Use of Multiple Peptide Identification Programs to Increase Peptide and Protein Identifications in Human Whole Saliva

Phillip A. Wilmarth[1], Brian C. Searle[2], D. Leif Rustvold[1], Michael A. Riviere[1], Larry L. David[1]

[1]Oregon Health & Science University, Portland, Oregon; [2]Proteome Software, Inc., Portland, Oregon

Introduction: Many proteomic studies of complex biological fluids, such as saliva or serum, attempt to identify as many proteins as possible. Enzymatic digestion followed by two- or three-dimensional liquid chromatography mass spectrometry has high sensitivity, but can generate in excess of 100,000 tandem mass spectra (MS/MS) per biological sample. This poses challenges for reliable peptide identification, and the related problem of consistent, non-redundant protein reporting. Trypsin digests of whole human saliva were separated by two-dimensional liquid chromatography, peptides identified using SEQUEST,[1] Mascot,[2] and X! Tandem[3,4], and results compared using several protein reporting programs.[5-7] Two techniques were found to increase the number of identified proteins: combining results from multiple peptide identification programs using a probabilistic model, and comparing proteins identified in three biological replicates. Evaluation of both techniques was facilitated by the new analysis software Scaffold™ (Proteome Software, Inc. Portland, OR).

Samples and data sets: Unstimulated human whole saliva was collected with IRB approval from three healthy non-smokers (male, non-pregnant female, and 32w gestation pregnant female). All samples were centrifuged to remove debris. The male sample was desalted using dialysis (10k MWCO) and female samples desalted using spin filtration (5k MWCO). Trypsin digests of 1.0 to 2.5 mg protein were performed, peptides separated offline by polysulfoethyl A cation exchange, and fractions collected (37 for male, 45 for females). Reverse phase LC-MS/MS was performed using either a 75 µm x 150 mm (male) or 0.5 mm x 250 mm (females) reverse phase columns packed with Zorbax SB-C18 stationary phase (Agilent) using a 7-30% ACN gradient at either 1.5 or 10 µl/min flow rate, respectively. An LCQ classic (ThermoFinnigan, San Jose, CA) ion trap was used to acquire tandem mass spectra utilizing data-dependent scanning and dynamic exclusion. The total number of DTA files was 27,494 for the male (one MS/MS scan per survey scan), 78,908 for the pregnant female (3 MS/MS per survey scan), and 91,151 for the non-pregnant female (also 3 MS/MS per survey scan).

Databases and search programs: Searches were performed against two human-only databases, downloaded February 22, 2005. One was the SwissProt database with 8 additional salivary proteins (12,069 total entries). The other was a combination of SwissProt and NCBI databases where identical, redundant sequences were removed (115,633 total entries). Both databases contained sequences of common contaminants. Three search programs were used: SEQUEST (v2.7, ThermoFinnigan), Mascot (v2.0.04, Matrix Science), and X! Tandem (release 2005.03.21, Beavis Informatics Ltd.). Three different search results compilers were evaluated: DTASelect (v1.9, The Scripps Research Institute),[5] PeptideProphet™/ProteinProphet™ (release 1.4-3, Institute for Systems Biology),[6,7] and Scaffold™ (release 1.0.04 Beta).

Acknowledgements: The support of the OHSU School of Dentistry and the NIH is gratefully acknowledged. We would like to thank Ashok Reddy for performing sample processing, and Surendra Dasari for running the Mascot searches.

References:

