

Improving Sensitivity by Combining Results from Multiple MS/MS Search Methodologies With the Scaffold Computer Algorithm

Brian C. Searle, James M. Brundage, and Mark Turner

Proteome Software Inc., 1336 SW Bertha Blvd, Portland, OR, 97219-2039, United States
E-mail: Brian.Searle@ProteomeSoftware.com, Telephone: (800) 944-6027, Fax: (503) 245-4910

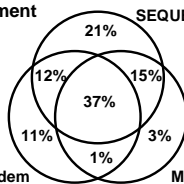
Abstract

Database-searching programs generally identify only a fraction of the spectra acquired in a standard LC/MS/MS study of digested proteins. Subtle variations in database-searching algorithms of MS/MS spectra have been known to provide different identification results. To leverage this variation, we developed Scaffold to probabilistically combine the results of multiple search engines, including SEQUEST, Mascot, and X!Tandem. Sensitivity is increased from 20% to 100% with each additional search engine, primarily due to increased confidence in identifications of lower scoring MS/MS spectra. In this experiment, SEQUEST, Mascot, and X!Tandem were used in combination to increase the number of highly confident spectrum identifications (>95%) between 1.5 and 4.5-fold in human lens tissue experiments. These increases allow us to confidently identify a substantially larger number of proteins in LC/MS/MS studies over single database searches alone.

Framing The Problem

SEQUEST (1), Mascot (2) and X! Tandem (3) are three commonly used database-searching tools for MS/MS driven proteomics. These three programs were employed to interpret the data from a mixture of protein standards. As shown in Figure 1, approximately a third of the matched peptides were identified by all of the search engines. However, due to subtleties in the different statistical methods applied, the results also contained subsets of peptides that were uniquely identified by each algorithm. Figure 2 shows that these differences can manifest as very different scores for the same identifications. Therefore, one easy way to increase the number of possible protein identifications in a sample is by combining the results of more than one search engine.

Ion Trap Control Experiment



Q-ToF Control Experiment

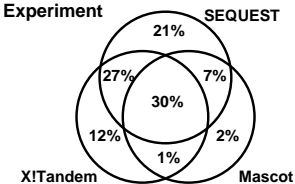


Figure 1: Venn diagram of correct spectrum identifications made by SEQUEST, X!Tandem, and Mascot at the 95% confidence level. The overlap between the three programs is surprisingly small.

Scaffold was developed to use this variation, as well as variation between biological samples, to provide increased confidence in results from MS/MS studies of enzymatically digested proteins. Scaffold accomplishes this task by:

- 1) Estimating peptide and protein identification probabilities using Bayesian statistics,
- 2) Combining the probabilities from various database-searching algorithms,
- 3) Considering protein identifications derived from multiple samples simultaneously

This increased confidence results in two outcomes. Probabilistic analysis allows researchers to control their own minimum confidence levels, such as 99% for rigorous investigations or 50% for fishing experiments. Secondly, the Bayesian statistical algorithms used automatically adjust for variation between samples, opening up the possibility of comparing multiple database-searching algorithms, instruments, and samples. By considering evidence from multiple sources, Scaffold identifies more peptides and proteins than single search engines at equivalent confidence levels.

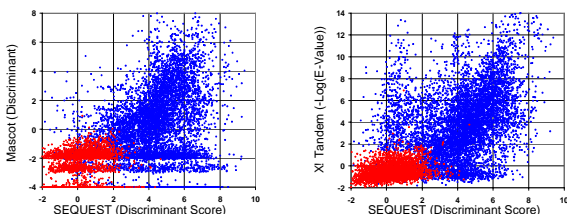


Figure 2: Scoring variation between SEQUEST, Mascot and X! Tandem when analyzing Ion Trap data from a mixture of protein standards. Correct identifications are labeled in blue while incorrect are labeled in red.

