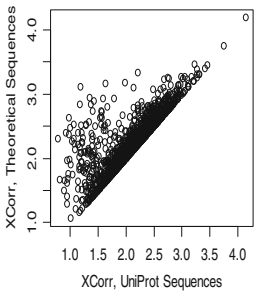
J. Am. Soc. Mass Spectrom. (2015) 26:1858–1864
DOI: 10.1007/s13361-015-1228-5

FOCUS: 20 YEAR ANNIVERSARY OF SEQUEST: RESEARCH ARTICLE

Using SEQUEST with Theoretically Complete Sequence Databases

Rovshan G. Sadygov[1,2]

[1]Department of Biochemistry and Molecular Biology, The University of Texas Medical Branch, Galveston, TX 77555, USA
[2]Sealy Center for Molecular Medicine, The University of Texas Medical Branch, Galveston, TX 77555, USA



Abstract. SEQUEST has long been used to identify peptides/proteins from their tandem mass spectra and protein sequence databases. The algorithm has proven to be hugely successful for its sensitivity and specificity in identifying peptides/proteins, the sequences of which are present in the protein sequence databases. In this work, we report on work that attempts a new use for the algorithm by applying it to search a complete list of theoretically possible peptides, a de novo-like sequencing. We used freely available mass spectral data and determined a number of unique peptides as identified by SEQUEST. Using masses of these peptides and the mass accuracy of 0.001 Da, we have created a database of all theoretically possible peptide sequences corresponding to the precursor masses. We used our recently developed algorithm for determining all amino acid compositions corresponding to a mass interval, and used a lexicographic ordering to generate theoretical sequences from the compositions. The newly generated theoretical database was many-fold more complex than the original protein sequence database. We used SEQUEST to search and identify the best matches to the spectra from all theoretically possible peptide sequences. We found that SEQUEST cross-correlation score ranked the correct peptide match among the top sequence matches. The results testify to the high specificity of SEQUEST when combined with the high mass accuracy for intact peptides.

Keywords: SEQUEST, Mass distribution of peptides, All theoretically possible peptides, De novo Peptide sequencing

Abbreviations Da Dalton; FDR False discovery rate; FFT Fast Fourier Transform; HCD Higherenergy collisional dissociation; mDa milliDalton; MS Mass spectrometry; PSM Peptide spectrum match; PTM Post-translational modification; Sp score Preliminary score; XCorr Cross-correlation score.

Received: 2 February 2015/Revised: 8 May 2015/Accepted: 17 June 2015/Published Online: 4 August 2015

Introduction

High throughput protein identification using tandem mass spectrometry coupled to liquid chromatography is a well-established and widely used technology for protein identification [1, 2]. The methodology has various implementations but can, in general, be classified into three major components, which are sample preparation (protein extraction, protein separation and digestion, peptide separation using chromatography) mass spectrometry and software for protein identification using tandem mass spectra and protein sequence databases [1]. The automated protein identification using software is a very important component in the methodology as the number of tandem mass spectra are in the tens of thousands and manual annotation of spectra is not feasible. SEQUEST [3] was one of the first database search engines developed to perform the task of the automating protein identification. Along with the other very few early search engines of the time, Mascot (probability based) [4], error tolerant [5], and high mass accuracy concept [6], it has contributed greatly to the development of the proteomics field and to its becoming widely accessible. Since the development of the original search engines, a number of new software have been developed that emphasized diverse and increasing needs of the field. We can only note a few: probabilistic OMSSA [7], X!Tandem [8], MyriMatch [9], Byonic [10], Inspect [11, 12], and high mass accuracy Andromeda [13]. The concepts of the probability-based peptide

Electronic supplementary material The online version of this article (doi:10.1007/s13361-015-1228-5) contains supplementary material, which is available to authorized users.

