POSTER R5-M / PROTEOMICS RESEARCH GROUP / ABRF-PRG02: IDENTIFICATION OF PROTEINS IN A SIMPLE MIXTURE

D. P. Arnott[1], M. Gawinowicz[2], R. A. Grant[3], W. S. Lane[4], L. C. Packman[5], K. Speicher[6], and K. Stone[7]

[1]Genentech, Inc., South San Francisco, CA, [2]Columbia University, New York, NY, [3]The Procter & Gamble Co., Cincinnati OH, [4]Harvard University, Cambridge, MA, [5]Cambridge University, Cambridge, United Kingdom, [6]The Wistar Institute, Philadelphia, PA, [7]Yale University, New Haven, CT

MISSION

The mission of the Proteomics Research Group (PRG) is to assist ABRF members in evaluating their capabilities to identify "unknown" proteins in order to establish realistic expectations for this technology.

ABSTRACT

Identification of multiple components in a protein mixture represents a continuing technological challenge. Whereas the main component of a simple mixture (e.g., a 1 or 2D PAGE band or spot) is often identifiable, a greater problem is posed when the proteins span a range of abundances and identification of the minor components is important. This is often the case with pull-down assays where bands may contain several components due to the complexity of the recovered protein profile, or from proteins that associate non-covalently.

The Proteomics Research Group therefore prepared and distributed a multi-component sample for analysis by ABRF members. This sample of up to 6 proteins was supplied as a tryptic digest. Some components were present at up to 2 pmol each; others at a several-fold lower level. Participants identified as many components as they could by any techniques available to them. The results should indicate which mass spectrometric or chemical approaches different laboratories applied, the levels of sensitivity that they achieved, and their ability to analyze results with different search programs. Problem areas identified in the analyses may provide a focus for future studies.

INTRODUCTION

The "completion" of the human and other genomes has created a shift in protein identification from obtaining information on highly purified novel proteins, to working with sub-pmol multiple component mixtures. SDS-PAGE (1 and 2 dimensional) is often used as the final purification step, but can still leave the protein of interest as a mixture (1) varying over a wide range of protein abundances. This is also often the case with approaches where gel bands may contain several components due to the complexity of the recovered protein profile, or from proteins that associate non-covalently. In addition, the need to handle complex protein mixtures has become more important due to the poor correlation between mRNA and protein expression levels (2, 3). Additional problems with 2D gel separation of large and small proteins, highly acidic or basic proteins, and low abundance proteins is encouraging the trend to analyze ever increasingly complex protein mixtures for protein expression studies (4).

For these reasons, the ABRF-PRG02 sample contained a protein mixture of bovine protein disulfide isomerase (~2 pmoles); *Schistosoma japonicum* glutathione-S-transferase (~2 pmoles); *E. coli.* GroEL (~200 fmoles); bovine serum albumin (~200 fmoles); and bovine superoxide dismutase (~200 fmoles). This combination of proteins was chosen since it could mimic a possible recombinant protein mixture that a membership lab might receive to analyze. For example, the bovine PDI (53kD) might be the recombinant protein of interest that is fused to GST (25kD); the GroEL (57kD) could be a co-purifying contaminant from the *E. coli* host; and the BSA (66kD) and SOD (24kD) are possible sample "contaminants".

REFERENCES

1. Gygi SP, Corthals GL, Zhang Y, Rochon Y and Aebersold R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci.* **97**, 9390-9395.

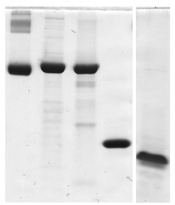
2. Gygi SP, Rochon Y, Franza BP, and Aebersold R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720-1730.

3. Lian Z, Yamaga S, Bonds W, Beazer-Barclay Y, Kluger Y, Gerstein M, Newburger PE, Berliner N, and Weissman SM (2001) Genomic and proteomic analysis of the myeloid differentiation program. *Blood* **98**, 513-524.

