AA_stat enables extensive characterization of artefact and post-translational modifications in bottom-up proteomics

Lev I. Levitsky[1,*], Julia A. Bubis[1,2,*], Mikhail V. Gorshkov[1], Irina A. Tarasova[1]

[1] V. L. Talrose Institute for Energy Problems of Chemical Physics, N. N. Semenov Federal Research Center for Chemical Physics, Russian Academy of Sciences, 119334 Moscow, Russia

[2] Moscow Institute of Physics and Technology, Dolgoprudny, Russia

* Contributed equally

Correspondence: iatarasova@yandex.ru

ABSTRACT

We report on AA_stat, a bioinformatic approach for panoramic profiling of artificial and post-translational modifications and their localization sites in large-scale proteomics data. Presented version of AA_stat provides validation of ultra-tolerant (open) search results followed by interpretation of the observed mass shifts and recommendation of the optimized sets of fixed and variable modifications for subsequent regular searches. Localization of modification sites is based on relative amino acid frequencies and analysis of tandem mass spectra. AA_stat determines groups of peptide identifications with mass shifts from the validated results of the open search and then scores each possible mass shift location by matching the MS/MS spectrum across the theoretical peptide isoforms. Here we demonstrate the utility of AA_stat for blind scanning of abundant and rare amino acid modifications of both artificial and biological origins and analyze advantages and limitations of open search strategies. AA_stat is implemented as an open-source command line tool available at https://github.com/SimpleNumber/aa_stat.

INTRODUCTION

Bottom-up proteomics is a method broadly applied for analysis of post-translational modifications (PTMs). Two main strategies exist: targeted PTM identification and the so-called unrestricted PTM search. Targeted approaches to PTM analysis are most commonly employed.[1–4] These methods rely on *a priori* knowledge of modification of interest and, thus, they are not able to characterize peptide modifications in proteomics samples in a blind manner. The latter is important for PTMs which have many types of functional groups, such as O-glycans. Unrestricted search for PTMs can also facilitate studying the crosstalk between PTMs, which typically remains hidden in targeted searches. In addition, blind scanning for unexpected artificial modifications coming from sample preparation can improve the results of peptide and protein identification and quantitation.

The concept of unrestricted, or blind PTM search appeared in the early 2000s.[5] However, the absence of powerful bioinformatics tools capable of coping with huge search spaces within reasonable data processing time restricted implementation of the strategy for a long time. MS-Alignment was among the first algorithms proposed for unrestricted PTM search.[6] Near the same time TwinPeaks,[7] SeMoP,[8] OpenSea,[9] SPIDER[10] and other tools were introduced.[11–13] Large-scale PTM profiling has usually implied some restrictions of the search space due to processing time issues. As a result, only partial detection of modifications of interest is usually attainable.

This situation has started to change with the appearance of ultrafast search engine MSFragger.[14] MSFragger became the first search engine that provides unrestricted PTM search in reasonable running time when combined with the ultra-tolerant search strategy.[15] Ultra-tolerant search assumes using a loose precursor mass tolerance (±500 Da and even more) for peptide identification. Here, we call this approach the “open search”, while conventional, narrow-tolerance search is called “closed search”. Peptide-spectrum matches (PSMs) identified with open search can be grouped by mass shift values that correspond to lost or attached functional groups. Recently, other engines, OpenPFind[16] and TagGraph,[17] were proposed as alternatives to MSFragger for open search.

Introduction of new powerful tools for ultra-fast peptide identification with open search calls for comprehensive algorithms for postsearch validation and interpretation of detected mass shifts. Few such tools are available, and their functionality is typically restricted to mass shift clustering followed by filtering, validation and mass shift localization at a particular amino acid residue (PTMiner[18], DeltaMass[19]). Interpretation of the mass shifts is typically based on annotations in the Unimod database.[20] PTM-Shepherd aims at addressing a lack of metrics necessary for proper determination of identities and origins of modifications discovered in the open search.[21] However, these tools do not provide practical recommendations for further optimization of the parameters of the closed searches.

