Published in final edited form as:
Anal Chem. 2011 August 15; 83(16): 6135–6140. doi:10.1021/ac2009806.

A statistical method for assessing peptide identification confidence in accurate mass and time tag proteomics

Jeffrey R. Stanley[†], **Joshua N. Adkins**[‡], **Gordon W. Slysz**[‡], **Matthew E. Monroe**[‡], **Samuel O. Purvine**[‡], **Yuliya V. Karpievitch**[‡], **Gordon A. Anderson**[‡], **Richard D. Smith**[‡,*], and **Alan R. Dabney**[†,*]

[†]Department of Statistics, Texas A & M University, College Station, TX 77840

[‡]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352

Abstract

Current algorithms for quantifying peptide identification confidence in the accurate mass and time (AMT) tag approach assume that the AMT tags themselves have been correctly identified. However, there is uncertainty in the identification of AMT tags, as this is based on matching LC-MS/MS fragmentation spectra to peptide sequences. In this paper, we incorporate confidence measures for the AMT tag identifications into the calculation of probabilities for correct matches to an AMT tag database, resulting in a more accurate overall measure of identification confidence for the AMT tag approach. The method is referred to as Statistical Tools for AMT tag Confidence (STAC). STAC additionally provides a Uniqueness Probability (UP) to help distinguish between multiple matches to an AMT tag and a method to calculate an overall false discovery rate (FDR). STAC is freely available for download as both a command line and a Windows graphical application.

Introduction

One of the unavoidable uncertainties inherent in mass spectrometry (MS)-based protein studies is the potential for incorrect peptide and protein identifications. As a result, statistically-based measures for estimating identification confidence are a critical foundation for quantitative proteomic studies. In liquid chromatography-tandem mass spectrometry (LC-MS/MS) approaches, a number of algorithms have been developed for estimating the confidence of individual peptide identifications obtained by searching spectra against a theoretical database.[1–5] Probability confidence metrics use peptide sequence search results[1–3] as inputs to algorithms,[6–9] while other scores may come directly from a search tool[2–5] or from secondary algorithms.[10,11] In addition, Protein-Prophet[12] and other approaches[2,9] compute protein-level confidence measures. Far less progress has been made towards assessing the confidence of identifications obtained using high resolution high mass measurement accuracy LC-MS-based approaches such as the accurate mass and time (AMT) tag strategy[13,14] or similar implementations[15–18] that employ mass and elution time as coordinates in the peptide identification process.

*To whom correspondence should be addressed rds@pnl.gov; adabney@stat.tamu.edu.

The authors state no conflict of interest.

Supporting Information Available

Mathematically detailed derivations of the equations presented in this paper are available in the Supplemental Materials. This material is available free of charge via the Internet at http://pubs.acs.org/.

In the high throughput AMT tag strategy, peptides are identified from high-resolution LC-MS spectra by matching measured mass and elution time features to a reference peptide database of theoretical masses and observed normalized elution times (NET).[19,20] The databases are generally populated with peptides identified using conventional software tools such as SEQUEST[1] or X!Tandem[2] and empirical cutoffs[21–23] used to control the number of false entries. The false discovery rate (FDR) of matching LC-MS features to the AMT tag database may be estimated by shifting the LC-MS features by 11 Da in a method similar to a decoy database search[24] or by using a histogram of the mass errors.[25] Statistical alternatives based on the Mahalanobis distance[26] and error mixture models[27] exist and allow scoring of individual identifications. However, in any of these FDR-estimation approaches, the AMT tag entries themselves are assumed to have been correctly identified. This means that the uncertainty associated with the LC-MS/MS-based identification process is not accounted for in standard FDR estimates for matches to an AMT tag database.

The method described in this paper, Statistical Tools for AMT tag Confidence (STAC), incorporates PeptideProphet posterior probabilities of correct LC-MS/MS peptide sequence identifications into the calculation of the confidence in matching LC-MS features to the resulting AMT tags. Experimental data, characterized by a measured mass and NET, is used to train a mixture model that calculates the confidence of the match to the database while simultaneously accounting for the probabilities that the theoretical mass and observed NET in the database represent a particular peptide. STAC is an evolution of an earlier algorithm referred to as SMART.[28] In addition to individual posterior probabilities of correct matches, we provide methodology for estimating the overall FDR, allowing for an interpretable evaluation of the overall quality of a collection of identified peptides. Finally, a specificity score is defined for discriminating between matches to multiple peptides that can result when matching to densely-populated AMT tag databases.