4. Peng J and Gygi SP. (2001) Proteomics: the move to mixtures. *J. Mass Spectrom.* **36,** 1083-1091.

METHODS

Preparation of Proteins for ABRF-PRG02



- Protein amounts determined by AAA
- 2 µg on analytical gel (shown)
  - 400 pmoles on preparative gel
- Stained gel with CBB R250

Digestion of Proteins for ABRF-PRG02

- Excised individual protein bands
- Reduced with 20 mM TCEP/ 25 mM ammonium bicarbonate (pH 8.0)
- Alkylated with 40 mM iodoacetamide/ 25 mM ammonium bicarbonate (pH 8.0)
- Digested with 0.02 µg/µl Trypsin (Promega modified)
  - 18 hr at 37ºC
- Mixed protein digests in 10 : 10 : 1 : 1 : 1 molar ratio
  - 2 : 2 : 0.2 : 0.2 : 0.2 moles
  - PDI : GST : GroEL : BSA : SOD
- Dried
- Tested by PRG member laboratories
- Mailed out to requesting laboratories

RESULTS

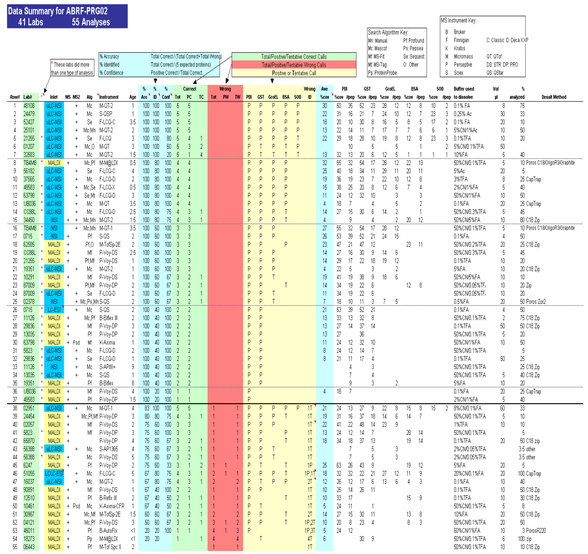


Table 1. Summary of data for ABRF-PRG02. The Key for this table is shown at the top in the boxes. This data has been sorted according to % Accuracy > % Identified > % Confidence > Average % Coverage. Analyses with all Correct assignments are in the top portion of the table.

Type of MS	# Analyses	Major Proteins		Minor Proteins			# Wrong calls
		PDI # correct	GST # correct	GroEL # correct	BSA # correct	SOD # correct	
µLC-NSI MSMS	21	21 P	18 P 3T	15 P 3 T	12 P 3 T	6 P 2 T	0 P 4 T
Nano ESI MSMS	4	4 P	4 P	2 P 1 T	1 T		
Nano ESI MS	1	1 P	1 P	1 P			
LC-ESI MSMS	1	1 P	1 P				
LCLC-ESI MSMS	1	1 P	1 P	1 P	1 T		1P 1T
MALDI-MS	25	23 P	18 P 2 T	5 P 2 T	3 P 7 T		3 P 19 T
MALDI-MS with PSD	2	2 P	1 P 1 T				0 P 1 T

Table 2. Breakdown of proteins identified as a function of MS techniques used. P = positively identified, T = tentatively identified, as decided by the individual investigators.

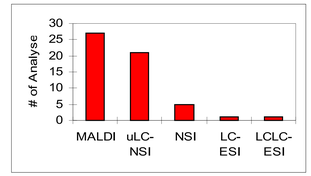


Figure 2. Summary of Mass Spectrometric Analysis of ABRF-PRG02. As summarized above, the type of MS analysis used most often was MALDI-MS with 49% (27/55) of the analyses done using this approach. µLC-NSI with MSMS analysis was used in 38% (21/55) of these analyses. Other types of MS used were NSI (9%), LCLC-ESI (2%) and LC-ESI (2%). Only the labs using µLC-NSI with MSMS were able to positively identify the SOD protein, which seemed to be the protein most difficult to identify, with only 11% of the analyses positively identifying this protein. MALDI-MS analyses had 4 positive wrong and 21 tentative wrong calls. This is not surprising due to the fact that MALDI-MS data (except for the 2 analyses that performed PSD) contains no sequence information.