Bayesian Method To Combine Searching Engines

To compare these scoring systems, we first normalize each search engine result independently using a new implementation of the PeptideProphet algorithm (4), which uses score distributions, expectation maximization, and Bayesian statistics to calculate the probability that an MS/MS spectrum was derived from a matching peptide sequence. Next, Scaffold combines the estimated identification probabilities from the various database-searching programs for a single spectrum using an algorithm based on ProteinProphet (5). Assuming that an identification is more likely to be correct if more than one search engine makes it, Scaffold employs the "number of sibling programs" score (NSP) to sum the peptide identification probabilities for the five top hits that each search program computed for a given spectrum:

$$NSP_{i,k} = \sum_{j=1}^5 p(+|D_{i,j,k}), \text{ where } p(+|D_{i,j}, NSP_{i,k}) = \frac{p(+|D_{i,j}) \cdot p(NSP_{i,k} | +)}{p(+|D_{i,j}) \cdot p(NSP_{i,k} | +) + p(-|D_{i,j}) \cdot p(NSP_{i,k} | -)}$$

A high NSP means that at least one other program came up with the same match for the spectrum; a low NSP suggests little agreement between the search results. The probability of a peptide being correct while having a particular score (D) and that NSP value is $p(+|D_{i,j}, NSP_{i,k})$.

$$p(+|D, NSP) = 1 - \prod_{j=1}^5 (1 - p(+|D_{i,j}, NSP_{i,k}))$$

Finally, an aggregate probability, $p(+|D, NSP)$, is determined by combining the score of individual matches with the relative frequency of the same identification for a particular spectrum being assigned by more than one algorithm.

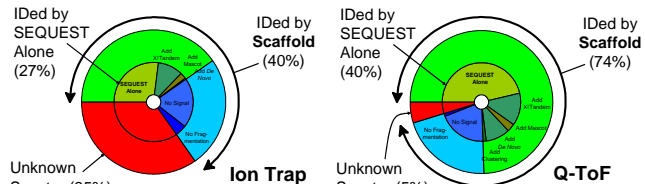


Figure 2: Percentage of MS/MS spectra identified by Scaffold. Green marks electronic noise or poor fragmentation, and red marks unidentified spectra. Scaffold identified approximately 50% more spectra than SEQUEST alone.

Testing The Approach With Shotgun Proteomics

Scaffold was tested using protein standards (Figure 2), which showed an increased identification rate of at least 50% more spectra when using Scaffold versus any single search engine. We also present how this increased peptide coverage can be useful in shotgun proteomics applications. In this study, Scaffold was used to interpret SEQUEST, Mascot, and X! Tandem results from soluble and insoluble human lens proteins separated using two-dimensional liquid chromatography and analyzed using quadrupole-time of flight (Q-ToF) and ion trap tandem mass spectrometers (6). Although crystallin proteins in the human lens have been previously examined using shotgun proteomics (6,7), Table 1 represents one of the first large catalogs of non-crystallin components in the lens. Using conservative probability thresholds (minimum of >99% protein probability and 2 peptide identifications >95%), 52 non-crystallin components were observed by Scaffold (49 in the Ion Trap dataset and 32 in the Q-ToF). Scaffold improved sensitivity by 20% and 19% in confidently identified proteins over SEQUEST alone for Ion Trap and Q-ToF datasets, respectively. Similarly, Scaffold provided a 32% and 100% gain over Mascot alone using the same criteria. Finally, by considering multiple technical replicates from different instruments, Scaffold can get somewhat deeper sample coverage than when considering only a single replicate.

Table 1: Confidently identified proteins (minimum requirement of >99% protein probability and 2 peptide identifications >95%) in both Ion Trap and Q-ToF analyses of the human lens proteome using Scaffold. The number of unique peptides identified from each protein are reported, as well as the total spectrum count in parentheses. Identifications marked in blue were missed by SEQUEST, whereas identifications marked in yellow were missed by Mascot. Identifications marked in green were missed by both. In Ion Trap and Q-ToF datasets, Scaffold improved sensitivity by 20% and 19% over SEQUEST alone and 32% and 100% gain over Mascot alone.