Methods

An LC-MS feature is correctly identified only when it matches closely to an AMT tag, and the AMT tag itself is ascribed the appropriate peptide sequence. In earlier work by our laboratory, Jaitly, et al.[28] proposed the SMART algorithm for incorporating PeptideProphet probabilities of correct AMT tag identification into calculations of posterior probabilities of correct LC-MS-based matches to the AMT tags. STAC represents an evolution of the original SMART algorithm, with the following enhancements: (1) Various tuning parameters have been modified to optimize performance and minimize computation time; (2) In addition to peptide-specific posterior probabilities of correct matches, we return an overall FDR estimate; and (3) We report a specificity score for discriminating between ambiguous matches to multiple peptides.

A match probability, which may be used to assess the confidence in the identification of its peptide sequence, will be calculated for each potential LC-MS feature. We will then use these probabilities as part of a score to distinguish high scoring matches and to calculate the FDR.

Statistical Tools for AMT tag Confidence (STAC)

Both AMT tags derived from LC-MS/MS analyses and features detected in LC-MS analyses can be described by an ordered pair that consists of NET and monoisotopic mass. Because the goal is to match LC-MS features to an AMT tag, our method considers the distance between any feature observed at (m_{obs}; t_{obs}) and an AMT tag located at (m_{amt}, t_{amt}). The difference in the match of a feature to an AMT tag is defined by the ordered pair $\delta = (\delta_m; \delta_{net})$, where $\delta_{net} = t_{obs} - t_{amt}$ is the unitless NET difference and $\delta_m = (m_{obs} - m_{amt})/m_{amt} * 1_E6$ is the mass error measured in parts per million (ppm). The differences in the NET and

mass for this matching are approximately normally distributed.[27,29,30] The distribution of (δ_m, δ_{net}) is approximately uniform over a reasonable tolerance ($\ll$100 ppm)[25] when matched against random, incorrect points. These differences account for experimental variation, but do not take into account the variability present when building the AMT tag database.

Matches to the AMT tag database may be classified as being in one of three states:

- d_1 – Correctly matched to the database and the AMT tag is assigned the correct peptide sequence.
- d_2 – Correctly matched to the database but the AMT tag is assigned an incorrect peptide sequence.
- d_3 – Incorrectly matched to the database.

This state is unknown for each match, so we use the Expectation-Maximization (EM) algorithm to compute the parameters of our model. Let α be the unknown proportion of correct matches to the AMT tag database. Let β_j be the LC-MS/MS-based probability that the *j*th AMT tag entry has been assigned the correct peptide sequence; in what follows, we use PeptideProphet posterior probabilities to estimate the β parameters. Using the above attributes, a three part mixture model is constructed. For each of the states d_1 and d_2, we assume bivariate Normal distributions, with mean μ, covariance matrix Σ and mean μ', covariance matrix Σ', respectively. For state d_3, we assume a bivariate uniform distribution on $[-\text{tol}_m, \text{tol}_m] \times [-\text{tol}_{net}, \text{tol}_{net}]$.

According to Bayes rule, the probability that match *i*, $i = 1, \ldots, n$, is correctly identified (i.e., that match *i* is in state d_1), is

$$\text{STAC}_i = \frac{\alpha\beta_i \text{P}_{d_1}(\delta_i;\theta)}{\alpha\beta_i \text{P}_{d_1}(\delta_i;\theta) + \alpha(1-\beta_i)\text{P}_{d_2}(\delta_i;\theta') + (1-\alpha)\text{P}_{d_3}(\delta_i)}, \quad (1)$$

where $\theta = \{\mu, \Sigma\}$ and $\theta' = \{\mu', \Sigma'\}$. Each of θ, θ', and α are unknown parameters of the probability model and are estimated via the EM algorithm. The β_i are estimated by the maximum PeptideProphet probability of correct sequence assignment for the AMT tag entry involved in this match. Plugging the resulting model estimates into Eq. (1) yields the STAC score for match *i*.

Note that if we assume all AMT tags have been correctly assigned to sequences, we have $\beta_i = 1$ for each AMT tag and the equation reduces to that of May et al. [27] It should also be pointed out that the STAC score is limited by the values of the β_i, so that we may be no more confident in the identification of the LC-MS feature than we have previously been in the associated LC-MS/MS identifications.