1. "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database." J.K. Eng, A.L. McCormack, and J.R. Yates, 3rd, *J. Am. Soc. Mass Spectrom.*, 5, 976-989 (1994).
2. "Probability-based protein identification by searching sequence databases using mass spectrometry data." D.N. Perkins, D.J.C. Pappin, D.M. Creasy, and J.S. Cottrell, *Electrophoresis*, 20, 3551-3567 (1999).
3. "A method for reducing the time required to match protein sequences with tandem mass spectra." R. Craig and R.C. Beavis, *Rapid Commun. Mass Spectrom.*, 17, 2310-2316 (2003).
4. "Open source system for analyzing, validating, and storing protein identification data." R. Craig, J.P. Cortens, and R.C. Beavis, *J. Proteome Res.*, 3, 1234-1242 (2004).
5. "DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics." D.L. Tabb, W.H. McDonald, and J.R. Yates, 3rd, *J Proteome Res*, 1, 21-26 (2002).
6. "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search." A. Keller, A.I. Nesvizhskii, E. Kolker, and R. Aebersold, *Analytical Chemistry*, 74, 5383-5392 (2002)
7. "A statistical model for identifying proteins by tandem mass spectrometry." A.I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold, *Analytical Chemistry*, 75, 4646-4658 (2003).
8. "Two-dimensional liquid chromatography study of the human whole saliva proteome." P.A. Wilmarth, M.A. Riviere, D.L. Rustvold, J.D. Lauten, T.E. Madden, and L.L. David, *J. Proteome Res.*, 3, 1017-1023 (2004).
9. "Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome." W-J Qian, T. Liu, M.E. Monroe, E.F. Strittmatter, J.M. Jacobs, L.J. Kangas, K. Petritis, D.G. Camp II, and R.D. Smith, *J. Proteome Res.*, 4, 53-62 (2005).
10. "The need for guidelines in publication of petide and protein identification data." S. Carr, R. Aebersold, M. Baldwin, A. Burlingame, K. Clauser, and A. Nesvizhskii, *Mol. Cell. Proteomics*, 3, 531-533 (2004).

Comparison of protein reports for a known sample: The male whole saliva data set was previously analyzed[8] and 102 proteins were reliably identified. This data set was used to evaluate the performance of a variety of results compilers using identification criteria listed in Table 1 below. Two representative databases (described above) were used to test dependence on database complexity and size. Sequence-reversed databases give reliable false positive identification estimates[9] and were used to test the criteria stringency. The results were subjectively ranked from poorest performance (red) to best performance (dark blue) with rainbow colors. Desirable characteristics would be little dependence on database choice, and a low false positive identification rate (a small number in parenthesis). Probabilistic models were generally superior. The trends apparent in Table 1 support recent guidelines[10] for publishing proteomic data. However, use of reversed databases should also be considered since it is a particularly good indicator of an inflated protein number.

Table 1. Comparison of identification criteria for protein results compilers capable of processing large data sets. The male saliva dataset (27,494 DTA files) was used in all cases.

Protein Compiler and Parameters	Number of Reported Proteins (Reversed Database Number)			
	One Peptide Minimum		Two Peptides Minimum	
	SwissProt	SwissProt + NCBI nr	SwissProt	SwissProt + NCBI nr
DTASelect defaults*, any tryptic status	613 (823)	956 (965)	170 (72)	397 (63)
DTASelect defaults*, any tryptic status, Occam's razor	584 (809)	655 (927)	153 (68)	167 (50)
DTASelect defaults*, full or half tryptic	288 (217)	578 (260)	134 (14)	367 (12)
DTASelect defaults*, full or half tryptic, Occam's razor	268 (215)	301 (254)	119 (12)	144 (8)
DTASelect defaults*, fully tryptic	185 (22)	410 (21)	113 (2)	284 (2)
DTASelect defaults*, fully tryptic, Occam's razor	173 (21)	192 (19)	104 (1)	115 (1)
DTASelect, Qian et al. criteria**	174 (12)	198 (14)	109 (1)	130 (1)
Protein Prophet Suit, Protein Prob. > 0.95	178 (0)	204 (8)	124 (0)	138 (1)
Protein Prophet Suit, Protein Prob. > 0.99	122 (0)	139 (8)	122 (0)	138 (1)
Scaffold, Prot. Prob. > 0.95, Pep. Prob. > 0.95	134 (2)	146 (0)	109 (1)	125 (0)
Scaffold, Prot. Prob. > 0.99, Pep. Prob. > 0.95	114 (2)	128 (0)	107 (1)	119 (0)
the GPM, log(e) < -3	209 (42)	503 (74)	185 (40)	227 (64)
the GPM, log(e) < -4	194 (33)	452 (52)	172 (31)	211 (47)

*Default thresholds (-1 1.8 -2 2.5 -3 3.5 -d 0.08).
**Tryptic dependent thresholds: (-1 2.0 -2 2.4 -3 3.7 -d 0.1) for fully tryptic, and (-1 3.0 -2 3.5 -3 4.5 -d 0.1) for half-tryptic taken from Ref. 9.