Recently, we introduced AA_stat, an open-source software for analysis and interpretation of the results of ultra-tolerant searches using amino acid frequency counts, which can be easily integrated into the open-search-based workflow.[22] Similar to the other algorithms,[18,19,21] AA_stat selects well-resolved mass shifts, followed by determination of accurate mass shift values using Gaussian fit and group-specific false discovery rate (FDR) estimation with target-decoy approach (TDA), as well as the Unimod-based interpretation. Here, we report on significantly renewed AA_stat that now includes an improved algorithm for amino acid frequency normalization, localization of mass shifts based on tandem mass spectra,[23] and provides recommendations on fixed and variable modifications to further optimize performance of subsequent closed searches. AA_search option of AA_stat automatically optimizes fixed modifications, initializing the open searches. We also present an analysis of advantages and limitations of the open-search-based strategies for large-scale profiling of artificial modifications and PTMs. Finally, we demonstrate the capabilities of AA_stat for analysis of both abundant modifications and rare PTMs.

EXPERIMENTAL SECTION

Public data were used for validation of AA_stat.[24–31] Detailed annotation of the datasets is provided in Supporting Information (**S-Table 1**). Selection was aimed at obtaining datasets containing a diverse number of modified amino acid residues per peptide. Data included

enrichments of S/T phosphorylation, N-glycosylation, O-glycosylation, K-acetylation and labelling with SILAC[32] and Tandem Mass Tags.[33]

MSFragger v2.3 was used for both open and closed searches. Parameters for open search were left at default settings, except for the precursor and fragment tolerances and fragment ion types (**S-Table 2)**. To retrieve fixed modifications, AA_stat with AA_search wrapper was run. SwissProt human canonical sequence database was used, retrieved on March 21, 2019. Decoy sequences were generated in reverse mode using Pyteomics.[34,35] Precursor and fragment tolerances, fragment ion types for subsequent closed searches were set as stated in **S-Table 3**. Isotope error correction, fixed and variable modifications for closed searches were set according to annotations (**S-Table 1**) or AA_stat recommendations (**S-Table 4**). Group-specific FDR filtering using target-decoy approach[36] and validation of identifications for closed searches were performed using Scavager.[37] Gene ontology (GO) enrichments were analyzed with STRING[38] (https://string-db.org/).

RESULTS AND DISCUSSION

Overview of general updates of AA_stat tool. AA_stat is an open-source command-line tool that processes the results of open search and composes a list of identified mass shifts that can be attributed to modifications. AA_stat calculates and visualizes amino acid statistics within the groups of peptides with shifted masses. This statistics allows to derive an initial hypothesis on possible modification sites. The latest version of AA_stat localizes identified mass shifts at specific amino acid residues using MS/MS spectra. Based on the localization results, AA_stat suggests a list of modifications to be considered for closed search. AA_search is a new command that runs an open search using MSFragger search engine and performs AA_stat analysis of the results. If AA_stat recommends any fixed modifications, AA_search can automatically apply them and repeat the open search for optimal results. General data processing flow of AA_stat and AA_search utilities is shown in **Figure 1**.