Uniqueness Probability (UP) and FDR

Because a feature may match to multiple AMT tags, a measure of specificity, referred to as the Uniqueness Probability (UP), is incorporated. UP weights AMT tag matches using two criteria–the STAC score and the number of times an AMT tag has been observed in the MS/MS analyses used in building the AMT tag database. Detailed explanations are given in the Supplemental Materials. Briefly, for any feature with multiple matches to the database, all high confidence matches are within a small area. Over this small area, the STAC scores are approximately uniformly distributed for random matches. A statistical test of these values would be based upon the null hypothesis that all of the matches are incorrect, and therefore randomly uniform. Utilizing the maximum probability with which an AMT tag has been

identified gives us a Dirichlet distribution as our null model. We use the marginal distributions of this null model to determine weights from each match's STAC score. UP is then the weighted probability of the match being correct as compared to other potential matches to the same peptide.

UP performs much the same function as the delta score in SEQUEST[1] and other algorithms. Using only matches with UP>0.5 filters out ambiguous identifications of a single LC-MS feature. Filtering matches at a higher UP has the same result as increasing the delta requirement, since both leave less ambiguous matches.

Both the STAC score and UP are match-level assessments which may be aggregated for analysis level results. When considering the additional level of identification error from LC-MS/MS analyses, we may obtain a more accurate FDR than is otherwise possible. Matching back to an AMT tag database can have no lower FDR than the database itself has, but with the use of STAC, we evaluate FDR based upon the AMT tags used rather than the entire database.

Through the use of Bayes rule, the STAC score is a posterior probability of being a correct identification. This means that we may calculate the expected FDR for a given STAC cutoff, c, by

$$\mathrm{FDR}=\frac{\sum_{i:\mathrm{STAC}_i>c}(1-\mathrm{STAC}_i)}{n_c}, \tag{2}$$

where STAC_i is the STAC score for match i and n_c is the number of matches at cutoff c. The numerator is the expected number of false matches for the cutoff value c. Eq. (2) considers all matches, so it is slightly conservative for densely populated databases. UP may be used as an additional criteria for the sum to reflect a more precise FDR calculation.

Data

The STAC algorithm was evaluated on two data sets. The first data set is used to show the validity of the distributional assumptions and consists of a representative Quality Control *Shewanella oneidensis* (QCShew) analysis matched to an internal standard AMT tag database. This data set is additionally used to demonstrate the performance of the FDR in a common setting. As an internal quality control, QCShew provides both a robust AMT tag database and data from a variety of preparations and instruments. The QCShew sample was prepared with a global trypsin digestion and the data was obtained using an LTQ-Orbitrap (Thermo Fisher Scientific, San Jose, CA). The QCShew database contains SEQUEST search results of ion trap MS/MS data, filtered using standard Washburn / Yates rules ($\Delta C_n \geq 0{:}10$, and XCorr ≥ 1:9 for 1+, 2.2 for 2+, or 3.75 for 3+ and higher charge states[21,22]), yet also requiring that the AMT tags were observed in at least 10 MS/MS spectra.

The second set of data consists of 16 high mass measurement accuracy QCShew LC-MS/MS analyses which are used to show the performance of the STAC algorithm. The samples were prepared with a global trypsin digestion and were analyzed on a high mass accuracy LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, San Jose, CA). VIPER[31] was used to extract LC-MS features from the parent spectra, and matching was performed against a more strict database. DeconMSn[32] was used to gain additional accuracy in the identification of the LC-MS/MS spectra. This restrictive AMT tag database was built from SEQUEST[1] results that met the above criteria for ΔC_n and XCorr and were identified in at least 4 of the 16 analyses. With this database, we are able to exert a greater level of control

over the population of the database as well as provide an experimental standard by which to examine our algorithm.

Because of its widespread use, PeptideProphet[6] scores are calculated for each AMT tag, and these are used to estimate the β_i in Eq. (1) for the analyses below. The datasets described above were also aligned to their respective AMT tag databases using LCMSWARP[29] to improve the quality of matching to the database. The data are available at http://omics.pnl.gov.

