J. Am. Soc. Mass Spectrom. (2015) 26:1814–1819
DOI: 10.1007/s13361-015-1201-3

The SEQUEST Family Tree

David L. Tabb

School of Medicine, Vanderbilt University, Nashville, TN 37232-8575, USA

Abstract. Since its introduction in 1994, SEQUEST has gained many important new capabilities, and a host of successor algorithms have built upon its successes. This Account and Perspective maps the evolution of this important tool and charts the relationships among contributions to the SEQUEST legacy. Many of the changes represented improvements in computing speed by clusters and graphics cards. Mass spectrometry innovations in mass accuracy and activation methods led to shifts in fragment modeling and scoring strategies. These changes, as well as the movement of laboratories and lab members, have led to great diversity among the members of the SEQUEST family.

Key words: SEQUEST, History, Intellectual property, Proteomics, Bioinformatics

Received: 30 January 2015/Revised: 14 May 2015/Accepted: 19 May 2015/Published Online: 30 June 2015

Database search algorithms are sufficiently ubiquitous in proteomics that the field is hard to imagine without this technology. At this time, more than 30 algorithms of this type have been published. These engines rely upon the same fundamental elements; they all read protein sequence databases, emulate enzymatic cleavage to peptides, extrapolate post-translational modifications (PTMs), require peptide masses to fall within a tolerance of observed precursor mass, predict fragment ions for each peptide sequence, and compare observed and expected fragments [1]. This Account and Perspective pulls back the curtain on the development of SEQUEST, the first of the database search algorithms [2], and it details both the evolution of that software over time and the relationship that later software packages bear to the original SEQUEST.

Achieving Version 1.0

As with most bioinformatics algorithms, SEQUEST had its origins in a cumbersome manual process. A seminal paper from Don Hunt et al. in 1986 illustrated the challenges of interpreting peptide tandem mass spectra [3]. John Yates, then a graduate student in the Hunt laboratory, began thinking of ways to apply computers in the process of spectral interpretation and built upon that experience during his early years as a faculty member [4]. Kevin Owens' 1992 review of correlation analysis in mass spectra [5] provided a mechanism by which tandem mass spectra could be compared with each other, and John Yates hired Jimmy Eng, an electrical engineer who had recently completed his Master's degree at the University of Washington, to begin software development in earnest.

SEQUEST was effective because of a series of shrewd judgment calls in software development. Sequence databases were miniscule, by today's standards (the *S. cerevisiae* genome was not completed until 1996 [6]). Dr. Yates, however, recognized early that using predicted protein sequences from genomic sequencing would drastically reduce the set of potential sequences to be compared with each tandem mass spectrum. Similarly, the group recognized that predicting the appearance of collision-induced dissociation (CID) tandem mass spectra accurately for peptide sequences was a daunting challenge, and they opted to employ very simple fragmentation models that predicted C-terminal y ions to be twice the intensity of N-terminal b ions. Each experimental spectrum was separated into 10 zones by *m/z*, with peak intensities normalized within each to make the experimental spectra look more like the theoretical ones. Finally, they recognized that cross-correlation required so much CPU power that a pre-scoring routine was necessary to retain only 500 candidate peptides for full scoring by cross-correlation. Taken together, these insights paved the way for fully automated peptide identification software.

Making SEQUEST widely available led through gates of intellectual property, commercialization, and publication. On March 14, 1994, the University of Washington filed for a pair of patents (US5538897A and US6017693A) that defined the use of database searching for amino acid and nucleotide sequences from tandem mass spectra collected in mixtures of proteins. In 1993, Dr. Yates had begun discussions with Adrian Land and Ian Jardine, researchers at Thermo Instrument Systems

Correspondence to: David L. Tabb; *e-mail:* david.l.tabb@vanderbilt.edu

(now Thermo Fisher Scientific), to commercially distribute the SEQUEST software. The University of Washington agreed to an exclusive license of the patents to Thermo Instrument Systems. Jim Shofstahl integrated the software into the DECUnix-based BioWorks for the TSQ 700 under the name "PepSearch" (the name "SEQUEST" was coined after the 1994 publication). At first, this appeared to be an ideal solution, but later implementations separated SEQUEST from BioWorks so that updates to the rapidly-changing SEQUEST could be incorporated more readily.