Comparison of individual search program results and statistical combinations of search program results: Scaffold™ extends the algorithms used in PeptideProphet™ [6] and ProteinProphet™ [7] to allow results from more than one search program to be combined at the MS/MS level. The male saliva data set (previously reported[8] to contain 102 proteins) was searched with non-enzymatic specificity using the three programs mentioned above with ion trap appropriate settings against the SwissProt database (12,069 entries). All results exceeding specified thresholds (protein probability >0.95, 2 unique peptides, and peptide probability >0.90, denoted "Total") were manually validated ("Valid"), contaminant and redundant immunoglobulin entries removed, and final results ("Final") shown in Figure 1 below. The number of protein identifications increased by 9% when using all three search results compared to using SEQUEST alone. The number of MS/MS spectra matched to identified proteins also increased from 5113 to 6606, a 29% increase.

Figure 1.

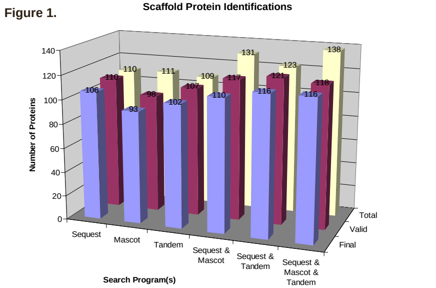
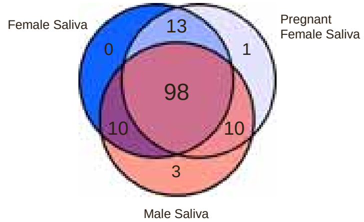
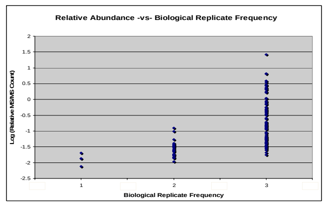


Figure 2. Distribution of identified proteins among three biological replicates.



Use of biological replicates to increase proteome depth: An alternative strategy of using biological replicates to increase the number of protein identifications in saliva was also explored. For simplicity, only non-enzymatic SEQUEST searches against the SwissProt database were used for the three samples (male, female, and pregnant female) and the results were compared with Scaffold™. Using strict filtering criteria (protein probability >0.95, 2 unique peptides per protein, and peptide probability >0.90) with manual validation, the numbers of proteins identified in each sample were 106 (male), 107 (pregnant), and 109 (female). The union of the identified protein lists was 135 distinct proteins (Table 2) and their distribution among the three samples is shown above in Figure 2. There were 131 proteins (98%) that were detected in two or more samples. More abundant proteins (estimated from MS/MS spectral counts) were observed with greater frequency and the proteins observed in only one sample were among the least abundant (Figure 3 below).

Figure 3. Trend of observation frequency with protein abundance.



Conclusions: Only more sophisticated protein results compilers such as DTASelect (with Occam's razor filter), ProteinProphet™, or Scaffold™ were capable of producing accurate lists of proteins from MudPIT data sets. Searches using the NCBI non-redundant database are particularly problematic, and results compilers should include an implementation of an Occam's razor algorithm, especially for larger databases. Lacking other experimental evidence, a minimum criterion of at least two peptides per protein should be adopted. The probabilistic combining of two or more search results from a given data set, now possible with Scaffold™, can increase both the number of identified proteins, and the number of MS/MS spectra matched to those proteins. The number of proteins reliably identified in human whole saliva was also increased 27% by analyzing two additional biological replicates, even though only a single search program (SEQUEST) was used. Additional proteins could be identified in each biological sample by combining multiple search programs with Scaffold™, and the resulting list of human whole saliva proteins further increased.

Table 2. Human whole saliva proteins identified using three biological replicates. The total number of matched MS/MS spectra are listed and proteins are ranked in decreasing order of average fractional MS/MS count.