Correspondence to: Rovshan Sadygov; *e-mail:* rovshan.sadygov@utmb.edu

identifications and databases have also been employed for modeling protein identifications from intact protein fragmentations [14, 15]. One of the important features of SEQUEST is its multiple scoring criteria. At first, it filters the database peptide sequences for candidate peptides using enzymatic specificity and experimental precursor mass including its accuracy. For each candidate peptide, a preliminary score, Sp, is computed. Sp is fast and all database peptides meeting the mass filtering criterion are assigned Sp scores. In the second stage, a certain number (500 by default) of top Sp scoring peptides are used for cross-correlation analysis with the experimental spectrum, to generate XCorr. This step involves multiple fast Fourier transformations (FFTs) per candidate peptide and is normally slower than Sp scoring. To accelerate this process for high mass accuracy data where the mass arrays are large, FFT libraries referred to as the fastest FFT in the West were adapted into the SEQUEST [16]. The XCorr reports the correlation values between the experimental spectrum and theoretical peptide sequence, Sp scoring accounts for total (explained) ion current. The other score, ΔCn, is the difference between the XCorr of a peptide and the highest ranked peptide, normalized by the XCorr of the latter. As the database sizes increased and more candidate sequences were correlated against the experimental spectra, it became necessary to provide a probability of a peptide identification being a true/false positive. A large number of research papers have explored different statistical approaches to employ the SEQUEST scores to assign the probability of false or true match [17–19]. SEQUEST-identified peptides have been used for further bioinformatics confirmations of post-translational modifications (PTMs), such as phosphorylations [20–23]. In brief, SEQUEST has stimulated a large number of studies in bioinformatics and statistical approaches to automate and advance protein identification, PTM determination, quantification, and many other diverse applications of the proteomics. This is reflected in the number of citations of the original SEQUEST paper, which is currently the most cited article in the JASMS. It has been serving as an inspiration for bioinformatics software development in the field of proteomics, metabolomics, and other research areas using mass spectrometry-based high throughput sequencing. In this issue of JASMS, Dr. David Tabb provides a comprehensive chronicle of the SEQUEST development and multiple software that it has influenced. Recent review papers describe protein identification [24] and interpretation of mass spectra [25].

In this paper, we report on our findings in using SEQUEST for a de novo-like sequencing. Originally, SEQUEST was designed as a database search engine to identify peptides from their tandem mass spectra and protein sequence databases. Here we adapt the algorithm for a small scale sequencing of all theoretically possible peptides by making use of our algorithm for generating amino acid compositions of all theoretically possible peptides from their intact masses and the mass accuracy of intact peptides [26, 27]. We sought to find out how SEQUEST scoring of a true match would fair with the large number of peptides that are analyzed in an unbiased de novo-like sequencing [28–35]. Our secondary purpose was to find out how large XCorrs can be obtained. The approach may also contribute to false discovery rate control [36, 37] based on the use of decoy databases.

In the Methods section, we describe the workflow and generation of theoretical peptide sequences. The Results section describes the application of the approach to study more than 1400 tandem mass spectra from a publicly available data set [38].

Methods

We start with identification of peptide sequences using their tandem mass spectra and protein sequence databases (UniProt) [39] utilizing SEQUEST. Then, given the mass of an intact peptide and the enzymatic specificity of protein digest, we generate the list of all theoretically possible amino acid compositions. The compositions are converted into peptide sequences using lexicographic ordering. The peptide sequences for each precursor mass are assembled into a theoretical database of candidate sequences. SEQUEST is used to search the theoretical database of sequences with the tandem mass spectra of the peptide. The procedure essentially amounts to the de novo-like sequencing—without consideration for PTMs.

Generating Theoretical Peptide Sequences

Here, we briefly review the procedure for generating peptide sequence for a given mass interval (determined by the mass of peptide and the mass accuracy of the measurement). A peptide is a sequence of letters from a 20-letter alphabet A, the letters of which correspond with the 20 amino acids. This sequence is a realization from a composition represented by a numerical vector (a_1, a_2, … , a_{20}), whose jth component is the number of occurrences of the jth letter (amino acid) in the sequence, j = 1, 2, … , 20. The number of the amino acid compositions of peptides of length L is given by the Bose-Einstein statistics:

$$\binom{N+L-}{L} = \frac{(N+L-1)!}{L!(N-1)!}$$

The number of all sequences of length L, with a given composition, is a multinomial coefficient:

$$\frac{L!}{a_1!a_2!\ \ldots\ a_N!}$$

and the number of all distinct sequences is N^L. Here N (=20) is the number of amino acids in the alphabet. The formulas are used to confirm the accuracy of the algorithms for determining amino acid compositions and the following sequence generations.