Publishing SEQUEST, however, proved to be a significant challenge. Initially, the manuscript was sent to the *Proceedings of the National Academy of Sciences*, but the reviewers found it to be a mismatch for the journal. Dr. Yates then turned to *Protein Science*, which consulted the same reviewers as for *PNAS* in order to speed the process of review. The speedy review, however, resulted in rejections there, as well. Dr. Yates consulted with his mentor, Don Hunt, who advised publication in the *Journal of the ASMS* after consulting with Michael Gross. *JASMS* received the manuscript along with its prior reviews, and the paper was accepted only 27 days after its receipt on June 29, 1994 [2]. Later in the same year, Mann and Wilm published the manual interpretation sequence-tagging approach to peptide identification [7]. That these two technologies were presented in the same year is no coincidence; tandem mass spectrometry was clearly the most promising data source for protein identification, and bioinformatics advances were critical to realizing its potential.

Interpreting SEQUEST results, of course, required additional tools. Thermo Instrument Systems had begun by licensing basic support tools, such as the "Display Ions" Peptide-Spectrum Match (PSM) viewer and "SEQUEST Summary" result table builder, from the University of Washington. They soon licensed the Harvard Proteomics Browser Suite (licensed as the SEQUEST Browser), a growing collection of scripts from the William S. Lane Laboratory [8]. These tools provided essential capabilities for the interpretation of data sets, such as the depth of protein sequence coverage in the Protein Report, between-experiment comparisons in IonQuest, and the recognition of variant peptide forms in MuQuest. The software assisted the manual interpretation of tandem mass spectra through the FuzzyIons tool [9] and combined SEQUEST scores for better discrimination in the ScoreFinal neural network. In several respects, the Suite prefigured later identification workflows such as the Trans-Proteomic Pipeline [10]. With these tools in place, the stage was set for large numbers of researchers to benefit from database searching.

Evolving New SEQUEST Capabilities and Applications

For the next 7 y, the Yates Lab worked closely with Thermo to update SEQUEST continuously with improvements (see Figure 1). The most essential boost came from the addition of "dynamic modifications" [11]. The software could be notified that certain amino acids may sometimes carry additional mass due to a post-translational modification (such as in a phosphorylation search, where Ser, Thr, or Tyr gain 79.97 Da). The initial searches with this feature were limited to dynamic PTMs on only two residues at a time (for context, the Intel Pentium Pro became available in late 1995). Soon thereafter, the number of modifiable residues was increased to three. With dynamic PTMs, SEQUEST came of age.

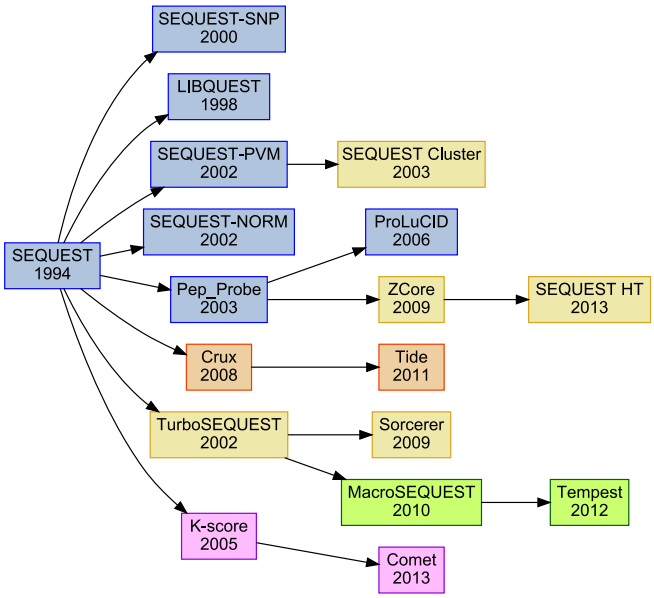

Figure 1. A tree representing the descendants of the original SEQUEST algorithm. Blue algorithms were produced in the Yates Laboratory, whereas yellow were produced in conjunction with commercial partners Thermo Fisher Scientific or Sage-N Research. Orange represents developments in the Noble Laboratory, with green denoting developments in the Gerber Laboratory, and purple marking advances from Jimmy Eng after the year 2000. An arrow does not imply direct use of source code

