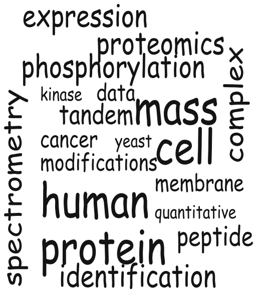
J. Am. Soc. Mass Spectrom. (2015) 26:1799–1803
DOI: 10.1007/s13361-015-1181-3

FOCUS: 20 YEAR ANNIVERSARY OF SEQUEST: ACCOUNT & PERSPECTIVE

The H-Index of 'An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database'

Michael P. Washburn[1,2]

[1]Stowers Institute for Medical Research, Kansas City, MO 64110, USA
[2]Departments of Pathology and Laboratory Medicine, University of Kansas Medical Center, Kansas City, KS 66160, USA



Abstract. Over 20 years ago a remarkable paper was published in the Journal of American Society for Mass Spectrometry. This paper from Jimmy Eng, Ashley McCormack, and John Yates described the use of protein databases to drive the interpretation of tandem mass spectra of peptides. This paper now has over 3660 citations and continues to average more than 260 per year over the last decade. This is an amazing scientific achievement. The reason for this is the paper was a cutting edge development at the moment in time when genomes of organisms were being sequenced, protein and peptide mass spectrometry was growing into the field of proteomics, and the power of computing was growing quickly in accordance with Moore's law. This work by the Yates lab grew in importance as genomics, proteomics, and computation all advanced and eventually resulted in the widely used SEQUEST algorithm and platform for the analysis of tandem mass spectrometry data. This commentary provides an analysis of the impact of this paper by analyzing the citations it has generated and the impact of these citing papers.

Keywords: SEQUEST, Tandem mass spectrometry, Proteomics, Yates

Received: 18 February 2015/Revised: 28 April 2015/Accepted: 29 April 2015/Published Online: 20 June 2015

Commentary

As of April 27, 2015 'An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database' [1] has 3667 citations according to www.scopus.com. From 2004 to 2014, this paper has averaged 265 citations per year, with 2014 data remaining incomplete. In 2001, the paper received more than 100 citations for the first time, and in 2004 the paper received more the 200 citations for the first time (Figure 1). Taking some liberties with the H-index [2] by assigning an H-index to a paper, this paper has an H-index of 185. This means 185 papers that cite this work published in 1994 each have at least 185 citations themselves. Seventeen of these 185 have been cited more than 1000 times, and 47 of these have been cited more than 500 times. Of the 17 that have been cited more than 1000 times, 15 are original research articles. The total number of citations of all citing papers is in excess of 200,000. A word cloud analysis of all the terms in all the titles of citing articles is shown in Figure 2a and a word cloud analysis of all the biological terms in all the titles of citing articles is shown in Figure 2b. Clearly, the work by Eng et al. [1] has had a lasting and enabling impact on the analysis of peptide fragmentation pattern interpretation, protein mass spectrometry, proteomics, and biological research. This paper laid the critical technical foundation that future work in peptide mass spectrometry and proteomics could build on, which is important to this day.

It is important to note that while Eng et al. [1] is widely associated with SEQUEST, the first time the term SEQUEST appeared in print was in a paper published in *Rapid Communications in Mass Spectrometry* entitled 'Direct database searching with MALDI-PSD spectra of peptides' [3], and this paper has been cited 71 times. The first time SEQUEST appeared in PubMed was in 1996 in another paper in *JASMS* [4], which has been cited 52 times. In between the two important papers in *JASMS*, the Yates lab published two papers in *Analytical Chemistry* that expanded on the landmark Eng et al. study in 1994 [1]. Both of these papers are highly cited in their own right with 298 citations for one [5] and 923 citations for the other [6]. A strong argument can be made that the less cited of these two papers, which describes a method to mine genomes by using a six-frame translation of a nucleotide sequence database to identify peptides [5], provided an