We have previously used our algorithm to build and study the mass distribution of all theoretically possible peptides [40] and applied them to distinguish phosphopeptides from unmodified peptides [41]. The

algorithm accounts for the digest specificity and number of missed cleavages. Here, we used this algorithm to generate amino acid compositions for all sequences, the mass of which fits the mass of an intact peptide with a given mass accuracy. The compositions are then used by a lexicographic algorithm to generate all possible unique peptide sequences from the compositions. Since the number of sequences is very large (20^L), we made use of the mass degeneracy of the Lue and Ile by using Lue only to reduce the complexity of theoretical databases. This effectively reduces the number of amino acids to 19. We used full trypsin digest specificity with no missed cleavages. To reduce the complexity, we have considered only peptides with intact mass less than 1200 Da, and have assumed mass window of 0.002 Da (2 mDa) centered on the precursor mass.

We used SEQUEST to search the theoretical sequence databases and identify the best matches to the spectra. Then we compared these peptides with the results that SEQUEST has identified from the UniProt database. No PTMs were considered in this study. Mass accuracy was 1 mDa for precursor ions. Figure 1 summarizes the workflow used in this study.

Results

To evaluate our approach, we used spectra obtained from first strong anion exchange fraction of MCF7 cell line, 20100719_Velos1_TaGe_SA_MCF7_01.raw [37]. The mass spectra were acquired using Orbitrap Velos, the product ions were generated using higher energy collisional dissociation (HCD). As mentioned above, because of the computational complexities, we have limited the range of peptides to those with masses less than 1200 Da. Only +2 charged peptides were considered. For each peptide, we then created a separate FASTA database of all theoretical peptide sequences that fit a 2 mDa mass window around the peptide's mass. We then used these databases in SEQUEST searches to determine the best match to the corresponding tandem mass spectra. In total, there were 1400 spectra in the data set.

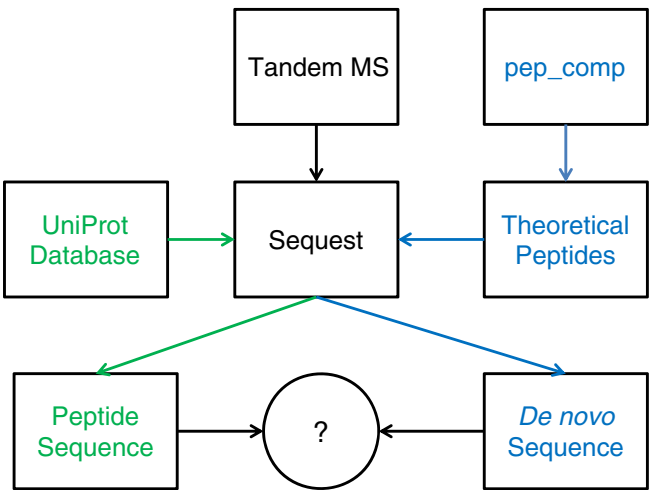


Figure 1. The workflow of the SEQUEST peptide identification using theoretically complete peptide sequences. The green colored path indicates normal database search procedure that SEQUEST is used for. The blue path indicates the generation of theoretical peptides, creation of the theoretical FASTA database, and de novo like sequence identification with SEQUEST

An example of the results is the peptide sequence, GAGTDDHTLIR, from human protein Annexin A5, with UniProt ID P08758. It has the mass of (monoisotopic mass of the amino acid sequence plus the mass of proton) 1155.57528 Da. SEQUEST identifies this peptide with XCorr value of 2.71. We used the peptide composition algorithm [26] to generate all amino acid compositions in the mass range of [1155. 574, 1155.576] Da. There were 802 unique compositions (after accounting for the Leu and Ile degeneracy). Using lexicographic ordering, from the compositions we generated a new peptide sequence database, specifically for this peptide. The size of the database was about 9 Gb. It had more than 600,000 candidate peptides for the spectrum. The best scoring peptide among the theoretical peptides was QGTDDHTLLR. It had an XCorr of 2.75. No other theoretical peptide sequence scored higher than the true peptide sequence, GAGT DDHTLIR. We note that the two sequences differ only on the prefix, "Q" in theoretical peptide versus "GA" in the true peptide. The annotated spectrum of the peptide is shown in Figure 2. Most of the y-ions of the peptide were observed in the tandem mass spectrum.