	Acc. No.	Protein Name	Female	Pregnant	Male
1	P04745	Salivary alpha-amylase precursor	2728	2793	923
2	P01036	Cystatin S precursor	432	517	476
3	P01037	Cystatin SN precursor	399	503	467
4	9 entries	Ig kappa chain	555	370	76
5	P02768	Serum albumin precursor	341	212	272
6	P01833	Polymeric-immunoglobulin receptor precursor	416	418	52
7	P02810	Salivary acidic proline-rich phosphoprotein 1/2 precursor	361	160	150
8	P09228	Cystatin SA precursor	246	276	135
9	P02812	Basic salivary proline-rich protein 2	290	252	121
10	P02814	Proline-rich protein 3 precursor	242	133	196
11	Q9HC84	Mucin 5B precursor	52	199	216
12	P01877	Ig alpha-2 chain C region	205	264	76
13	P01876	Ig alpha-1 chain C region	230	219	83
14	P28325	Cystatin D precursor	64	94	224
15	P04280	Basic salivary proline-rich protein 1 precursor	310	121	74
16	P02788	Lactotransferrin precursor	252	199	57
17	P25311	Zinc-alpha-2-glycoprotein precursor	189	172	63
18	P12273	Prolactin-inducible protein precursor	118	190	88
19	4 entries	Ig lambda chain	190	205	31
20	Q04118	Basic salivary proline-rich protein 3 precursor	151	65	46
21	Q96DR5	Short palate, lung & nasal epithelium carcinoma assoc. protein 2	25	39	121
22	P22079	Lactoperoxidase precursor	69	73	55
23	P10161	Basic salivary proline-rich protein 4 allele M	115	33	43
24	P61626	Lysozyme C precursor	24	33	91
25	P02808	Statherin precursor	56	31	61
26	gi|16307266	Similar to common salivary protein 1	31	88	39
27	P23280	Carbonic anhydrase VI precursor	77	62	29
28	P04080	Cystatin B	75	64	26
29	P02787	Serotransferrin precursor	42	57	32
30	P01871	Ig mu chain C region	79	55	7
31	P10163	Basic salivary proline-rich protein 4 allele S precursor	75	23	20
32	P01857	Ig gamma-1 chain C region	54	25	30
33	P01034	Cystatin C precursor	48	40	24
34	P01859	Ig gamma-2 chain C region	58	19	26
35	P54108	Cysteine-rich secretory protein-3 precursor	29	50	21
36	P01591	Immunoglobulin J chain	61	42	6
37	P15515	Histatin 1 precursor	10	20	41
38	Q8N4F0	Bactericidal/permeability-increasing protein-like 1 precursor	27	41	19
39	P61769	Beta-2-microglobulin precursor	47	30	9
40	P63261	Actin, cytoplasmic 2	35	16	22
41	P80303	Nucleobindin 2 precursor	24	25	20
42	P31025	Von Ebner's gland protein precursor	10	12	26
43	gi|5733598	Deleted in malignant brain tumors 1	12	32	13
44	gi|52546027	Hypothetical protein	11	2	29
45	P20061	Transcobalamin I precursor	21	23	5
46	P06733	Alpha enolase	22	8	11
47	P01009	Alpha-1-antitrypsin precursor	13	23	6
48	Q08380	Galectin-3 binding protein precursor	19	15	5
49	2 entries	Ig heavy chain	24	16	1
50	P63267	Actin, gamma-enteric smooth muscle	12	9	11
51	P06702	Calgranulin B	2	6	17
52	P09211	Glutathione S-transferase P	18	12	5
53	Q01469	Fatty acid-binding protein, epidermal	14	12	6
54	P01033	Metalloproteinase inhibitor 1 precursor	7	16	6
55	Q07654	Trefoil factor 3 precursor	7	14	7
56	P06870	Kallikrein 1 precursor	19	8	4
57	P02042	Hemoglobin delta chain	1		18
58	P00738	Haptoglobin precursor	4	2	15
59	Q9UBC9	Small proline-rich protein 3	14	11	4
60	gi|1944352	Fc fragment of IgG binding protein	1	2	16
61	P10909	Clusterin precursor	12	14	3
62	P60174	Triosephosphate isomerase	15	7	5
63	P69905	Hemoglobin alpha chain	1		14
64	P03973	Antileukoproteinase 1 precursor	3	7	8