Early efforts in proteogenomics were also demonstrated in 1995 with the new ability to search nucleotide databases through six-frame translation [12]. Dr. Yates was able to demonstrate that protein identification was feasible using the chromosome II, III, and IX sequences produced by the in-progress *S. cerevisiae* genome project. This paper also supplied an early answer for how relative scoring can be used to determine which spectra had been successfully identified; the paper specified that PSMs in which the best match scored 10% better than the second (a ΔC_n or DeltaCN greater than 0.1) could be trusted. Leveraging genomic data would later be augmented in SEQUEST-SNP, which introduced non-synonymous single nucleotide polymorphisms to nucleotide databases for recognizing amino acid variants [13].

SEQUEST had become increasingly associated with the Thermo LCQ 3D ion trap after its release in 1996. The software was included in the new "XCalibur" interface for Windows NT. In an effort to make the software more broadly applicable,

the Yates Lab added the capability to look for a broader set of fragment ions associated with high-energy CID [14]; similarly, they examined post-source decay spectra as a source of identifications [15]. The Yates team also turned their efforts toward adapting the technique for spectral library searching. Their introduction of LIBQUEST [16] applied the same PSM scoring system from SEQUEST to the matching of previously identified spectra with recently collected MS/MS scans.

Algorithm Efficiency and Parallelization

Improving the speed of SEQUEST execution was a priority from early in development. Modifications to the initial C++ codebase targeted both Windows and UNIX platforms. Cross-correlation is a powerful match discriminator, but it requires a computationally expensive fast Fourier transform (FFT) operation. The initial implementation was based on code from *Numerical Recipes in C* [17] and from *Dr. Dobb's Journal*. At the time, floating-point performance for Intel processors was relatively slow, and so 64-bit DEC Alpha processors were investigated to improve execution. Over the next several years, though, Intel and AMD greatly improved performance by switching to 64-bit datapaths, accelerating math by operating on vectors of numbers, and increasing processor frequencies. These shifts have benefited SEQUEST performance even as MS/MS data sets have dramatically grown in size; during the time required for MS/MS scan rates to quintuple from an LCQ (1996) to an LTQ (2003), the number of transistors in Intel CPUs rose by an order of magnitude from the 200 MHz Pentium Pro (1995) to the 2.4 GHz Pentium 4 (2002).

The pressures to improve search speeds continued, however, and Yates Lab and Thermo worked together and separately to address the problem. Jimmy Eng and Bill Lane each worked on strategies for pre-indexing FASTA sequence databases to sort the masses of tryptic peptides prior to search. Jim Shofstahl adapted this code to produce indexes that could be exploited for PTM searches, releasing "TurboSEQUEST" in BioWorks 3.0. Jimmy Eng was able to leverage the Parallel Virtual Machine package from ORNL [18] to distribute the identification task across multiple computers, bridging between Windows master nodes and UNIX slave nodes. The SEQUEST-PVM software [19] was able to accelerate these searches by a factor that scaled linearly with the number of computers in the cluster. Jim Shofstahl at Thermo tuned this software for more robust operation, and end users were able to purchase SEQUEST Cluster licenses with BioWorks 3.1 in which IBM provided computers and Thermo contributed necessary software. Under license with Thermo, Sage-N Research produced "Sorcerer," an FPGA (field-programmable gate array) that had been configured to accelerate FFT in hardware [20]. Over time, Sage-N switched to source code optimizations in TurboSEQUEST to improve performance in ×86 systems provided by the company.

Thermo and the University of Washington have occasionally licensed the TurboSEQUEST source code to universities. Vanderbilt University, for example, compiled the source to produce specialized executables; the campus supercomputing facility employed IBM JS20 blades that used PowerPC 970 processors rather than ×86 or Alpha CPUs. The Gerber Lab at Dartmouth, however, had more ambitious ideas for their collaboration. Under their source license, the group produced MacroSEQUEST, a streamlined build of the software that searched all spectra simultaneously rather than using the spectrum-at-a-time approach of the original SEQUEST [21]. A key modification made in MacroSEQUEST allowed for users to adjust the FFT bin size, which permitted users of HCD (a collision cell fragmentation that is similar to beam-type CID) high-resolution tandem mass spectra to profit from high fragment mass accuracy in XCorr computation. The group continued their modifications in the Tempest project to off-load cross-correlation to a graphical processing unit (GPU) or employ extremely fast dot product computation for scoring instead [22].