Protein Name	Accession Number(s)	Ion Trap			Q-ToF				
		Scaffold	SEQUEST	Mascot	X!Tandem	Scaffold	SEQUEST	Mascot	X!Tandem
1 Beta crystallin B1	CRB1_HUMAN	59 (1545)	23 (1163)	22 (749)	67 (927)	42 (1161)	23 (983)	18 (558)	39 (922)
2 Beta crystallin B2	CRB2_HUMAN	53 (698)	23 (476)	19 (348)	47 (401)	28 (278)	15 (228)	13 (119)	26 (191)
3 Beta crystallin B3	CRB3_HUMAN	38 (888)	18 (737)	7 (378)	33 (268)	27 (488)	14 (408)	7 (134)	24 (254)
4 Beta crystallin A1 (Contains: Beta crystallin A1)	CRBA_HUMAN	36 (1022)	15 (768)	15 (527)	34 (424)	32 (559)	19 (487)	12 (250)	29 (383)
5 Alpha crystallin B	CRAB_HUMAN	34 (840)	14 (494)	15 (243)	29 (256)	34 (540)	17 (425)	10 (282)	32 (486)
6 Alpha crystallin A	CRAA_HUMAN	32 (746)	15 (611)	13 (336)	27 (288)	31 (571)	14 (448)	9 (181)	28 (293)
7 Aldheyde dehydrogenase	BFDH1_HUMAN	31 (149)	20 (120)	16 (48)	22 (62)	22 (74)	15 (58)	7 (28)	18 (42)
8 Fibrinogen	FBN1_HUMAN	29 (178)	21 (102)	21 (100)	21 (95)	24 (136)	18 (110)	13 (80)	18 (142)
9 Beta crystallin A4	CRB4_HUMAN	22 (362)	10 (242)	9 (146)	21 (190)	16 (233)	10 (190)	8 (90)	15 (142)
10 Actin, cytoplasmic 2	ACTC2_HUMAN	20 (108)	14 (84)	13 (80)	19 (99)	9 (27)	8 (24)	5 (9)	8 (17)
11 Gamma crystallin C	CRGC_HUMAN	18 (164)	13 (135)	7 (52)	11 (85)	14 (90)	11 (58)	4 (40)	11 (59)
12 Gamma crystallin D	CRGD_HUMAN	13 (156)	11 (148)	8 (75)	10 (81)	14 (105)	11 (72)	7 (27)	11 (69)
14 Brain acid soluble protein 1	BASP_HUMAN	13 (38)	8 (28)	8 (24)	13 (28)	4 (8)	4 (8)	0 (0)	1 (1)
15 Gamma crystallin B	CRGB_HUMAN	12 (88)	9 (89)	6 (29)	8 (28)	11 (35)	9 (25)	2 (5)	6 (19)
16 Glyceraldehyde 3-phosphate dehydrogenase	GAPDH_HUMAN	11 (65)	7 (47)	9 (38)	10 (31)	10 (34)	9 (30)	6 (10)	9 (25)
17 Sorbitol dehydrogenase	SDH3_HUMAN	11 (37)	10 (34)	8 (17)	5 (11)	8 (19)	6 (18)	5 (10)	5 (14)
18 Fructose-bisphosphate aldolase A	PFKB_HUMAN	10 (28)	7 (16)	9 (11)	7 (10)	6 (13)	4 (8)	2 (4)	4 (6)
19 Alpha enolase	ENO3_HUMAN	10 (19)	8 (14)	7 (13)	8 (14)	6 (9)	5 (8)	4 (6)	6 (10)
20 Fructose-bisphosphate aldolase C	PFKB_HUMAN	9 (27)	5 (7)	4 (12)	6 (17)	4 (11)	2 (7)	4 (5)	4 (6)
21 Phosphoglycerate kinase 1	PFKP_HUMAN	9 (23)	8 (19)	1 (14)	6 (12)	6 (11)	2 (4)	2 (4)	5 (9)
22 Phosphoglycerate mutase 1	PFMG1_HUMAN	9 (20)	8 (18)	6 (8)	8 (12)	3 (8)	3 (5)	0 (0)	2 (3)
23 Actin, gamma-esteric smooth muscle	ACTA_HUMAN	8 (48)	8 (40)	7 (22)	8 (15)	4 (12)	4 (12)	2 (3)	4 (6)
24 Transketolase	TKT_HUMAN	8 (27)	4 (23)	5 (16)	7 (17)	4 (7)	4 (6)	4 (6)	3 (5)
25 Beta crystallin B3	CRB3_HUMAN	8 (12)	6 (5)	5 (7)	7 (10)	4 (8)	3 (6)	2 (3)	2 (4)
26 Vimentin	VIME_HUMAN	8 (12)	5 (5)	5 (7)	7 (10)	4 (8)	3 (6)	2 (3)	2 (4)
27 Pyruvate kinase, M2 isozyme	PKPF_HUMAN	7 (21)	6 (18)	5 (9)	5 (9)	5 (7)	4 (6)	2 (3)	5 (7)
28 Tubulin alpha-1 chain	TBA1_HUMAN	5 (18)	5 (18)	2 (2)	5 (8)	5 (8)	4 (8)	1 (1)	5 (7)
29 Heat-shock protein beta-1	HSPB1_HUMAN	5 (15)	5 (15)	4 (8)	4 (7)	1 (2)	1 (2)	1 (2)	1 (2)
30 Glutathione transferase	HSZP_HUMAN	5 (10)	5 (10)	4 (7)	4 (8)	4 (8)	3 (8)	2 (2)	4 (4)
31 Spectrin beta chain, brain 1	SPOC_HUMAN	5 (9)	5 (9)	4 (7)	4 (7)	2 (3)	1 (2)	1 (1)	2 (3)
32 Phosphatidylinositol-3-OH kinase class I	PIK3C_HUMAN	4 (15)	3 (13)	3 (11)	3 (9)	3 (8)	3 (8)	2 (3)	3 (7)