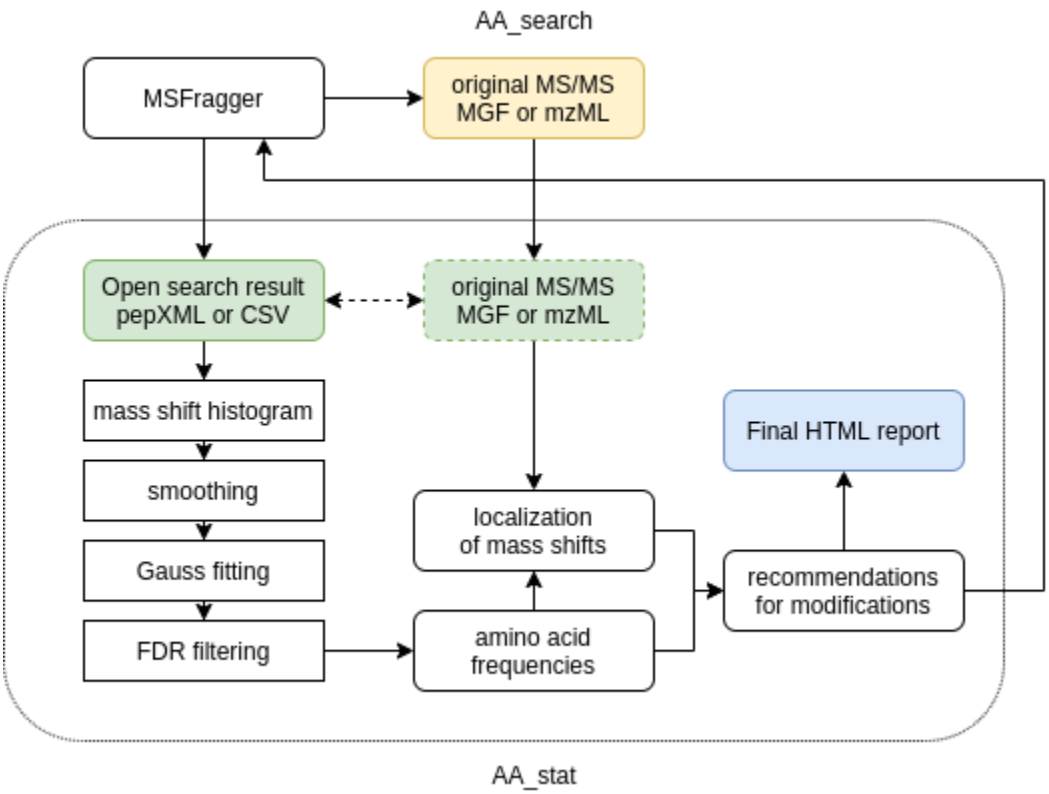

Figure 1. Overview of AA_stat and AA_search algorithms. AA_stat starts with open search results (pepXML or CSV), then groups PSMs by mass shift values. Next, smoothing, Gauss fitting and group-specific FDR filtering are performed, followed by calculation of normalized amino acid frequencies. After composing the amino acid candidate list, AA_stat localizes mass shifts (using mzML or MGF files), concludes on fixed and variable modifications and produces a final HTML report. AA_search is a wrapper that runs an open search with MSFragger, followed by post-processing of the results with AA_stat. Performing in a cycle, AA_search then applies fixed modifications recommended by AA_stat and repeats the open search for optimal results.

Data processing in AA_stat involves Gaussian fit to group potentially modified peptides, group-specific FDR filtering, amino acid counting, and frequency normalization for the particular mass shifts.[22] In the first version of AA_stat,[22] the zero mass shift (containing unmodified peptides and/or peptides with fixed modifications) was always used for normalization (*reference mass shift*). The latter restricted applicability of the tool to label-free and TMT-based proteomics workflows. For example, in cases of metabolic labeling, zero mass shift (0.0 *m/z*) can be rarely observed in the open search results and consequently provides poor normalization as a reference. The updated AA_stat dynamically selects the reference mass shift using the following algorithm. First, AA_stat determines if the zero mass shift exists in the window of [-

0.05 Da, 0.05 Da] (here and throughout, default configuration parameters are specified). Then, if such a mass shift is found and it contains more peptides than 5% of the most abundant mass shift, it becomes the reference mass shift. Otherwise, the most abundant mass shift is chosen as a reference. Additionally, if the zero mass shift is found, all mass shift values are adjusted so that the value for the zero mass shift is exactly zero. This calibration procedure is performed individually for every file analyzed, to correct for systematic errors in precursor *m/z* measurements. After mass calibration and determination of the reference mass shift, AA_stat calculates normalized amino acid frequencies, and percentage of peptides containing each specific amino acid for each mass shift ("peptide percentage"). This information is further employed in the mass shift localization algorithm followed by recommendation of fixed and variable modifications.

Localization of modification sites by AA_stat. The latest version of AA_stat localizes identified mass shifts and their superpositions at specific sites of a peptide using MS/MS spectra. The localization procedure is outlined below.

First, a list of possible amino acid localizations for each mass shift is formed. There are three sources of amino acid candidates (eligible residues) for a specific mass shift: (1) all amino acids (including N- and C-terminus) from Unimod database that correspond to the mass shift with a given mass tolerance; (2) amino acids whose normalized frequency (eq. 2 in our previous study[22]) exceeds a given threshold; (3) if mass shift can be potentially a ^{13}C isotope within a given mass tolerance, all candidates of monoisotopic mass shift are considered.

At the next step, the isoforms for each peptide are generated according to the amino acid candidate list. If the considered mass shift equals the sum of two other observed mass shifts within a given tolerance, isoforms for the combination of modifications are also generated.