Thermo, of course, has continued to invest in development. SEQUEST-HT, which became available as part of Proteome Discoverer 1.4, is a reimplementation of the TurboSEQUEST algorithm using the Microsoft .NET framework. SEQUEST-HT is multi-threaded to take advantage of multi-core CPUs, now commonplace. It benefits from sequence database management in the ZCore algorithm [23] along with its handling of ETD and HCD fragmentation with the small FFT bin sizes like those of MacroSEQUEST.

Search Engines from the Diaspora

After the Yates Lab moved from the University of Washington to The Scripps Research Institute in the year 2000, intellectual property issues prevented the group from producing new variants of the software and publishing them as SEQUEST (which is a trademark owned by the University of Washington). Similarly, Jimmy Eng moved to the Institute for Systems Biology, shifted to the Fred Hutchinson Cancer Research Center in 2004, and then returned to the University of Washington in 2007.

At Scripps, the Yates Lab was increasingly encountering very large data sets as it employed fractionated sample techniques such as MudPIT [24]. SEQUEST had been crafted to read individual MS/MS scans from DTA files and to write PSMs for individual spectra to OUT files, a strategy that led to significant file system problems when combining tens or hundreds of high-scan-rate LC-MS/MS experiments. Because of the bloat associated with XML file formats, the Yates Lab adopted delimited text formats for storing information: MS1 (mass spectra), MS2 (tandem mass spectra), and SQT (SEQUEST outputs) [25]. Thinking along similar lines, Jim Shofstahl had created the binary SRF format for storing DTA and OUT data structures for the commercial SEQUEST release. Support for the HUPO-PSI mzML format [26] was added to SEQUEST-HT via an importer in Proteome

Discoverer, and output from SEQUEST-HT can be converted to mzIdentML format [27] within that framework, as well.

Within the Yates Lab, peptide identification included insights from Michael MacCoss, Rovshan Sadygov, and Tao Xu. In 2002, Michael MacCoss introduced SEQUEST-NORM, a variant of the SEQUEST code that could produce peptide length-independent cross-correlation scores [28]. Dr. Sadygov incorporated the "Fastest Fourier Transform in the West" library into SEQUEST to accelerate FFT computation [29] and added a dot product score for use with accurate mass MS/MS scans, with the enthusiastic support of Michael Senko at Thermo Fisher Scientific. Dr. Sadygov's experiments on improving the pre-scoring routines of SEQUEST led to an altogether new search engine: Pep_Probe employed a hypergeometric distribution rather than cross-correlation as its primary match score [30]. This software was a useful test-bed for exploring other scoring functions. In 2005, Dr. Sadygov demonstrated the implementation in Pep_Probe of a scoring model based on accounting for larger fractions of total fragment ion intensity for an MS/MS, compared with cross-correlation and the original hypergeometric implementation [31]. After Dr. Sadygov was employed by Thermo Fisher Scientific, he turned those skills to the identification of ETD spectra. In collaboration with the Coon Laboratory at the University of Wisconsin-Madison, he published the ZCore algorithm, which combined his hypergeometric assessment of matched peak counts with an assessment of the matched fragment ion intensities [23].

Tao Xu produced the current algorithm employed for database search in the Yates Lab. ProLuCID employs the binomial distribution for determining the best 500 peptide sequences from a protein database and then applies cross-correlation to this set [32, 33]. The software predicts fragments with improved isotope models for better cross-correlation scoring discrimination. For each spectrum, ProLuCID determines the Z-score for the highest XCorr against the distribution produced by the top 500 candidates, determining the extent to which the best match falls outside the distribution produced by random matches. In addition to dynamic modifications on particular residues, the software adds the capability for peptide N-terminal and C-terminal modifications. ProLuCID is written in Java and can be deployed on individual computers or on Linux clusters.