Software Implementation of STAC

The STAC algorithm was implemented as both a command-line and a Windows graphical user application, STACker. Minimally, the user supplies two files: a text file containing AMT tag data, and a text file containing experimentally-derived LC-MS feature data. Optionally, the user may adjust STAC-related parameters, such as the matching tolerances tol_m and tol_{net} and filters for the AMT tag database. Multiple LC-MS feature file inputs are also possible within STACker, allowing automated processing of hundreds of data files. For every LC-MS feature file input, the STAC algorithm creates two text files: one containing the STAC score for each match, and a second file showing aggregate FDR data, including the number of matches at a given STAC cutoff and the corresponding FDR value. Both the command-line application and STACker are available at http://omics.pnl.gov/software/STAC.php.

Results and Discussion

Assessment of distributional assumptions

To begin the evaluation of the STAC algorithm, we must first check the validity of the distributional assumptions. Figure 1 shows a scatterplot and marginal histograms of the mass and NET errors when identifying representative QCShew data. The data were matched to the AMT tag database using tolerances of 20 ppm in mass and 0.05 in NET. Figure 1(a) shows the clustering of the data at the center of the scatterplot. Points with STAC scores greater than 0.90 are shown in black to emphasize the cluster. The histograms shown in Figure 1(b) and (c) show the mass and NET differences, respectively. All matches with a positive probability of being correct are shown in black. Though neither distribution is perfectly Normal, both appear to be reasonably so. The red bars in Figure 1(b) and (c) represent the random noise in the matching and appear to have approximately the same height across the regions, indicating that the uniformity of noise is a reasonable assumption.

Evaluation of STAC

Next, we compare STAC scores obtained from a representative high quality QCShew dataset to the previously established 11 Da shift method.[24] The 11 Da shift is an empirical method similar to a decoy database search and is based on all shifted matches being randomly incorrect. All of the AMT tag masses are shifted by 11 Da so that the masses are different by an amount not possible through modification, and all matches with a defined tolerance should then be truely random. Figure 2 shows the central region of this analysis; all matches are shown in red, high confidence STAC matches in black, and the tolerance bounds of the region used for the 11 Da shift is shown by an ellipse. The tolerance region used for the 11 Da shift, within the blue ellipse, is chosen to be the same area as that over which high quality STAC matches were determined. Since we constrained our analysis to a small dataset and a carefully constructed database, we expect very few false discoveries. This is indeed the case as the FDR for the 11 Da shift, calculated as number of decoy matches/(number of decoy and real matches),[22] is 0.95% compared to 0.54% calculated by

Eq. (2) for STAC. Eq. (2) allows control over the cutoff which is used, but it remains constant in its form unlike FDR calculations for decoy database searches.[22,24,33]

The STAC score performs well, as demonstrated above, when matching a data set to its own LC-MS/MS identifications, but even for this strict example there were cases of LC-MS features matching to more than one AMT tag identified in the LC-MS/MS searches. The red crosses shown within the blue ellipse are either ambiguous matches (UP≤0.5) that are less likely than another match or are features that match back to AMT tags that were not identified with sufficient confidence in the LC-MS/MS analyses. The impact of ambiguous points can also be seen in Figure 3(a) which shows the result of matching the representative QCShew data set used for checking the distributional assumptions to a much more densely populated database.

Figure 3 shows the effect of filtering out non-unique matches, as given by UP<0.5. The problematic points in Figure 3(a) are solely related to ambiguity and are marked in red to contrast with the unique high confidence matches shown in black. The STAC score takes into account low LC-MS/MS identification confidences, but UP should also be used to account for ambiguity. The extensively developed QCShew database used for matching is filtered to remove low confidence AMT tags, but the database is still densely populated. The effect of UP on FDR is shown in Figure 3(b). STAC alone is shown as a black line while STAC+UP is shown as the blue line. The red lines, indicating 0.85 and 0.90 STAC cutoffs, show that as we accept more matches, the FDR rises more quickly without accounting for the ambiguity of the matches. The use of STAC+UP provides an effective means for assessing match quality and for controlling the FDR of an analysis, permitting greater numbers of matches at a given FDR.

The Statistical Tools for AMT tag Confidence (STAC) algorithm presented here provides a statistical method for assessing peptide identification confidences in AMT tag proteomics. STAC combines confidence measures from LC-MS/MS and LC-MS analyses into a single confidence measure. The algorithm also provides a Uniqueness Probability (UP) for disambiguation of identifications and a means for calculating the false discovery rate, which may be used as an assessment of overall dataset quality. Though STAC, as demonstrated here, uses PeptideProphet[6] probabilities, it is readily adaptable to incorporate other LC-MS/MS identification probabilities.