Correspondence to: Michael Washburn; *e-mail:* mpw@stowers.org

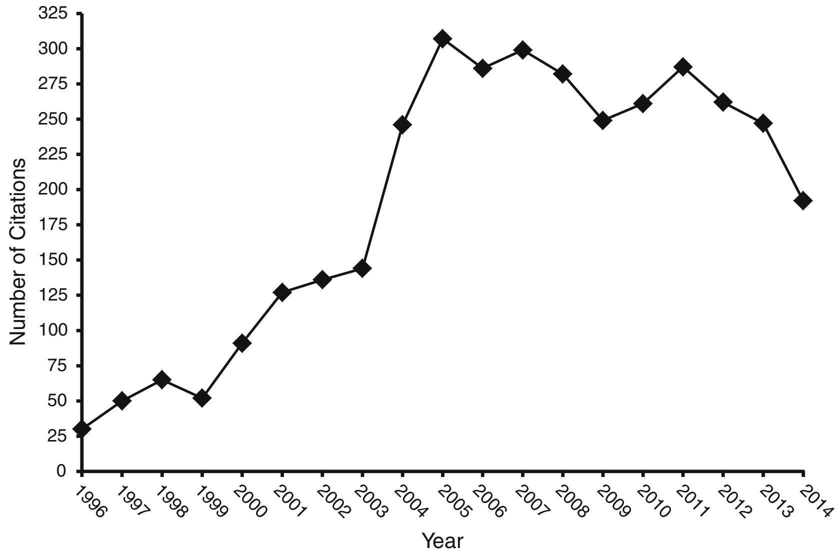


Figure 1. Citations per year for Eng et al. Shown are the citations from www.scopus.com as of April 27, 2015 with the number of citations per year for this landmark paper

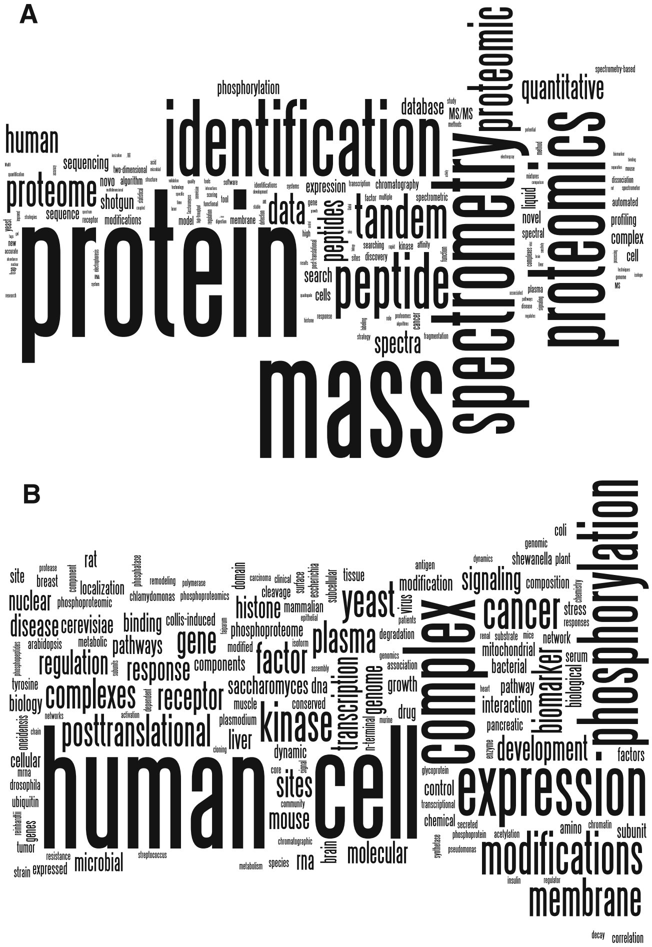


Figure 2. Word clouds of all titles of all citing papers. From www.wordle.net (**a**) shows the word cloud using all the terms in all the titles of all citing papers, and (**b**) shows the word cloud using only biological terms after the mass spectrometry and proteomics related terms have been removed