At the Institute for Systems Biology, Jimmy Eng began work on the Comet search engine in 2001. At first, it took the form of a search engine that scored PSMs by inferring a Z-score from a distribution of dot-product scores instead of the costly cross-correlation of SEQUEST (much as Tempest did for HCD several years later) [10]. The approach gained broader use as the "K-score" in X!Tandem [34]. Upon his return to the University of Washington in 2007, Jimmy Eng returned to the SEQUEST scoring approach to discover a method by which FFT could be entirely bypassed in high-speed computation of cross-correlation scores [35], a technique incorporated into SEQUEST-HT. Four years later, he had written the Comet search engine from the ground up to support standard file formats for inputs and outputs, support a variety of activation methods, and distribute processing over multiple threads [36]. Comet reports expectation values that estimate how many PSMs might have been expected to score as well as the best match by random chance alone.

The University of Washington continued development in the SEQUEST family after the departure of the Yates Lab, principally in the laboratory of William Noble. Christopher Park introduced Crux in 2008 [37], featuring efficient peptide indexing for FASTA databases, on-the-fly decoy generation, and distribution fitting for top XCorrs. Crux paired efficiently with the Percolator algorithm from the Noble Lab for improved PSM discrimination [38]. Benjamin Diament applied a wide variety of optimization strategies to create the highly efficient Tide algorithm [39], showing considerable improvements in search times compared with 1993 and 2009 builds of SEQUEST and to Crux. Further refinements in 2014 enabled the calculation of accurate *P*-values from XCorr scores [40]. These tools were combined in the broader framework of the Crux Toolkit in 2014 [41].

Mapping the Future

The algorithms detailed above account for a substantial fraction of the publications in proteomics over the last two decades. The family would be even larger if the roster included algorithms that employ very different strategies than SEQUEST and yet compute XCorr scores [42]. SEQUEST has stood the test of time for two main reasons; cross-correlation has demonstrated itself to be an excellent discriminator in the presence of noise peaks, and a variety of fully automated processing pipelines can work from SEQUEST identifications to simplify determining which spectra were confidently identified and to assemble protein inferences from the peptide-spectrum matches. SEQUEST is one of many search engines, but it continues to command considerable mind-share.

In considering the search engines appearing in Figure 1, a reader might reasonably ask which algorithm is the "true" SEQUEST. John Yates contends that "SEQUEST is an approach." In effect, any algorithm that predicts spectra from database-derived peptides and compares the predictions to uninterpreted tandem mass spectra is following the SEQUEST paradigm. Just as the term "Xerox" has come to mean "to photocopy," one may reasonably "SEQUEST an LC-MS/MS experiment," even when employing an algorithm that never shared source with the original SEQUEST.

In the years since SEQUEST's publication, many search engines have been published, both within and without its lineage. The years 2013 and 2014, for example, saw the publication of Comet [36], EasyProt [43], Morpheus [44], MS Amanda [45], MS-GF+ [46], and Peppy [47]. Mass spectrometrists are faced with an embarrassment of riches. For a young bioinformaticist, however, the ability to make a distinctive mark by creating a faster, more flexible, or more accurate search engine for proteomics continues to diminish.

Helpfully, the fields of glycomics, lipidomics, and other systems biologies are awaiting a similar transformation.

Acknowledgment

D.L.T. was supported by U24 CA159988. He greatly appreciates interviews and/or comments from Jimmy K. Eng, Scott Gerber, Bill Lane, Bill Noble, Rovshan Sadygov, Jim Shofstahl, Tao Xu, and John R. Yates. He acknowledges Scott Wasson of TechReport.com for providing CPU technology insights, and Jay D. Holman for producing the graphical abstract image.