Next, the theoretical spectrum for each peptide isoform is generated by considering singly-charged fragment ions. AA_stat uses a simplified renormalized Hyperscore scoring function introduced earlier.[39] Each isoform is scored against the experimental spectrum within a given fragment mass tolerance. If there is no clear winner among the isoforms, the mass shift is treated as "non-localized". Only a single top isoform is considered as a successful localization

for each peptide. If a modification is localized at a terminal residue, then it is also considered to be localized at the corresponding peptide terminus. Localization score is calculated as the difference between the top isoform score and the second best one, divided by the top score. If the isoform list consists of only one entity, the peptide is considered localized with zero score.

Finally, AA_stat produces comprehensive output files. It generates a CSV table which contains amino acid candidates for localization, isotope information, and localization results (full description of columns is provided at https://github.com/SimpleNumber/aa_stat). Text-format tables named after mass shifts are also reported. They contain top isoform, experimental spectrum ID and localization score for each identified peptide at the considered mass shift. Localization statistics for specific mass shifts along with normalized amino acid frequencies and peptide percentages are shown in self-titled figures. These tables and figures are gathered in a single HTML report file for convenient access. An example of the AA_stat output is shown at https://levitsky.github.io/aa_stat_report_example/.

Selection of fixed and variable modifications. New version of AA_stat uses MS/MS localization results to recommend fixed and variable modifications. "Fixed modifications" are recommended to be switched on in the open search for the purposes of producing an easily interpretable AA_stat result and increasing sensitivity of modification profiling. Indeed, one of the main outcomes of AA_stat analysis are normalized amino acid frequencies. The normalization procedure for a given amino acid is only sound if a sufficient amount of peptides containing that amino acid are detected with the reference mass shift. However, if a certain residue is present only in modified form, this condition is not satisfied, then, normalization procedure requires correction. Details of the algorithm for selection of fixed modifications are shown in **S-Figure 1A**. Briefly, the algorithm starts with checking the reference mass shift determined by AA_stat. If the reference mass shift is *m/z* 0.0, a list of amino acid candidates bearing fixed modifications is composed of amino acids whose occurrence at zero shift is below a threshold. Then, for each residue candidate, the mass shift containing the maximum total peptide percentage is determined. If the peptide percentage for this mass shift satisfies selection criterion, the fixed modification is recommended. If the reference shift is not zero, preferable mass shift localizations in the non-zero reference shift are ranked. If the most

preferable localization site is determined, then this mass shift is recommended as fixed modification on the preferable site. Otherwise, the procedure is repeated for the mass shift next by abundance.

AA_search routine has an option that implements the recommended fixed modifications and runs another round of open search automatically. This is repeated until no more fixed modifications are needed. This takes from one to four steps, depending on the complexity of the data.

Variable modifications are recommended to summarize the abundant modifications retrieved by AA_stat. AA_stat aims at producing the most effective list of variable modifications for a follow-up closed search. Recommendation is done solely on the basis of successful MS/MS localizations, preferring the mass shifts and modification sites with the highest absolute localization count. Already recommended fixed modifications are excluded. The user can also instruct AA_stat to not recommend multiple modifications of the same residue. Additionally, a recommendation is made to allow for the possibility of isotope mass error; it is also accounted for in the selection of mass shifts as variable modifications. Exact procedure is outlined in **S-Figure 1B**.

Limitations of the open search strategies decrease sensitivity of modification profiling. To assess the limitations of the open search strategies for profiling of amino acid modifications, public data were considered (**S-Table 1**). Datasets were generated using either label-free or label-based proteomics workflows and PTM enrichments. **S-Figure 2** demonstrates that phosphorylation enrichment followed by TMT labeling resulted in a peptide set in which more than 90% of peptides bear 3 and more modifications. Such datasets represent the highest level of complexity for the open search strategies for uncovering PTMs and artificial modifications. The presence of two or even more modifications in the peptide sequence shifts the fragments in tandem spectra. The shifted fragment series have especially profound effects if modification sites are distantly located from each other in the peptide sequence, because the chance to match the MS2 spectrum of a modified peptide to the theoretical spectrum of its unmodified isoform significantly decreases. Particularly, less than 10% of tandem spectra in such “over-modified” datasets result in successful identifications in the open searches (**S-Table 2**). Thus,

the latter observation represents an intrinsic limitation of open-search strategies aimed at panoramic uncovering modifications in a blind manner. However, the problem can be partially resolved by either enabling the fixed modifications expected from the sample preparation or step-by-step inclusion of AA_stat-predicted fixed modifications.