The peptide SGGGGGGGGSSWGGR of heterogeneous nuclear ribonucleoprotein A0, UniProt ID Q13151, was one of the higher mass peptides with the mass of 1192.50899 Da. It had XCorr value of 4.22. The [1192.508, 1192.510] Da mass interval was used to generate theoretical peptide compositions for this peptide. There were 983 unique compositions. After converting the compositions to sequences, the database size of the theoretical peptides exceeded 16 Gb. It had more than 1.2 million candidate sequences. The best scoring peptide among the theoretical peptides was the sequence, GSGGGGGGGS SWNR. It had XCorr score of 4.2. SEQUEST correctly identified this peptide among all theoretically possible peptides for this tandem mass spectrum. In this case as well, we see that there is long subsequence, GGGG GGGSSW, common to the true peptide and best scoring theoretical peptide sequences.

Table 1 summarizes the results for a sample of six spectra that were used in this study. The peptides that we have chosen did not have very high XCorr values, in general. In spite of this, SEQUEST produced results where the true peptides were always amongst the top highest scoring peptides in the large, unbiased databases comprising all theoretically possible peptides. This testifies to high specificity of SEQUEST when combined with the high mass accuracy for intact peptides. Among the small number of peptides in this table, the misassignments by SEQUEST included replacement of Ala and Gly by Gln,