References

1. Eng, J.K., Searle, B.C., Clauser, K.R., Tabb, D.L.: A face in the crowd: recognizing peptides through database search. Mol. Cell. Proteomics **10**, R111.009522 (2011)
2. 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)
3. Hunt, D.F., Yates, J.R., Shabanowitz, J., Winston, S., Hauer, C.R.: Protein sequencing by tandem mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. **83**, 6233–6237 (1986)
4. Yates III, J.R., Griffin, P., Hood, L., Zhou, J.: Computer aided interpretation of low energy MS/MS mass spectra of peptides. Tech. Protein Chem. II. **46**, 477–485 (1991)
5. Owens, K.G.: Application of correlation analysis techniques to mass spectral data. Appl. Spectrosc. Rev. **27**, 1–49 (1992)
6. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G.: Life with 6000 genes. Science **274**(546), 563–567 (1996)
7. Mann, M., Wilm, M.: Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal. Chem. **66**, 4390–4399 (1994)
8. Chittum, H.S., Lane, W.S., Carlson, B.A., Roller, P.P., Lung, F.D., Lee, B.J., Hatfield, D.L.: Rabbit beta-globin is extended beyond its UGA stop codon by multiple suppressions and translational reading gaps. Biochemistry **37**, 10866–10870 (1998)
9. Lane, W.S., Eng, J., Yates, J.R., Baker, M.A.: Fuzzy ions: a web-based workbench for de novo MS/MS sequence interpretation of peptides. Proceedings of the ASMS Conference on Mass Spectrometry and Allied Topics, May 31–June 4, Orlando, FL, pp. 121–121 (1998)
10. Keller, A., Eng, J., Zhang, N., Li, X., Aebersold, R.: A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol. Syst. Biol. **1**, E1–E8 (2005)
11. Yates, 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)
12. Yates, 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)
13. Gatlin, C.L., Eng, J.K., Cross, S.T., Detter, J.C., Yates, J.R.: Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. Anal. Chem. **72**, 757–763 (2000)
14. 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)
15. Griffin, P.R., MacCoss, M.J., Eng, J.K., Blevins, R.A., Aaronson, J.S., Yates, J.R.: Direct database searching with MALDI-PSD spectra of peptides. Rapid Commun. Mass Spectrom. **9**, 1546–1551 (1995)
16. Yates, J.R., Morgan, S.F., Gatlin, C.L., Griffin, P.R., Eng, J.K.: Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. Anal. Chem. **70**, 3557–3565 (1998)
17. Press, W.H. (ed.): Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge (1992)
18. Geist, A. (ed.): PVM—Parallel Virtual Machine: a Users' Guide and Tutorial for Networked Parallel Computing. MIT Press, Cambridge (1994)
19. Sadygov, R.G., Eng, J., Durr, E., Saraf, A., McDonald, H., MacCoss, M.J., Yates, J.R.: Code developments to improve the efficiency of automated MS/MS spectra interpretation. J. Proteome Res. **1**, 211–215 (2002)
20. Lundgren, D.H., Martinez, H., Wright, M.E., Han, D.K.: Protein Identification Using Sorcerer 2 and SEQUEST. In: Baxevanis, A.D., Petsko, G.A., Stein, L.D., Stormo, G.D. (eds.) Current Protocols in Bioinformatics. John Wiley and Sons, Inc, Hoboken (2009)
21. Faherty, B.K., Gerber, S.A.: MacroSEQUEST: efficient candidate-centric searching and high-resolution correlation analysis for large-scale proteomics data sets. Anal. Chem. **82**, 6821–6829 (2010)
22. Milloy, J.A., Faherty, B.K., Gerber, S.A.: Tempest: GPU-CPU computing for high-throughput database spectral matching. J. Proteome Res. **11**, 3581–3591 (2012)
23. Sadygov, R.G., Good, D.M., Swaney, D.L., Coon, J.J.: A new probabilistic database search algorithm for ETD spectra. J. Proteome Res. **8**, 3198–3205 (2009)
24. Washburn, M.P., Wolters, D., Yates, J.R.: Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. **19**, 242–247 (2001)
25. McDonald, W.H., Tabb, D.L., Sadygov, R.G., MacCoss, M.J., Venable, J., Graumann, J., Johnson, J.R., Cociorva, D., Yates III, J.R.: MS1, MS2, and SQT—three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. Rapid Commun. Mass Spectrom. **18**, 2162–2168 (2004)
26. Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W.H., Römpp, A., Neumann, S., Pizarro, A.D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., Deutsch, E.W.: mzML—a community standard for mass spectrometry data. Mol. Cell. Proteomics. **10**, R110.000133 (2011)
27. Seymour, S.L., Farrah, T., Binz, P.-A., Chalkley, R.J., Cottrell, J.S., Searle, B.C., Tabb, D.L., Vizcaíno, J.A., Prieto, G., Uszkoreit, J., Eisenacher, M., Martínez-Bartolomé, S., Ghali, F., Jones, A.R.: A standardized framing for reporting protein identifications in mzIdentML 1.2. Proteomics **14**(21–22) 2389–2399 (2014)
28. MacCoss, M.J., Wu, C.C., Yates, J.R.: Probability-based validation of protein identifications using a modified SEQUEST algorithm. Anal. Chem. **74**, 5593–5599 (2002)
29. Frigo, M., Johnson, S.G.: FFTW: an adaptive software architecture for the FFT. Proceedings of the 1998 I.E. International Conference on Acoustics, Speech, and Signal Processing, May 12–15, pp. 1381–1384. (1998)
30. Sadygov, R.G., Yates, J.R.: A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal. Chem. **75**, 3792–3798 (2003)
31. Sadygov, R., Wohlschlegel, J., Park, S.K., Xu, T., Yates, J.R.: Central limit theorem as an approximation for intensity-based scoring function. Anal. Chem. **78**, 89–95 (2006)
32. Xu, T., Venable, J.D., Park, S.K., Cociorva, D., Lu, B., Liao, L., Wohlschlegel, J., Hewel, J., Yates J.R. III: ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. Molecular & Cellular Proteomics, pp. S174–S174. American Society of Biochemistry Molecular Biology Inc., Bethesda (2006)
33. Lu, B., Xu, T., Park, S.K., Yates, J.R.: Shotgun Protein Identification and Quantification by Mass Spectrometry. In: Reinders, J., Sickmann, A. (eds.) Proteomics, pp. 261–288. Humana Press, Totowa (2009)
34. MacLean, B., Eng, J.K., Beavis, R.C., McIntosh, M.: General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. Bioinformatics **22**, 2830–2832 (2006)
35. Eng, J.K., Fischer, B., Grossmann, J., MacCoss, M.J.: A fast SEQUEST cross correlation algorithm. J. Proteome Res. **7**, 4598–4602 (2008)
36. Eng, J.K., Jahan, T.A., Hoopmann, M.R.: Comet: an open-source MS/MS sequence database search tool. Proteomics **13**, 22–24 (2013)
37. Park, C.Y., Klammer, A.A., Käll, L., MacCoss, M.J., Noble, W.S.: Rapid and accurate peptide identification from tandem mass spectra. J. Proteome Res. **7**, 3022–3027 (2008)
38. Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., MacCoss, M.J.: Semisupervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods **4**, 923–925 (2007)
39. Diament, B.J., Noble, W.S.: Faster SEQUEST searching for peptide identification from tandem mass spectra. J. Proteome Res. **10**, 3871–3879 (2011)
40. Howbert, J.J., Noble, W.S.: Computing exact *P*-values for a cross-correlation shotgun proteomics score function. Mol. Cell. Proteomics **13**, 2467–2479 (2014)
41. McIlwain, S., Tamura, K., Kertesz-Farkas, A., Grant, C.E., Diament, B., Frewen, B., Howbert, J.J., Hoopmann, M.R., Käll, L., Eng, J.K., MacCoss,