Table 1. Top 20 Cited Papers that Cite ‘An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database’ (with citations as of April 27, 2015 from www.scopus.com)

Authors	Title	Year	Journal	Citations	Reference
Aebersold, R., Mann, M.	Mass spectrometry-based proteomics	2003	Nature	3768	[7]
Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., Aebersold, R.	Quantitative analysis of complex protein mixtures using isotope-coded affinity tags	1999	Nature Biotechnology	3569	[10]
Washburn, M.P., Wolters, D., Yates, J.R.	Large-scale analysis of the yeast proteome by multidimensional protein identification technology	2001	Nature Biotechnology	3149	[18]
Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R.	Correlation between protein and mRNA abundance in yeast	1999	Molecular and Cellular Biology	2468	[11]
Keller, A., Nesvizhskii, A.I., Kolker, E., Aebersold, R.	Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search	2002	Analytical Chemistry	2361	[13]
Nesvizhskii, A.I., Keller, A., Kolker, E., Aebersold, R.	A statistical model for identifying proteins by tandem mass spectrometry	2003	Analytical Chemistry	1965	[14]
Hayashi, F., Smith, K.D., Ozinsky, A., Hawn, T.R., Yi, E.C., Goodlett, D.R., Eng, J.K., Akira, S., Underhill, D.M., Aderem, A.	The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5	2001	Nature	1880	[26]
Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., Yates, J.R. III	Direct analysis of protein complexes using mass spectrometry	1999	Nature Biotechnology	1732	[19]
Pandey, A., Mann, M.	Proteomics to study genes and genomes	2000	Nature	1506	[9]
Liu, H., Sadygov, R.G., Yates, J.R. III	A model for random sampling and estimation of relative protein abundance in shotgun proteomics	2004	Analytical Chemistry	1387	[21]
Elias, J.E., Gygi, S.P.	Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry	2007	Nature Methods	1356	[15]
Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J., Hunt, D.F.	Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry	2004	Proceedings of the National Academy of Sciences of the United States of America	1295	[24]
Ficarro, S.B., McCleland, M.L., Stukenberg, P.T., Burke, D.J., Ross, M.M., Shabanowitz, J., Hunt, D.F., White, F.M.	Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae	2002	Nature Biotechnology	1216	[23]
Wolters, D.A., Washburn, M.P., Yates, J.R. III	An automated multidimensional protein identification technology for shotgun proteomics	2001	Analytical Chemistry	1168	[20]
Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J., Gygi, S.P.	Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome	2003	Journal of Proteome Research	1071	[16]
Shevchenko, A., Jensen, O.N., Podtelejnikov, A.V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H., Mann, M.	Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels	1996	Proceedings of the National Academy of Sciences of the United States of America	1037	[25]
Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y., Aebersold, R.	Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology	2000	Proceedings of the National Academy of Sciences of the United States of America	1007	[12]
Beausoleil, S.A., Jedrychowski, M., Schwartz, D., Elias, J.E., Villén, J., Li, J., Cohn, M.A., Cantley, L.C., Gygi, S.P.	Large-scale characterization of HeLa cell nuclear phosphoproteins	2004	Proceedings of the National Academy of Sciences of the United States of America	991	[17]
Mann, M., Jensen, O.N.	Proteomic analysis of post-translational modifications	2003	Nature Biotechnology	939	[8]
Yates, J.R. III, Eng, J.K., McCormack, A.L., Schieltz, D	Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database	1995	Analytical Chemistry	923	[6]

underappreciated foundation for the current practice of proteogenomics. The second of these two papers is significant because it describes the analysis of covalently modified peptides, including post-translational modifications like phosphorylation [6]. Both proteogenomics and the analysis of post-translational modifications are important applications of modern proteomics, and these two papers played important early roles in these fields.