Thus, inclusion of fixed modifications in open search is beneficial for the following reasons. First, an increase in the total number of peptide identifications in open search results significantly improves the sensitivity of open search methods meant for large-scale modification profiling. Second, inclusion of fixed modifications usually reduces the number of observed mass shifts in the open search, which simplifies interpretation of the results. Additionally, accounting for fixed modifications in post-processing with AA_stat is required to improve accuracy of calculating the normalized amino acid frequencies, which are used further to hypothesize localizations of the functional groups at particular residues.

AA_stat correctly recommends abundant modifications for subsequent closed searches. AA_stat correctly retrieves abundant mass shifts from the results of open search and recommends profiles of fixed and variable modifications. Numbers of peptide identifications in open and subsequent closed searches with AA_stat-recommended modification sets are comparable with or even higher than the ones obtained with annotated modifications (**Figure 2, S-Table 2-3**).

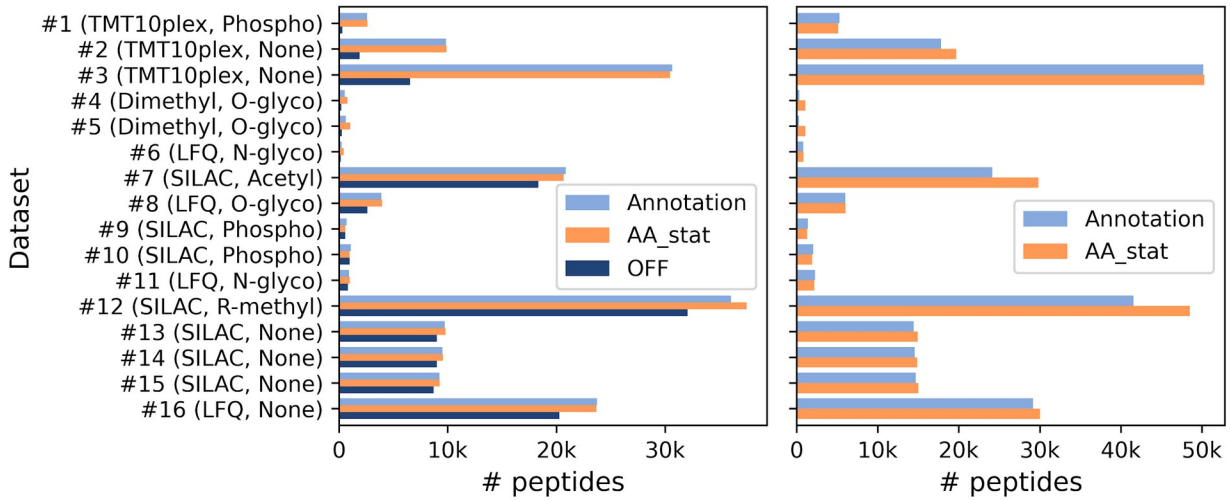

Figure 2. Numbers of peptide identifications in open (A) and closed (B) searches are comparable or even higher when the AA_stat recommended modification sets are used. Y-axis labels

correspond to subsets as listed in **S-Tables 3-4**. "None" means no enrichment. For open searches, the number of peptides is the sum of peptide counts in all mass shifts as reported by AA_stat. For closed searches, the number of peptides as reported by Scavager is shown. Search results were obtained using either the annotated (from a respective study) or AA_stat-recommended sets of modifications. To retrieve modifications, AA_stat was run with the AA_search wrapper.

S-Table 4 summarizes the sets of fixed and variable modifications recommended by AA_stat in comparison with the modification sets from annotation of the respective studies.[24–31] Comparing the annotated and AA_stat-retrieved modifications, both expected and unexpected modifications can be mentioned. For example, Cys-carbamidomethylation was annotated for all data, however, two subsets with enrichments of phosphorylated peptides (#9, #10 in **Figure 2**) and one subset with R-methylation enrichment (#12 in **Figure 2**) contained no or few peptides with +57.02 *m/z* mass shift. At the same time, AA_stat recommends fixed modification of Met residue for some subsets. For example, nearly all methionines in #6, #11 and #12 subsets are present in oxidized form (+15.99 *m/z*), while #16 has +43 *m/z* localized at Met. Variable modifications recommended by AA_stat for each subset match the annotated profiles well. Analysis shows that datasets with phosphorylation enrichment preferably contain peptides with modified Ser and Thr, but not Tyr. Dataset with N-glycosylation has revealed more than 95% of peptides contained +0.98@Asparagine, the asparagine residues deglycosylated with PNGase F. HexNAc with *m/z* of +203@ST were detected in a subset with O-glycosylation enrichment. Necessity for correction of monoisotopic peak selection was revealed for 3 of 16 considered subsets. Among the unexpected artificial modifications, overalkylation of CPTAC subset can be mentioned, for which +57.02 and +114.04 *m/z* were localized at peptide N-terminus.