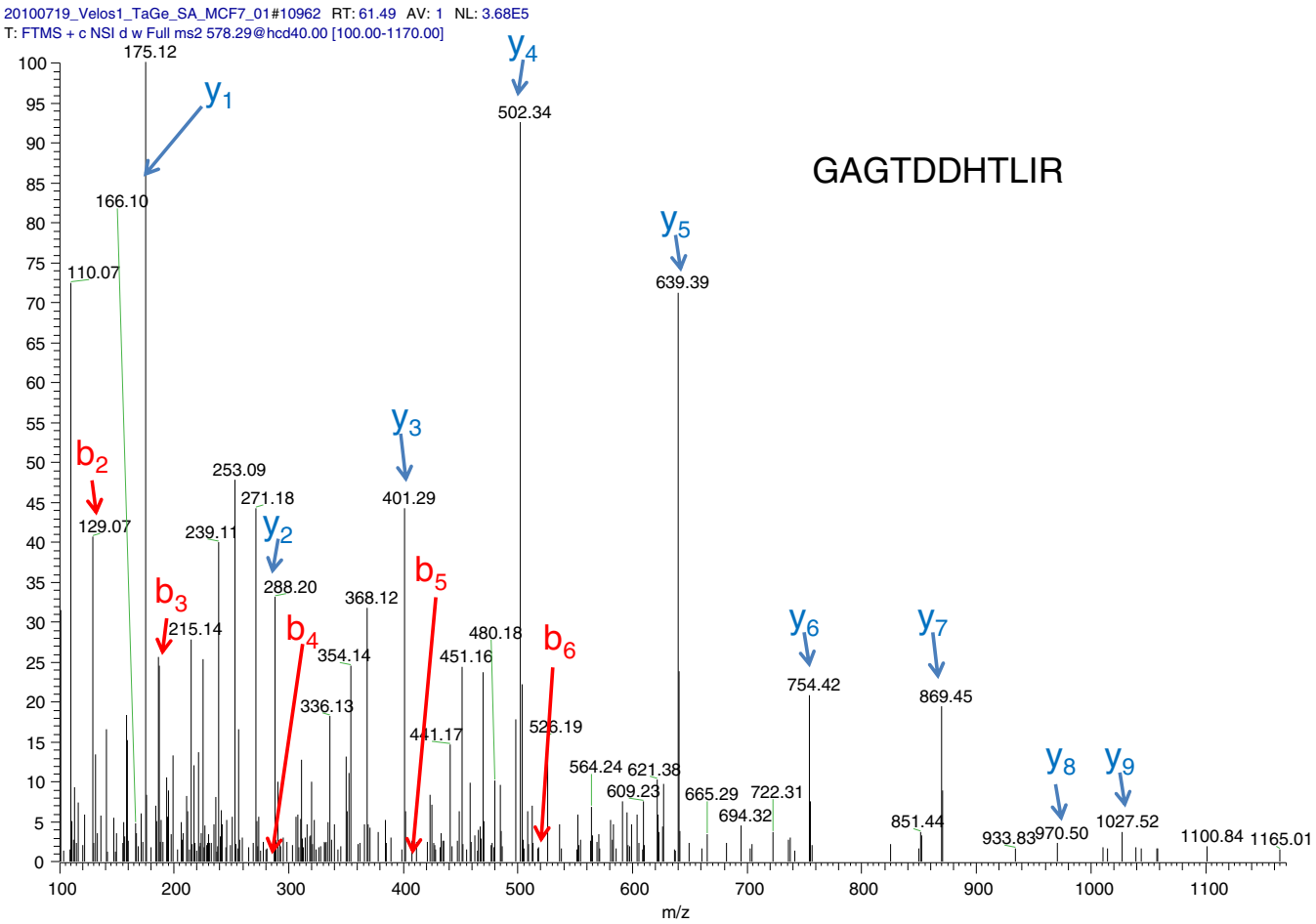


Figure 2. Annotated HCD spectrum of the peptide GAGTDDHTLIR. The blue color indicates y-ions and the red color indicates the b-ions. All y-ions, except for y_{10} have been observed in the spectrum. The XCorr value of this peptide was 2.71. The search of the all theoretically possible peptides using SEQUEST returns a slightly different sequence as the highest XCorr peptide, QAGTDDHTLIR, XCorr = 2.75. This was the only theoretical peptide to score higher than the true peptide

two Glys by Asn, and in some cases, amino acid scrambling.

In Figure 3, we show the scatter plot of XCorrs computed for the peptides identified from UniProt and theoretical sequence databases for all of the spectra used in this study (1413 spectra). For 465 spectra (~33% of all spectra) the sequences identified from the theoretical and UniProt databases were identical (as mentioned above, we did not differentiate between Leu and Ile). In addition, 157 peptide sequences (11% of the total) in UniProt and the corresponding theoretical peptides had the same amino acid compositions. The complete list of all scan numbers, identified sequences, and their XCorrs are provided in the Supplementary Materials. The XCorrs for theoretical peptides are always higher than or equal to the corresponding values for UniProt database peptides. For SEQUEST identifications, an important value has been the

Table 1. Summary for the Peptide Sequences, Their Tandem MS Scan Numbers (from the raw file 20100719_Velos1_TaGe_SA_MCF7_01.raw [37]), and Corresponding XCorrs

Peptide	Scan	Mass[b]	XCorr[a]	Theoretical peptide
GSGGGSSGGSIGGR	5202	1092.503	3.76/3.73	GSGGGSSGGSLNR
SGGGGGGGGSSWGGR	6946	1192.509	4.22/4.2	GSGGGGGGGSSWNR
GAGTDDHTLIR	10962	1155.575	2.71/2.75	QGTDDHTLLR
LGSLVENNER	13339	1130.580	2.23/2.31	LGSLVENGGER
IVQMTEAEVR	15962	1175.609	2.5/2.68	VLAGMTEEAVR
LTMQVSSLQR	18267	1162.624	2.5/2.4	TLAMGVSSGALR

Underlined are the common subsequences between the actual and theoretical peptide sequences. All precursors were +2 charged. All true peptides were among the three highest scoring peptides in their respective SEQUEST searches against the theoretical peptide databases

[a]The XCorrs are for the true peptide (the first score) and the best scoring theoretical peptide (the second score)

[b]Shown is the mass of a peptide's monoisotopic mass plus the proton mass

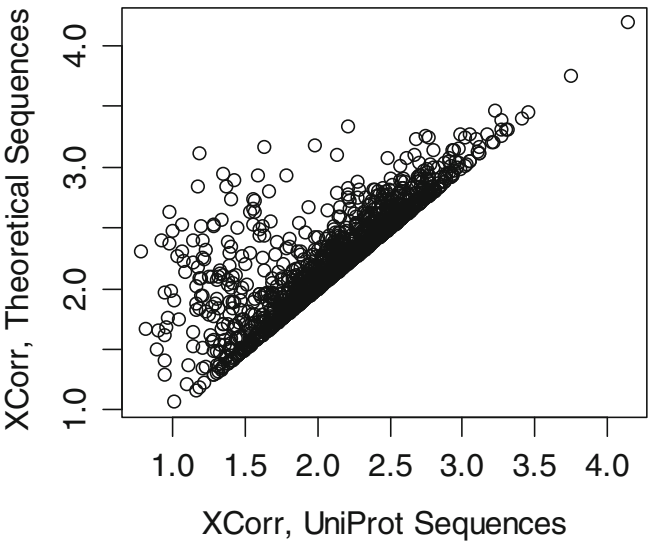


Figure 3. The scatter plot of the XCorr values for the theoretical and UniProt sequences. There were 465 (from the total of 1413) spectra for which the theoretical and database sequences were identical