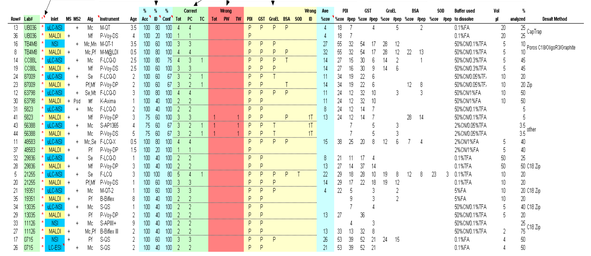


Table 3. Results for Labs that ran 2 Types of MS Analyses. There were 14 labs that ran 2 types of MS on the sample. Each of these labs ran NSI, with 13 responses also running MALDI-MS. Only 1 of these labs called all 5 proteins correctly (4PC, 1TC); 4 labs ID'd 4 proteins. Overall, it is difficult to determine the impact that 2 different types of MS analysis had on the interpretation of the data.

	# Labs
NSI & MALDI same # IDs	4
NSI more Correct IDs than MALDI	6
MALDI more correct IDs than NSI	2
NSI & MALDI differ in the proteins ID'd	1
NSI more Correct IDs than LC-ESI	1

- A total of 41 labs participated in the study with 14 labs performing 2 types of MS analyses
- Of the 55 analyses:
  - 53/55 (96%) identified PDI correctly (53 PC, 0 TC)
  - 50/55 (91%) identified GST correctly (44 PC, 6 TC)
  - 49/55 (89%) identified BOTH correctly (44 PC, 5 TC)
  - 30/55 (55%) identified GroEL correctly (24 PC, 6 TC)
  - 27/55 (49%) identified BSA correctly (15 PC, 12 TC)
  - 8/55 (15%) identified SOD correctly (6 PC, 2 TC)
  - 7/55 (13%) identified all 5 proteins correctly (Positive Correct & / or Tentative Correct), with no Wrong calls (all used LC-MSMS)
  - 4/55 (7%) identified all 5 proteins as Positive Correct, with no Tentative or Wrong calls
  - 8/55 (15%) identified all 5 proteins correctly (includes one analysis that also made a Tentative Wrong call)
  - 18/55 (33%) analyses resulted in the assignment of a protein(s) that was not in the mixture (4 PW, 25 TW)

Figure 1. A brief summary of the Results.

- µLC-NSI with MSMS and MALDI-MS were the most common types of MS performed with 21 and 27 analyses respectively
- The highest number of MALDI-MS positive correct calls, with no positive or tentative wrong, was 4, submitted by one lab that also analyzed the sample by µLC-NSI. The % coverage of the known sequences was 55% PDI; 54% GST; 28% GroEL; and 22% BSA. 3 other MALDI-MS analyses had 3 proteins identified positively correctly (PDI, GST and GroEL).
- The solvent used to dissolve the digest varied from a low % acid to a low % acid/60% acetonitrile mix. There appears to be no correlation between the solvent used and the quality of the analysis results, although the top 5 labs used a low % acid with no organic.
- 24 analyses were desalted prior to analysis in some manner, with a C18 ZipTip (Millipore) being the most common. Desalting did not appear to aid in positively identifying a correct protein and for the 2 analyses that did not ID the PDI, this may have hindered their analysis. No lab that desalted identified SOD.
- 29% of the instruments used for analysis were ≤ 1year old; 35% were 1-2 years old; 33% were 2-5 years old; and 3.6% (2 instruments) were 8-9 years old. There was no clear correlation between instrument age and the proteins identified.