33 Carbonyl reductase (NADPH) 1	CAR1_HUMAN	4 (4)	3 (4)	3 (4)	4 (4)	3 (4)	3 (4)	1 (1)	3 (4)
34 Filamin A	FLNA_HUMAN	4 (7)	1 (2)	3 (4)	3 (5)	0 (0)	0 (0)	0 (0)	0 (0)
35 Spectrin alpha chain, brain 1	SPOC_HUMAN	4 (4)	1 (1)	2 (4)	2 (2)	4 (4)	2 (4)	0 (0)	1 (1)
36 4-hydroxyphenylpyruvate dioxygenase	HPPD_HUMAN	4 (4)	4 (4)	1 (1)	3 (3)	2 (2)	2 (2)	0 (0)	2 (2)
37 Ubiquitin carboxyl-terminal hydrolase isozyme 8	UBH1_HUMAN	4 (4)	3 (3)	2 (2)	2 (2)	0 (0)	0 (0)	0 (0)	0 (0)
38 Fatty acid-binding protein, epidermal	FABP4_HUMAN	3 (8)	2 (6)	2 (4)	3 (4)	1 (1)	1 (1)	1 (1)	1 (3)
39 Triosephosphate isomerase	TPI1_HUMAN	3 (8)	3 (6)	3 (6)	3 (6)	1 (1)	1 (1)	0 (0)	1 (1)
40 DJ-1 protein (Chacogen DJ-1)	PRK7_HUMAN	3 (8)	0 (0)	3 (8)	0 (0)	1 (2)	0 (0)	1 (2)	0 (0)
41 Nuclein-1 (nucleoprotein)	NUC1_HUMAN	3 (4)	2 (2)	2 (3)	2 (2)	2 (2)	2 (2)	0 (0)	2 (2)
42 Peroxisome biogenesis factor 1	PDXF_HUMAN	3 (4)	1 (1)	2 (3)	2 (3)	2 (3)	2 (3)	1 (1)	2 (3)
43 14-3-3 protein zeta	PPP1R1_HUMAN	3 (3)	2 (2)	1 (1)	2 (2)	1 (1)	1 (1)	0 (0)	2 (2)
44 Peptidyl-prolyl cis-trans isomerase A	CYP11_HUMAN	3 (3)	2 (2)	0 (0)	1 (1)	1 (1)	1 (1)	0 (0)	1 (1)
45 Beta crystallin A2	CRBA_HUMAN	3 (3)	3 (3)	2 (2)	1 (1)	1 (2)	1 (2)	1 (1)	1 (2)
46 Plasmodium falciparum inhibitor	ODI1_HUMAN	3 (3)	3 (4)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)
47 Cytidine deaminase	ODD_HUMAN	2 (5)	2 (5)	2 (4)	2 (4)	1 (1)	1 (1)	0 (0)	1 (1)
48 Ras GDP dissociation inhibitor alpha	GDI1_HUMAN	2 (4)	2 (4)	2 (2)	2 (4)	2 (2)	2 (2)	1 (1)	2 (2)
49 Galectin-1	GAL1_HUMAN	2 (4)	2 (4)	2 (2)	2 (2)	1 (1)	1 (1)	1 (1)	1 (1)
50 Vacuolar ATP synthase catalytic subunit A	VAAC_HUMAN	2 (4)	2 (4)	0 (0)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)
51 Ubiquitin	UBIQ_HUMAN	2 (4)	1 (1)	0 (0)	2 (2)	0 (0)	0 (0)	0 (0)	0 (0)
52 Myotrophin	MTPN_HUMAN	2 (4)	2 (4)	1 (1)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)
53 Susac syndrome dismutase	SODC_HUMAN	2 (3)	0 (0)	2 (3)	1 (1)	1 (1)	1 (1)	1 (1)	0 (0)
54 Plectin 1	PLE1_HUMAN	2 (3)	2 (3)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)
55 Guanine nucleotide exchange factor, mitochondrial precursor	GGEF_HUMAN	2 (3)	2 (3)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)
56 6-phosphofructokinase, liver type	PFKB_HUMAN	2 (3)	2 (3)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)
57 Malate dehydrogenase, cytoplasmic	MDH1_HUMAN	2 (3)	2 (3)	2 (2)	2 (2)	0 (0)	0 (0)	0 (0)	0 (0)
58 Gamma-aminobutyrate transaminase	GABA1_HUMAN	2 (3)	2 (3)	2 (2)	2 (2)	0 (0)	0 (0)	0 (0)	0 (0)
59 Glycogen phosphorylase, brain form	PHS3_HUMAN	2 (2)	0 (0)	1 (1)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)
60 NADP-dependent malic enzyme	ME3C_HUMAN	1 (4)	1 (4)	1 (2)	1 (1)	2 (3)	2 (3)	1 (2)	2 (2)
61 Phosphoglycerate migration inhibitory factor	PGIF_HUMAN	1 (2)	1 (2)	1 (2)	1 (2)	1 (1)	1 (1)	1 (1)	1 (2)
62 Glucose-6-phosphate isomerase	GPI_HUMAN	1 (1)	1 (1)	1 (1)	1 (1)	2 (3)	2 (3)	1 (1)	2 (3)
63 6-Carbonic hydratase	CA1A_HUMAN	1 (1)	1 (1)	0 (0)	0 (0)	1 (1)	1 (1)	1 (1)	1 (1)
64 Collagen alpha 1(VI) chain	C1A1A_HUMAN	0 (0)	0 (0)	0 (0)	0 (0)	2 (2)	2 (2)	0 (0)	1 (1)