ΔCn. This is the XCorr difference between the two highest ranked sequences, scaled by the XCorr of the highest ranked sequence. In Figure 4, we show the distribution for a similar value, which is the XCorr difference between the theoretical, $XCorr^{TH}$ and UniProt, $XCorr^{Uni}$, database peptides, scaled by the XCorr of the theoretical peptide. The overall correlation between the $XCorr^{TH}$ and $XCorr^{Uni}$ was 0.82 (Pearson's correlation). Pearson's correlation coefficient between the adapted ΔCn and $XCorr^{TH}$ is very small, 0.06, as can be seen from Figure 4.

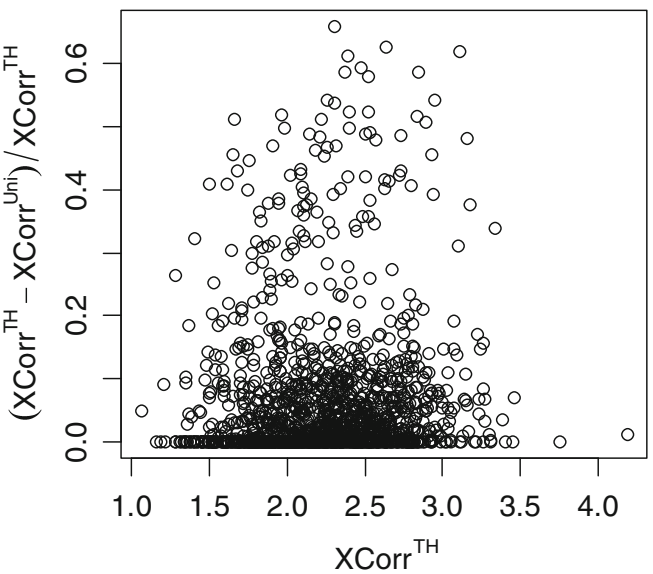


Figure 4. The distribution of the XCorr differences between the theoretical, $XCorr^{TH}$, and UniProt, $XCorr^{Uni}$, sequences scaled by the theoretical sequence's XCorr

We compared the two results from the two sequencing strategies when a combined (forward and reverse) database is used to control the false discovery rate (FDR) in the database searching. For this small dataset, 643 peptide spectrum matches (PSMs) passed the 1% FDR threshold; 202 of these PSMs had identical sequences to those obtained from the de novo-like sequencing; 87 of these PSMs (passing 1% FDR threshold) had identical amino acid compositions, thus differing only by amino acid scrambling from the corresponding sequences identified in our approach. Among the rest of the PSMs filtered at 1% FDR, there were 80 sequences that had subsequences of at least three amino acids long that were common to both results. We note again that the size of the dataset is very small, and while FDR filtering helps to control some erroneous matches, the distribution of XCorrs is not likely to represent the true sample distribution for this system. We also tested using ΔCn as a cutoff criterion (ΔCn > 0.1) in addition to FDR. The relative statistics of the PSMs identified in the de novo-like sequencing and database searching did not change substantially (about 5%).

Combined, forward and reverse, database searching is commonly used to control false discovery rate in large-scale peptide identifications. As the peptide size increases, normally in the species specific protein sequence databases, there are less peptides with the similar mass, particularly when precursor masses are determined in high resolution and mass accuracy instruments. The current study accounted for all possible theoretical peptides, as it generated a comprehensive list of all peptides. We used a smaller mass window, 2 mDa, centered on the peptide mass to control the size of the theoretical databases. In most of the cases that we studied, there were long common subsequences between the best theoretical match and the true peptide match. The long common subsequence is important as Blast searches of the theoretical peptides will likely map to correct proteins if the common subsequences (with the true peptides) are long. The study shows that for relatively short peptides (<1200 Da), peptide mass accuracy is very important and it will lead to correct peptide identifications even if the protein sequence database is unbiased (nonspecies-specific) and very large (includes all theoretically possible peptides).