M.J., Noble, W.S.: Crux: rapid open source protein tandem mass spectrometry analysis. J. Proteome Res. **13**, 4488–4491 (2014)
42. Dasari, S., Chambers, M.C., Codreanu, S.G., Liebler, D.C., Collins, B.C., Pennington, S.R., Gallagher, W.M., Tabb, D.L.: Sequence tagging reveals unexpected modifications in toxicoproteomics. Chem. Res. Toxicol. **24**, 204–216 (2011)
43. Gluck, F., Hoogland, C., Antinori, P., Robin, X., Nikitin, F., Zufferey, A., Pasquarello, C., Fétaud, V., Dayon, L., Müller, M., Lisacek, F., Geiser, L., Hochstrasser, D., Sanchez, J.-C., Scherl, A.: EasyProt—an easy-to-use graphical platform for proteomics data analysis. J. Proteome **79**, 146–160 (2013)
44. Wenger, C.D., Coon, J.J.: A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. J. Proteome Res. **12**, 1377–1386 (2013)
45. Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S., Mechtler, K.: MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. J. Proteome Res. **13**, 3679–3684 (2014)
46. Kim, S., Pevzner, P.A.: MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. **5**, 5277 (2014)
47. Risk, B.A., Spitzer, W.J., Giddings, M.C.: Peppy: proteogenomic search software. J. Proteome Res. **12**, 3019–3025 (2013)