Total Spectra Identified: 641 (7538) 396 (5734) 340 (3504) 523 (3663) 443 (4548) 302 (3790) 185 (1757) 374 (3060)

References

- (1) Eng, J. K., et al. *Am. Soc. Mass Spectrom.* 1994, 5, 976-989.
- (2) Perkins, D. N., et al. *Electrophoresis* 1999, 20, 3551-3567.
- (3) Field, H. I., et al. *Proteomics* 2002, 36, 47.
- (4) Keller, A., et al. *Anal. Chem.* 2002, 74, 5383-5392.
- (5) Nesvizhskii, A. I., et al. *Anal. Chem.* 2003, 75, 4646-4658.
- (6) Searle, B. C., et al. *J. Proteome Res.* 2005, 4, 546E-554.
- (7) MacKoss, M. J., et al. *Proc. Natl. Acad. Sci. U.S.A.* 2002, 99, 7900-7905

Scaffold Availability

A free trial version of Scaffold is available for download at www.ProteomeSoftware.com. Other availability inquiries should be made to Mark.Pitman@ProteomeSoftware.com

Acknowledgements

We thank Phil Wilmarth for helpful discussions as well as Srinvasa Nagalla and Larry David for providing us with the Q-ToF and Ion Trap MS/MS data sets used in this study.