We note that in the current implementation of this approach, there are large computational resource requirements. It is possible to automate the approach and generate the theoretical sequence databases on the "fly." However, the databases are still large and the computation takes considerably longer time compared with the regular database search.

Conclusions

We have implemented a workflow that allowed us to use SEQUEST scoring techniques for a de novo-like peptide identification. For every spectrum search, we have generated sequences of all possible peptides, using the intact peptide mass with the mass accuracy of 1 mDa. For a given mass interval (centered on intact peptide's mass) we first determined all

possible compositions. From the compositions, we generated all theoretical sequences using a lexicographic ordering. SEQUEST then was used to search the theoretically created database against the experimental spectrum. We have applied this approach to peptides with a mass less than 1200 Da. We found that when used with high mass accuracy for intact peptide mass, SEQUEST was highly specific; 33% of peptides identified in the theoretical sequence databases were the same as the corresponding original sequences in UniProt. In general, only a few theoretical sequences scored higher than the true peptide sequence in each case. In many cases, there were long common subsequences between the theoretically identified sequences and the true peptides. The current results testify to the high specificity of SEQUEST.

Acknowledgments

The author acknowledges support by the National Institute of General Medical Sciences of the National Institutes of Health under award no. R01GM112044.

References

1. Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.C., Yates III, J.R.: Protein analysis by shotgun/bottom-up proteomics. Chem. Rev. **113**(4), 2343–2394 (2013)
2. Walther, T.C., Mann, M.: Mass spectrometry-based proteomics in cell biology. J. Cell Biol. **190**(4), 491–500 (2010)
3. 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**(11), 976–989 (1994)
4. Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S.: Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis **20**(18), 3551–3567 (1999)
5. Mann, M., Wilm, M.: Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal. Chem. **66**(24), 4390–4399 (1994)
6. Clauser, K.R., Baker, P., Burlingame, A.L.: Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal. Chem. **71**(14), 2871–2882 (1999)
7. Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., Bryant, S.H.: Open mass spectrometry search algorithm. J. Proteome Res. **3**(5), 958–964 (2004)
8. Craig, R., Beavis, R.C.: TANDEM: matching proteins with tandem mass spectra. Bioinformatics **20**(9), 1466–1467 (2004)
9. Tabb, D.L., Fernando, C.G., Chambers, M.C.: MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J. Proteome Res. **6**(2), 654–661 (2007)
10. Bern, M., Kil, Y.J., Becker, C.: Byonic: advanced peptide and protein identification software. Curr. Protoc. Bioinformatics (2012). doi:10.1002/0471250953.bi1320s40.
11. Tanner, S., Pevzner, P.A., Bafna, V.: Unrestrictive identification of post-translational modifications through peptide mass spectrometry. Nat. Protoc. **1**(1), 67–72 (2006)
12. Tanner, S., Shu, H., Frank, A., Wang, L.C., Zandi, E., Mumby, M., Pevzner, P.A., Bafna, V.: InsPecT: identification of post-translationally modified peptides from tandem mass spectra. Anal. Chem. **77**(14), 4626–4639 (2005)
13. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., Mann, M.: Andromeda: a peptide search engine integrated into the MaxQuant environment. J. Proteome Res. **10**(4), 1794–1805 (2011)
14. Meng, F., Cargile, B.J., Miller, L.M., Forbes, A.J., Johnson, J.R., Kelleher, N.L.: Informatics and multiplexing of intact protein identification in bacteria and the archaea. Nat. Biotechnol. **19**(10), 952–957 (2001)
15. Johnson, J.R., Meng, F., Forbes, A.J., Cargile, B.J., Kelleher, N.L.: Fourier-transform mass spectrometry for automated fragmentation and identification of 5-20 kDa proteins in mixtures. Electrophoresis **23**(18), 3217–3223 (2002)
16. Sadygov, R.G., Zabrouskov, V.: Database search of high mass resolution data. J. Biomol. Tech. **18**(6), 1 (2007)