Uses and future directions

STAC provides a means to assess the quality of matches to an AMT tag database, and the use of only higher quality matches improve downstream analyses. Using only high quality STAC matches removes less certain identifications and improves the quality of the data in quantitative studies, where misidentifications may confound significant results. STAC may also be used as the probability of peptide identification for protein identification algorithms, such as ProteinProphet.[12] In this way, STAC serves as a bridge to carry LC-MS/MS search results to higher-level evaluations. The use of STAC for downstream analyses can therefore improve results whether conclusions are drawn at the peptide or protein level.

The STAC algorithm combines information from multiple sources into a single confidence metric. Further refinements are envisioned for improved performance and account for additional relations between the variables. A reduction in the number of model assumptions, such as the independence of the mass and NET errors, would improve the robustness of the algorithm. As shown in Figure 1, the data are not always exactly Normal, so the use of more flexible (non-parametric) models should be given consideration. Consideration should also be given to methods of rolling up quantitative peptide level results for protein inferences with STAC used to weight intensities within a protein or for ambiguous proteins.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Drs. Roslyn Brown and Kristin Burnum, and Navdeep Jaitly for helpful discussions and assistance with parts of the data analysis. This research was supported by the NIH National Center for Research Resources (RR18522) and the National Institute of Allergy and Infectious Diseases NIH/DHHS through Interagency agreement Y1-AI-8401 and award number U54AI081680. Work was performed in the Environmental Molecular Science Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory in Richland, Washington. Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC05-76RL0 1830.

References

1. Eng JK, McCormack AL, John R, Yates I. J. Am. Soc. Mass Spectrom. 1994:976–989.
2. Craig R, Beavis RC. Bioinformatics. 2004:1466–1467. [PubMed: 14976030]
3. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. J. Proteome Res. 2004:958–964. [PubMed: 15473683]
4. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Electrophoresis. 1999:3551–3567. [PubMed: 10612281]
5. Tanner S, Shu H, Frank A, Wang L-C, Zandi E, Mumby M, Pevzner PA, Bafna V. Anal. Chem. 2005:4626–4639. [PubMed: 16013882]
6. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Anal. Chem. 2002:5383–5392. [PubMed: 12403597]
7. Choi H, Ghosh D, Nesvizhskii AI. J. Proteome Res. 2008:286–292. [PubMed: 18078310]
8. López-Ferrer D, Martìnez-Bartolomé S, Villar M, Campillos M, Martìn-Maroto F, Vàzquez J. Anal. Chem. 2004:6853–6860. [PubMed: 15571333]
9. Shen C, Wang Z, Shankar G, Zhang X, Li L. Bioinformatics. 2008:202–208. [PubMed: 18024968]
10. Sadygov RG, John R, Yates I. Anal. Chem. 2003:3792–3798. [PubMed: 14572045]
11. Kim S, Gupta N, Pevzner PA. J. Proteome Res. 2008:3354–3363. [PubMed: 18597511]
12. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. Anal. Chem. 2003:4646–4658. [PubMed: 14632076]
13. Smith RD, Anderson GA, Lipton MS, Paša-Tolic̀ L, Shen Y, Conrads TP, Veenstra TD, Udseth HR. Proteomics. 2002:513–523. [PubMed: 11987125]
14. Zimmer JSD, Monroe ME, Qian W-J, Smith RD. Mass Spectrom. Rev. 2006:450–482. [PubMed: 16429408]
15. Jaffe JD, Mani DR, Leptos KC, Church GM, Gillette MA, Carr SA. Mol. Cell. Proteomics. 2006:1927–1941. [PubMed: 16857664]
16. Xu D, Suenaga N, Edelmann MJ, Fridman R, Muschel RJ, Kessler BM. Mol. Cell. Proteomics. 2008:2215–2228. [PubMed: 18596065]
17. Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, Brusniak M-Y, Vitek O, Aebersold R, Müller M. Proteomics. 2007:3470–3480. [PubMed: 17726677]
18. May D, Fitzgibbon M, Liu Y, Holzman T, Eng J, Kemp CJ, Whiteaker J, Paulovich A, McIntosh M. J. Proteome Res. 2007:2685–2694. [PubMed: 17559252]
19. Petritis K, Kangas LJ, Ferguson PL, Anderson GA, Paša-Tolic̀ L, Lipton MS, Auberry KJ, Strittmatter EF, Shen Y, Zhao R, Smith RD. Anal. Chem. 2003:1039–1048. [PubMed: 12641221]
20. Strittmater EF, Ferguson, Tang K, Smith RD. J. Am. Soc. Mass Spectrom. 2003; 14:980–991. [PubMed: 12954166]
21. Washburn MP, Wolters D, John R, Yates I. Nat. Biotechnol. 2001:242–247. [PubMed: 11231557]
22. Qian W-J, Liu T, Monroe ME, Strittmatter EF, Jacobs JM, Kangas LJ, Petritis K, David G, Camp I, Smith RD. J. Proteome Res. 2005:53–62. [PubMed: 15707357]