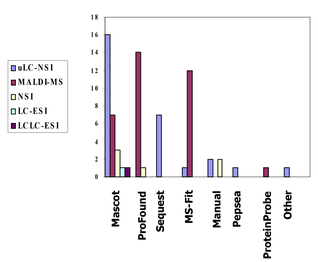


Figure 3. Bar Graph of Search Programs. The type of search program used depended on the type of analysis performed. For µLC-NSI, Mascot was used most often, while for MALDI-MS, ProFound and MS-FIT were the most used search programs. There was no clear correlation between the program used and the proteins identified.

Table 4

LC/MS - Instrumentation														Protein ID								
Lab #	*	Mass Spectrometer	Age (yrs)	HPLC	Column	Packing	Size, mm	Flow ul/min	Flow Splitter Type	Split	PDI	GST	GroEL	BSA	SOD	Incorrect ID	
48108	*	M-QTZ	1	MT-LiiPS	LCP	PMC18	0.075x150	0.2	Valco	98%	P	P	P	P	P		
52427		F-LCQ-C	3.5	HOME	HOME	Vyd C18	0.150x50	0.2			P	P	P	P	P		
02951		M-QT1	4	W-capLc	LCP	PMC18	0.075x150	0.2	Home	15:1	P	P	P	P	P	1T	
24479		S-QSP	1	HP-Ap1100	Ag	Zorb	0.300x150	4			P	P	P	P	P		
21255		F-LCQ	3	MC-MG2002	NO	C18-5	0.075x100	0.4	MC	100:1	P	P	P	P	P		
25101		M-QT2	2	W-capLC	LCP	C18	0.075x150	0.3			P	P	P	P	P		
32583		M-QT2	1.5	W-capLC	LCP	PMC18	0.075x150	0.375	LCP	90%	P	T	T	T	T		
01207				MT	LCP	C18	0.050	0.25	UPCH	98%	P	T	P	T	P		
56182		F-LCQ-C	4	ABI-140B	HOME	Phenom	0.075x80	0.3	Home	99:8%	P	P	P	P			
37565		F-LCQ-D	2	W-capLC	NO	Picofrit	0.075x150	0.3	Home	40:1	P	P	P	P			
LBO36	*	M-QT	3.5	W-capLC	LCP	PMC18	0.075x150	0.2-0.4	Home	95%	P	P	P	P			
48583	*	F-LCQ-XP	0.5	F-Surv	MC	MC	0.1x50	1	Valco	99%	P	P	P	P			
63798	*	F-LCQ-D	3	LCP-ULT	LCP	PMC18	0.1x150	0.5	LCP		P	P	P	P			
CC88L	*	F-LCQ-D	2.5	LCP-ULT	LCP	PMC18	0.3x50	4			P	P	P	T			
51055	*	F-LCQ	5	SHI-LC-VP	2Dm	1Ph1YLC PLYSULA	1x50	80			P	P	P	T		1P, 1T	
						2Keystone beta basic	1x100	10									
16037		M-QT2	1	W-capLC	NO	Picofrit	0.075x50	0.250	Valco	15:1	P	T	P	P		2T	
0715		S-QS	2	MC-MG2002	MC	C18	0.2x50	2			P	P	P				
19351	*	M-QT2	1	W-capLc	LCP	PMC18	0.18x150	1			P	P	P				
87009	*	F-LCQ-D	2	HP-1090	LCP	Vyd-C18	0.3x150	3	LCP	99.25%	P	P	T				
56388	*	S-APi1365	4								P	P	T				
5823	*	F-LCQ-D	2	LCP-ULT	LCP	PMC18	0.075x150	0.2	LCP		P	P	P				
29836	*	F-LCQ-D	1	ABI-140D	HOME	2BXC18	0.05x50	0.250	Valco	98%	P	P					
13035	*	S-QS	1	MT-II	MC		0.10x150	0.3	MT	98%	P	P					

M	Micromass (QT Q-tof)	MC	Michrom
F	ThermoFinningan (X'P D Deca C Classic)	W	Waters
S	Sciex (QS Q-Star)	ABI	Applied Biosystems
LCP	LC Packings	SHI	Shimadzu
MT	Microtech	NO	New Objective

