 | 10 | 0 | 10 | 0 | | | | | 0 |
| 15119 | 1 | E | | B | 21 | 10 | 0 | | | 10 | 0 | 10 | 0 | 1 |
| 14125 | 1 | | | B | 53 | 0 | 10 | 0 | 10 | 0 | 10 | 0 | 10 | 1 |
| 25058 | | | | B | 29 | 0 | 10 | 0 | 10 | 0 | 0 | | | 0 |
| 12107 | | | | B | 73 | 0 | 10 | 0 | 10 | | | | | 1 |
| 82930 | 1 | | | B | 35 | 0 | 10 | 0 | 10 | | | | | 1 |
| 26402 | 1,2 | | | B | 46 | 0 | 10 | 0 | 10 | | | | | 2 |
| 38357 | | | | B | 54 | 0 | 10 | 0 | 10 | | | | | 5 |
| 21970 | | | | B | 71 | 0 | 9 | 0 | 9 | | | | | 0 |
| 41684 | | | | B | 70 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 10 | 0 |
| 19737 | | | | B | 63 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 18984 | 2 | | | B | 24 | 0 | 0 | 0 | 0 | 0 | 0 | | | 0 |
| 73254 | 1 | | | B | 79 | 0 | 0 | 0 | 0 | 0 | 0 | | | 1 |
| 19743 | 2 | | | B | 10 | 0 | 0 | 0 | 0 | 0 | 0 | | | 3 |
| 27406 | 1 | | | B | 63 | 0 | 0 | 0 | 0 | | | | | 0 |
| 48583 | | | | B | 55 | 0 | 0 | 0 | 0 | | | | | 0 |
| 16131 | 2 | | T | | 38 | 0 | 0 | 0 | 0 | | | | | 0 |
| 21543v | 2 | | | B | 38 | 0 | 0 | 0 | 0 | | | | | 0 |
| 94661 | | | | B | 0 | 0 | 0 | 0 | 0 | | | | | 0 |
| 31508 | | | | B | 62 | 0 | 0 | 0 | 0 | | | | | 1 |
| 12016 | | | | B | 41 | 0 | 0 | 0 | 0 | | | | | 1 |
| 16132 | | | | B | 41 | 0 | 0 | 0 | 0 | | | | | 1 |
| 27467v | 1 | | | B | 35 | 0 | 0 | 0 | 0 | | | | | 1 |
| 12059 | | | | B | 56 | 0 | 0 | 0 | 0 | | | | | 2 |
| 34364 | 1 | | | B | 45 | 0 | 0 | 0 | 0 | | | | | 3 |
| 19402 | 1 | | | B | 47 | 0 | 0 | 0 | 0 | | | | | 5 |
| 94713 | | | T | | 100 | 0 | 0 | | | | | | | 0 |
| 19573 | 1 | | | B | 63 | 0 | 0 | | | | | | | 0 |
| 17406 | 1,2 | | T | B | 0 | 0 | 0 | | | | | | | 0 |
| 36189 | | | | B | 44 | 0 | 0 | | | | | | | 1 |
| 62904 | 1 | | | B | 33 | | | 10 | 0 | | | | | 0 |
| 29107 | 1 | | | B | 0 | | | | | 10 | 10 | | | 0 |
| 31231 | | | | | 36 | | | | | | | | | 0 |
| 57821 | | | | B | 14 | | | | | | | | | 0 |
| 12246 | 2 | | | B | 0 | | | | | | | | | 0 |
| 19702 | 1 | | | B | 0 | | | | | | | | | 0 |
| 50248 | | | T | | 0 | | | | | | | | | 0 |
| 79745 | | | | B | 0 | | | | | | | | | 0 |

Participant identifier (ID); Gel (1, 1D SDS PAGE; 2, 2D SDS PAGE); Edman sequencing (E); top-down analysis (T); bottom-up analysis (B). The number of correct residues for the first 10 amino acids of each terminus is shown. The right-hand column shows the number of proteins reported that were not correct. All entries within a column transition from green (best) to yellow (worst). Dark blue indicates that no protein sequence was returned.


## Acknowledgements

The PRG gratefully acknowledges the assistance of the following people: Brain Dattilo, Susan Meyn and Walter Chazin (Vanderbilt Center for Structural Biology) for supplying the RAGE samples; abd Evelin Szakal (UCSF) who served as the anonymizer.