23. Keller A, Purvine S, Nesvizhskii AI, Stolyar S, Goodlett DR, Kolker E. OMICS. 2002; 6:207–212. [PubMed: 12143966]
24. Petyuk VA, Qian W-J, Chin MH, Wang H, Livesay EA, Monroe ME, Adkins JN, Jaitly N, Anderson DJ, David G, Camp I, Smith DJ, Smith RD. Genome Res. 2007:328–336. [PubMed: 17255552]
25. Tolmachev AV, Monroe ME, Jaitly N, Petyuk VA, Adkins JN, Smith RD. Anal. Chem. 2006:8374–8385. [PubMed: 17165830]
26. Norbeck AD, Monroe ME, Adkins JN, Anderson KK, Daly DS, Smith RD. J. Am. Soc. Mass Spectrom. 2005:1239–1249. [PubMed: 15979333]
27. May D, Liu Y, Law W, Fitzgibbon M, Wang H, Hanash S, McIntosh M. J. Proteome Res. 2008:5148–5156. [PubMed: 19367719]
28. Jaitly N, Adkins JN, Monroe ME, Norbeck AD, Mottaz HM, Dabney AR, Lipton MS, Anderson GA, Smith RD. A Statistical Approach to Quantifying Uncertainties in the AMT Tag Pipeline. 2008
29. Jaitly N, Monroe ME, Petyuk VA, Clauss TRW, Adkins JN, Smith RD. Anal. Chem. 2006:7397–7409. [PubMed: 17073405]
30. Anderson KK, Monroe ME, Daly DS. Proteome Sci. 2006
31. Monroe ME, Tolic̀ N, Jaitly N, Shaw JL, Adkins JN, Smith RD. Bioinformatics. 2007:2021–2023. [PubMed: 17545182]
32. Mayampurath AM, Jaitly N, Purvine SO, Monroe ME, Auberry KJ, Adkins JN, Smith RD. Bioinformatics. 2008; 24:1021–1023. [PubMed: 18304935]
33. Elias JE, Gygi SP. Nat. Methods. 2007:207–214. [PubMed: 17327847]

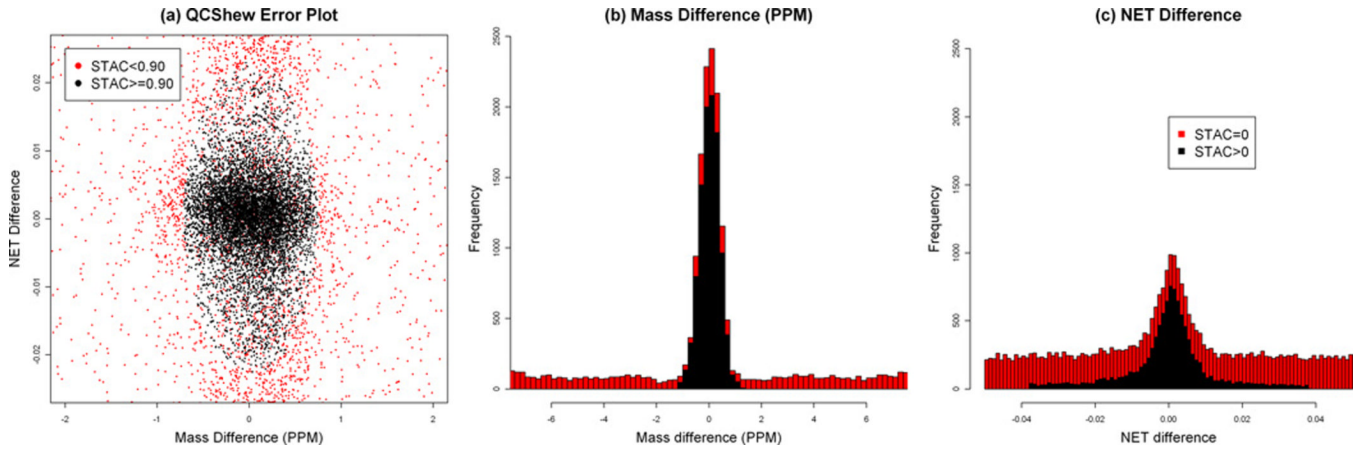

Figure 1.