17. 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**(20), 5383–5392 (2002)
18. Anderson, D.C., Li, W., Payan, D.G., Noble, W.S.: A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J. Proteome Res. **2**(2), 137–146 (2003)
19. Kall, L., Canterbury, J.D., Weston, J., Noble, W.S., MacCoss, M.J.: Semisupervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods **4**(11), 923–925 (2007)
20. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J., Gygi, S.P.: A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat. Biotechnol. **24**(10), 1285–1292 (2006)
21. Taus, T., Kocher, T., Pichler, P., Paschke, C., Schmidt, A., Henrich, C., Mechtler, K.: Universal and confident phosphorylation site localization using phosphoRS. J. Proteome Res. **10**(12), 5354–5362 (2011)
22. Savitski, M.M., Lemeer, S., Boesche, M., Lang, M., Mathieson, T., Bantscheff, M., Kuster, B.: Confident phosphorylation site localization using the Mascot Delta Score. Mol. Cell. Proteomics **10**(2), M110 (2011)
23. Vandenbogaert, M., Hourdel, V., Jardin-Mathe, O., Bigeard, J., Bonhomme, L., Legros, V., Hirt, H., Schwikowski, B., Pflieger, D.: Automated phosphopeptide identification using multiple MS/MS fragmentation modes. J. Proteome Res. **11**(12), 5695–5703 (2012)
24. 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**(11), R111 (2011)
25. Ma, B., Johnson, R.: De novo sequencing and homology searching. Mol. Cell. Proteomics **11**(2), O111 (2012)
26. Nefedov, A.V., Mitra, I., Brasier, A.R., Sadygov, R.G.: Examining troughs in the mass distribution of all theoretically possible tryptic peptides. J. Proteome Res. **10**(9), 4150–4157 (2011)
27. Nefedov, A.V., Sadygov, R.G.: A parallel method for enumerating amino acid compositions and masses of all theoretical peptides. BMC Bioinformatics **12**(1), 432 (2011)
28. Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G.A., Ma, B.: PEAKS DB: De Novo sequencing assisted database search for sensitive and accurate peptide identification. Mol. Cell. Proteomics **11**(4), M111.010587 (2011)
29. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G.: PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. **17**(20), 2337–2342 (2003)
30. Kim, S., Gupta, N., Bandeira, N., Pevzner, P.A.: Spectral dictionaries: integrating de novo peptide sequencing with database search of tandem mass spectra. Mol. Cell. Proteomics **8**, 53–69 (2009)
31. Frank, A.M., Savitski, M.M., Nielsen, M.L., Zubarev, R.A., Pevzner, P.A.: De novo peptide sequencing and identification with precision mass spectrometry. J. Proteome Res. **6**(1), 114–123 (2007)
32. Johnson, R.S., Taylor, J.A.: Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. Mol. Biotechnol. **22**(3), 301–315 (2002)
33. Chi, H., Sun, R.X., Yang, B., Song, C.Q., Wang, L.H., Liu, C., Fu, Y., Yuan, Z.F., Wang, H.P., He, S.M., Dong M.Q.: pNovo: de novo peptide sequencing and identification using HCD spectra. J. Proteome Res. **9**(5), 2713–2724 (2010)
34. Tabb, D.L., Saraf, A., Yates III, J.R.: GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal. Chem. **75**(23), 6415–6421 (2003)
35. Yan, Y., Kusalik, A.J., Wu, F.X.: NovoHCD: de novo peptide sequencing from HCD spectra. IEEE Trans. Nanobiosci. **13**(2), 65–72 (2014)
36. Moore, R.E., Young, M.K., Lee, T.D.: Qscore: an algorithm for evaluating SEQUEST database search results. J. Am. Soc. Mass Spectrom. **13**(4), 378–386 (2002)
37. Elias, J.E., Gygi, S.P.: Target-decoy search strategy for mass spectrometry-based proteomics. Methods Mol. Biol. **604**, 55–71 (2010)
38. Geiger, T., Wehner, A., Schaab, C., Cox, J., Mann, M.: Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but

varying expression of most proteins. Mol. Cell. Proteomics **11**(3), M111 (2012)
39. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al.: The universal protein resource (UniProt). Nucleic Acids Res. **33**(Database issue), D154–D159 (2005)
40. Mitra, I., Nefedov, A.V., Brasier, A.R., Sadygov, R.G.: Improved mass defect model for theoretical tryptic peptides. Anal. Chem. **84**(6), 3026–3032 (2012)
41. Sadygov RG: Use of singular value decomposition analysis to differentiate phosphorylated precursors in strong cation exchange fractions. Electrophoresis **35**(24), 3498–3503 (2014)