65	Q14508	WAP four-disulfide core domain protein 2 precursor	6	11	4
66	P62937	Peptidyl-prolyl cis-trans isomerase A	13	7	2
67	P29508	Squamous cell carcinoma antigen 1	10	8	3
68	P07602	Proactivator polypeptide precursor	3	10	4
69	P04075	Fructose-bisphosphate aldolase A	9	4	3
70	P02774	Vitamin D-binding protein precursor	3	11	2
71	P01024	Complement C3 precursor	5	2	6
72	Q14515	SPARC-like protein 1 precursor	3	7	4
73	P07339	Cathepsin D precursor	9	4	2
74	P04406	Glyceraldehyde-3-phosphate dehydrogenase, liver	1	2	7
75	P08571	Monocyte differentiation antigen CD14 precursor	8	4	2
76	P07108	Acyl-CoA-binding protein	6	7	1
77	P05109	Calgranulin A		2	7
78	Q8NBJ4	Golgi phosphoprotein 2	1	8	3
79	P07858	Cathepsin B precursor	3	9	1
80	Q9UBX7	Kallikrein 11 precursor	3	7	2
81	P02679	Fibrinogen gamma chain precursor	3	4	3
82	P06396	Gelsolin precursor	4	1	4
83	P02763	Alpha-1-acid glycoprotein 1 precursor	5	5	1
84	P80188	Neutrophil gelatinase-associated lipocalin precursor	4	4	2
85	P02647	Apolipoprotein A-I precursor	3	3	3
86	P30086	Phosphatidylethanolamine-binding protein	6		3
87	P13796	L-plastin	3	1	4
88	P00450	Ceruloplasmin precursor	3	6	1
89	P22528	Cornifin B	5	4	1
90	P02750	Leucine-rich alpha-2-glycoprotein precursor	5	2	2
91	P23284	Peptidyl-prolyl cis-trans isomerase B precursor		3	4
92	P02765	Alpha-2-HS-glycoprotein precursor		1	5
93	P30740	Leukocyte elastase inhibitor	5	1	2
94	Q02487	Desmocollin-2 precursor	5	1	2
95	P14618	Pyruvate kinase, isozymes M1/M2		2	4
96	P02766	Transthyretin precursor	2		4
97	P02790	Hemopexin precursor	3	4	1
98	P18510	Interleukin-1 receptor antagonist protein precursor	7	2	
99	P16870	Carboxypeptidase E precursor		5	2
100	P10599	Thioredoxin	3	2	2
101	P31947	14-3-3 protein sigma	4	1	2
102	P01028	Complement C4 precursor	3		3
103	Q96QR1	Uteroglobin-related protein 2 precursor	1	7	
104	P52209	6-phosphogluconate dehydrogenase, decarboxylating	3	3	1
105	P32926	Desmoglein 3 precursor	2	2	2
106	P37802	Transgelin-2	3	1	2
107	P02675	Fibrinogen beta chain precursor	4		2
108	O43852	Calumenin precursor	2	3	1
109	P61916	Epididymal secretory protein E1 precursor	3	2	1
110	P02749	Beta-2-glycoprotein I precursor		3	2
111	P07237	Protein disulfide-isomerase precursor		1	3
112	P01040	Cystatin A		1	3
113	P29401	Transketolase	1		3
114	P07737	Profilin-1	2	4	
115	P13987	CD59 glycoprotein precursor	3	3	
116	Q02818	Nucleobindin 1 precursor	1	1	2
117	P31151	S100 calcium-binding protein A7			3
118	P01011	Alpha-1-antichymotrypsin precursor	1	4	
119	P15328	Folate receptor alpha precursor	3	2	
120	Q9UIV8	Hurpin	3	2	
121	P01023	Alpha-2-macroglobulin precursor	4	1	
122	P08107	Heat shock 70 kDa protein 1	1	2	1
123	P04217	Alpha-1B-glycoprotein precursor	3		1
124	Q06828	Fibromodulin precursor		1	2
125	P27482	Calmodulin-related protein NB-1	1		2
126	P55058	Phospholipid transfer protein precursor	1	3	
127	Q96BQ1	Protein FAM3D precursor	2	2	
128	P36952	Maspin precursor	3	1	
129	Q96S96	PEBP family protein precursor	3	1	
130	P00558	Phosphoglycerate kinase 1		2	1
131	P06744	Glucose-6-phosphate isomerase	2		1
132	P18669	Phosphoglycerate mutase 1			2
133	P35579	Myosin heavy chain, nonmuscle type A			2
134	P11142	Heat shock cognate 71 kDa protein	2	1	
135	Q9UKR0	Kallikrein 12 precursor		2	