(a) A scatter plot of mass and NET differences for a typical quality control *Shewanella oneidensis* dataset, with high confidence matches in black, shows the distributional assumptions and the central clustering of matches. (b) and (c) Histograms of mass (b) and NET (c) differences for the same data show the approximate Normality of the distributions and the approximate uniformity of random matches. All matches with positive probability are shown in black and show the distributions of the purportedly correct matches.

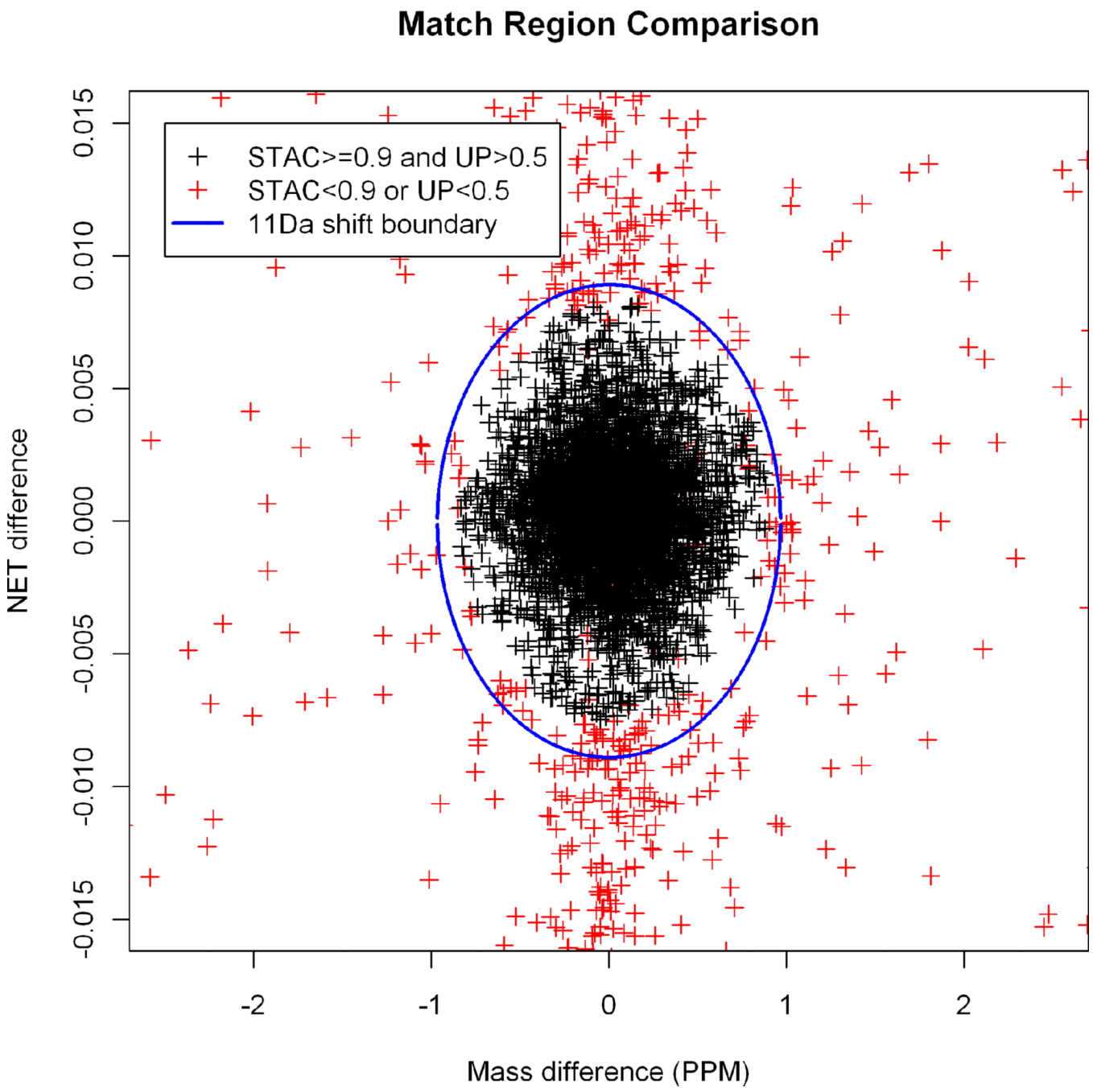

Figure 2.
Scatterplot showing the high confidence STAC matches (black crosses) and a similarly sized region used as a tolerance for an 11 Dalton shift FDR calculation. STAC yields a smaller FDR on approximately the same region when comparing LC-MS data to its matching LC-MS/MS data.