The top 20 most cited papers that cite Eng et al. are listed in Table 1. They range in citations from 923 to 3768. Three of these are highly cited review articles [7–9], and the rest are important original research papers. The most highly cited original research paper that cites Eng et al. is a publication from the Aebersold and Gelb labs that described the isotope coded affinity tag (ICAT) [10]. Although ICAT is not widely used today, this work published in *Nature Biotechnology* was a landmark study in quantitative proteomics that has garnered more than 3500 citations. Other highly cited papers from the Aebersold group that cite Eng et al. are their work on the comparison of mRNA and protein expression in yeast with 2468 citations [11], an evaluation of the limits of two-dimensional gel electrophoresis with 1007 citations [12], on statistical modeling of tandem mass spectrometry data from Keller et al. [13] with 2361 citations, and Nesvizhskii et al. [14] with 1965 citations. SEQUEST played a major role in enabling all these papers, which have become landmark papers in the field on their own. Additional highly cited papers in the top 20 came from the Gygi lab once he left the Aebersold lab and started his own group at Harvard. This includes a paper describing the widely used target-decoy strategy for minimizing false discovery rates in proteomics analyses [15] with 1356 citations, a paper on multidimensional chromatography in proteomics [16] with 1071 citations, and an analysis of HeLa cell nuclear phosphopeptides [17] with 991 citations.

The second most highly cited original research paper that cites Eng et al. is one of three papers from the Yates lab that describes the development of multidimensional protein identification technology (MudPIT) with 3149 citations [18]. I personally owe my scientific career to this paper, and this work would have not been possible without the prior development of SEQUEST. The first description of the MudPIT approach was the groundbreaking paper by Link et al. in 1999 [19] with 1732 citations, and another important paper taking an analytical approach to dissecting MudPIT and how it worked was by Wolters et al. in 2001 [20] with 1168 citations. These three papers all described the development of the direct coupling of a two-dimensional chromatography system that included strong cation exchange and reversed phase chromatography to a tandem mass spectrometer, which has proven to be a powerful technique for the analysis of proteomes [18–20]. All three papers were made possible by the availability of SEQUEST. Another paper from the Yates lab in the top 20 is from Liu et al. [21], which has 1387 citations. This paper was key to laying the foundation for the use of spectral counting in label-free quantitative proteomics [21]. Although spectral counting is controversial to some, when coupled with MudPIT, spectral counting is a highly effective and statistically robust way to carry out an experiment [22]. What is interesting to note is that from 1993 to 1999, the Aebersold and Yates labs were adjacent to each other in the Department of Molecular Biotechnology at the University of Washington. An amazing number of high impact and long lasting impact papers originated during this time and in this department.

Four original research papers remain from the top 20. Two are from the Hunt lab, which describes the IMAC approach for phosphopeptide enrichment [23] and electron transfer dissociation [24]. One is an important study from the Mann lab on the analysis of proteomes from two-dimensional gel electrophoresis [25], and the final paper is on the innate immune response [26]. In total, the top 20 papers citing Eng et al. have a combined total of 34,788 citations, and 28,575 of these are from the 16 original research articles in the top 20. What does the future hold for this paper? In terms of citations, they will only grow. It is likely that as of this writing, the 2014 citation data for the paper is incomplete and the paper will continue to have more than 200 citations per year for several years to come. Clearly, 'An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database' has had a profound impact on the practice of protein mass spectrometry and proteomics.

Acknowledgments

This work was supported by the Stowers Institute for Medical Research.