Table 5

MALDI-Tof Instrumentation				Protein ID						
Lab #	*	Mass Spectrometer	Age (yrs)	PSD	PDI	GST	GroEL	BSA	SOD	Incorrect ID
UBO36	*	PB-Voy DE-STR	4		P					
48583	*	PB-Voy DE-PRO	1.5		P					
63798	*	K-Axima	1	Y	P	P				
21255		PB-Voy DE-STR	1		P	P	P			
TBAM8	*	M-MgBiDI	0.5		P	P	P			
CC88L	*	PB-Voy DE-STR	2.5		P	P	P			
24454		PB-Voy DE-PRO	3		P	P	T		1T	
19351	*	BR-Biflex	8		P	P				
10291		PB-Voy STR	1		P	P	T			
87009	*	PB-Voy DE-PRO	2		P	P		T		
5823	*	PB-Voy DE-PRO	3		P	P		P	1T	
02057		PB-Voy DE-STR	2		P	P	P		(1T)	
56388	*	PB-Voy DE-STR	5		P	P	T		1T	
65870		PB-Voy DE-PRO	4		P	P		T	1T	
62585		M-ToFSp-2E	2		P	P	P			
04121		PB-Voy DE-STR	3		P	P		T	1P,2T	
30967		M-ToFSp-2E	1.5		P	P	P	T	2T	
6247		PB-Voy DE-PRO	2		P	P	T	T	1P	
29836	*	PB-Voy DE-PRO	3		P	P	P			
12510		BR-Reflex III	2		P			T	1T	
11126	*	BR-Biflex III	2		P	P				
90991		PB-Voy DE-STR	1		P	P	P		1T	
10461		K-Axima-CFR	1		P	T			1T	
13035	*	PB-Voy DE-PRO	2		P	P				
48011		BR-AutoFlx	<1		P				1P,3T	
18273		M-MgBiDI	<1			T			4T	
06443		M-ToF Spc II	2						1T	

PB	Perseptive Biosystems		
K	Kratos		
BR	Bruker		
M	Micromass		

Table 6

Nanospray Instrumentation				Protein ID					
Lab #	*	Mass Spectrometer	Needle	PDI	GST	GroEL	BSA	SOD	Incorrect ID
TBAM8	*	M-QT	Micromass-F	P	P	P			
34450		M-QT2	Protana	P	P	P	T		
02378		SX-GSTR	Protana	P	P	T			
11126	*	SX-APIIII+	Protana	P	P				

Tables 4, 5, 6. Detailed information on instrumentation used in study

CONCLUSIONS

One of the goals of this study was to compile protein identification data on a representative unknown in order to help establish realistic expectations for proteomic analysis. From this study, it appears that samples at 2 pmoles can be identified with high % accuracy using a variety of MS approaches. At a 10-fold lower amount (~200 fmoles), it became increasingly difficult for MALDI-MS analyses to identify the 3 lower-abundance proteins. This may be due in part to the complex mixture (i. e., some peptides may have been suppressed, or the % coverage of the known protein was too low for an identification), and not to the sensitivity of the instrumentation used. µLC-NSI with MS/MS analysis did better than MALDI-MS at the 200 fmole level, perhaps due to the fact that in MS/MS analysis, individual peptides are interpreted instead of a mass list as in MALDI-MS.

The 7 labs that identified all 5 proteins correctly (tentative and positive) used LC-MS/MS. Although there were several times more positive and tentative wrong identifications using MALDI-MS than LC-MS/MS, it should be emphasized that the vast majority of these wrong calls were classified as tentative.

Overall, the current study shows a marked improvement from the Protein Identification Research Group (PIRG) study done in 1999 (which was the last study of this type). In that study, the sample contained a mixture of 2 proteins at the 10 and 2 pmole levels. 97% of the calls for the major protein were correct, while 77% of the calls for the minor protein were *incorrect*. Four years later, 96% of the calls for PDI at the 2 pmole level are correct.