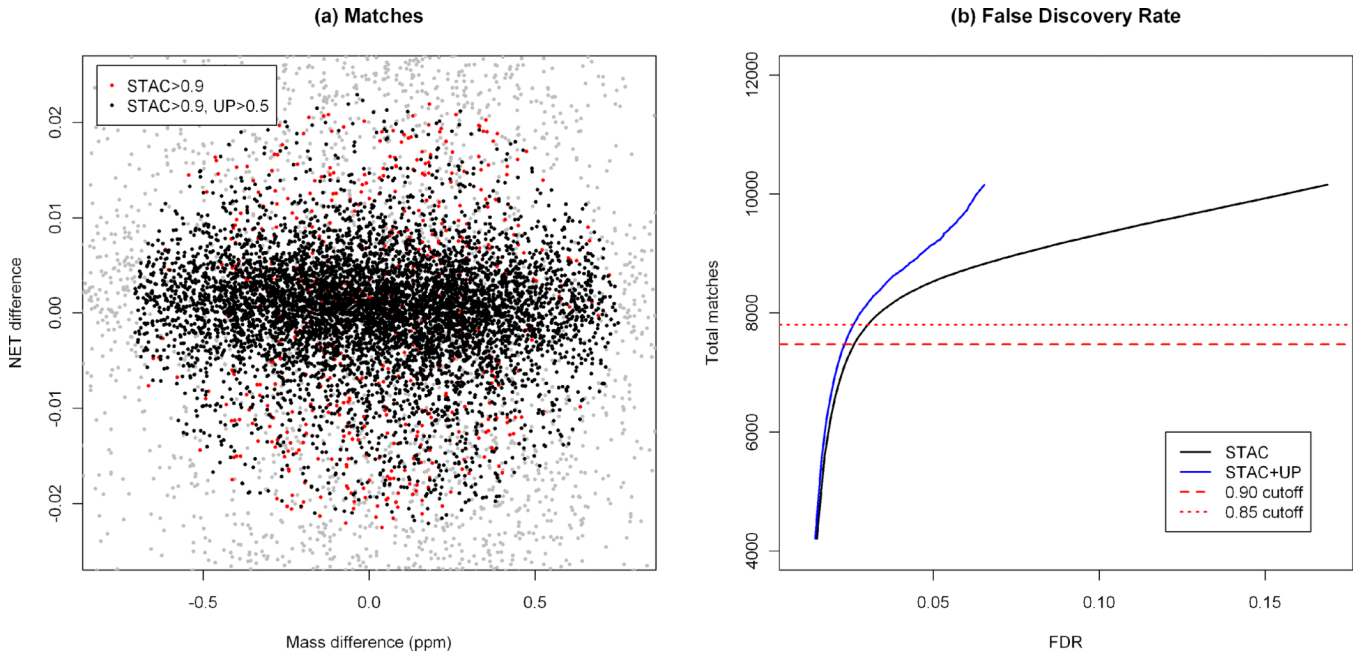

Figure 3.

(a) A scatterplot of the central region of matching a representative QCShew data set to the standard QCShew database. Matches shown in black are both uniquely identified (UP≥ 0:5) and highly confident (STAC> 0:9). Matches shown in red are highly confident, but were excluded due to match ambiguity, while low quality matches are shown in gray. This plot shows the increased specificity of the STAC algorithm by including UP as a filter criteria. (b) This plot shows the relation of the number of matches to the database and the FDR. Indicators are given at 0.85 and 0.90 STAC cutoffs to show the effect of the chosen cutoff on FDR.