References

1. Eng, J.K., McCormack, A.L., Yates, J.R.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. **5**, 976–989 (1994)
2. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. U. S. A. **102**, 16569–16572 (2005)
3. Griffin, P.R., MacCoss, M.J., Eng, J.K., Blevins, R.A., Aaronson, J.S., Yates III, J.R.: Direct database searching with MALDI-PSD spectra of peptides. Rapid Commun. Mass Spectrom. **9**, 1546–1551 (1995)
4. Yates, J.R., Eng, J.K., Clauser, K.R., Burlingame, A.L.: Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides. J. Am. Soc. Mass Spectrom. **7**, 1089–1098 (1996)
5. Yates III, J.R., Eng, J.K., McCormack, A.L.: Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal. Chem. **67**, 3202–3210 (1995)
6. Yates III, J.R., Eng, J.K., McCormack, A.L., Schieltz, D.: Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal. Chem. **67**, 1426–1436 (1995)
7. Aebersold, R., Mann, M.: Mass spectrometry-based proteomics. Nature **422**, 198–207 (2003)
8. Mann, M., Jensen, O.N.: Proteomic analysis of post-translational modifications. Nat. Biotechnol. **21**, 255–261 (2003)
9. Pandey, A., Mann, M.: Proteomics to study genes and genomes. Nature **405**, 837–846 (2000)
10. Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., Aebersold, R.: Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. **17**, 994–999 (1999)
11. Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R.: Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. **19**, 1720–1730 (1999)

12. Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y., Aebersold, R.: Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc. Natl. Acad. Sci. U. S. A. **97**, 9390–9395 (2000)
13. Keller, A., Nesvizhskii, A.I., Kolker, E., Aebersold, R.: Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. **74**, 5383–5392 (2002)
14. Nesvizhskii, A.I., Keller, A., Kolker, E., Aebersold, R.: A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. **75**, 4646–4658 (2003)
15. Elias, J.E., Gygi, S.P.: Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods **4**, 207–214 (2007)
16. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J., Gygi, S.P.: Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J. Proteome Res. **2**, 43–50 (2003)
17. Beausoleil, S.A., Jedrychowski, M., Schwartz, D., Elias, J.E., Villen, J., Li, J., Cohn, M.A., Cantley, L.C., Gygi, S.P.: Large-scale characterization of HeLa cell nuclear phosphoproteins. Proc. Natl. Acad. Sci. U. S. A. **101**, 12130–12135 (2004)
18. Washburn, M.P., Wolters, D., Yates III, J.R.: Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. **19**, 242–247 (2001)
19. Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., Yates III, J.R.: Direct analysis of protein complexes using mass spectrometry. Nat. Biotechnol. **17**, 676–682 (1999)
20. Wolters, D.A., Washburn, M.P., Yates III, J.R.: An automated multidimensional protein identification technology for shotgun proteomics. Anal. Chem. **73**, 5683–5690 (2001)
21. Liu, H., Sadygov, R.G., Yates III, J.R.: A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal. Chem. **76**, 4193–4201 (2004)
22. Pavelka, N., Fournier, M.L., Swanson, S.K., Pelizzola, M., Ricciardi-Castagnoli, P., Florens, L., Washburn, M.P.: Statistical similarities between transcriptomics and quantitative shotgun proteomics data. Mol. Cell. Proteom. **7**, 631–644 (2008)
23. Ficarro, S.B., McCleland, M.L., Stukenberg, P.T., Burke, D.J., Ross, M.M., Shabanowitz, J., Hunt, D.F., White, F.M.: Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. Nat. Biotechnol. **20**, 301–305 (2002)
24. Syka, J.E., Coon, J.J., Schroeder, M.J., Shabanowitz, J., Hunt, D.F.: Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. **101**, 9528–9533 (2004)
25. Shevchenko, A., Jensen, O.N., Podtelejnikov, A.V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H., Mann, M.: Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two-dimensional gels. Proc. Natl. Acad. Sci. U. S. A. **93**, 14440–14445 (1996)
26. Hayashi, F., Smith, K.D., Ozinsky, A., Hawn, T.R., Yi, E.C., Goodlett, D.R., Eng, J.K., Akira, S., Underhill, D.M., Aderem, A.: The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. Nature **410**, 1099–1103 (2001)