Some search engines restrict the search space by limiting the number of variable modifications per peptide. Based on the results in the AA_stat's HTML report, a grounded decision on the value for the parameter can be made. Detailed instructions for such a decision are provided using example of AA_stat results for subset #7 (**S-Figure 3**). This subset combines post-translational Lys acetylation (+42.01 *m/z*), metabolic labeling of Lys and Arg residues (+8.01 and 10.01 *m/z*) and methionine oxidation (+15.99 *m/z*). Specifying one missed cleavage suggests the presence of up to two modified lysines per peptide. Thus, taking into account

relative percentage of peptides bearing together methionine oxidation, SILAC labels, and acetylation, allowing at least three sites per peptide can be recommended.

AA_stat allows comprehensive mass shift interpretation. AA_stat assists user's decisions on the following issues: (1) choice of modified residues and respective mass shifts (according to the algorithms in **S-Figure 1AB**), (2) interpretation of each mass shift by interactive linking to Unimod (http://www.unimod.org), (3) whether modification should be set as fixed or variable, (4) maximum number of modifications allowed per peptide and/or amino acid residue (according to the scheme in **S-Figure 3**), and, (5) need for correction of erroneous choice of monoisotopic peak. Altogether these features protect the user from loss of identifications due to unexpected modifications and prevent an extension of the search space due to inclusion of variable modifications absent in the sample. A dramatic example is identification of O-glycans, for which possible variations of attached sugars are counted in hundreds.[40] Recently, several studies have demonstrated a high potential of open search strategies for glycosylation research.[41,42] However, we argue that even the rather routine identification of phosphorylated peptides is often performed with excessive variable modification sets. Typically Ser, Thr and Tyr are set as variable phosphorylation sites in proteomics studies, often regardless of experimental conditions and enrichment protocols. At the same time, identification of representative sets of Tyr-phosphorylation is prominent only for protocols allowing specific extraction of Tyr-kinase substrates. In phosphoproteomes from whole cell lysates, peptide identifications with Tyr phosphorylation are rare.

AA_stat demonstrates a potential for uncovering rare modifications. One of the common strategies for mass-spectrometry-based discovery of rare post-translational modifications is the analysis of public data followed by validation of the collected rare events for significance and biological relevance.[43] We tested AA_stat for the ability to accumulate peptides with rare PTMs when analyzing a subset of data and clearly demonstrated its potential for uncovering rare PTMs. Oxidation of proline was used as an example. Based on AA_stat reports, we selected from the CPTAC subset (#16 in **Fig. 2**) the mass shift of 15.99 *m/z* containing 69 peptides with localization of hydroxyl group at Pro. These peptides predominantly originated from collagens and collagen-like proteins, a well-characterized

protein family which needs hydroxylation of prolines for correct folding.[44] GO analysis of the corresponding proteins revealed enrichment of collagen trimer and endoplasmic reticulum (ER) where hydroxylation takes place (**S-Table 5**, **S-Figure 4**). Localization of hydroxyl groups at particular protein sites was additionally verified against the Uniprot database. 26 of 69 localizations reported by AA_stat are at sites listed in Uniprot (**S-Table 6**). Consensus-motif recognition demonstrated the presence of XpG motif around localization of hydroxyproline for 55 of 69 cases. XpG (any amino acid - hydroxyproline - glycine) is a well-described conserved amino acid motif for the prolyl 4-hydroxylases, the enzymes catalyzing modification of collagens in ER.[45]

CONCLUSION

New implementations differentiate AA_stat from analogous tools intended for post-processing of the open search results. Updated AA_stat performs comprehensive large-scale profiling of modifications in the open search results, complemented with specific recommendations for the subsequent closed search. Currently, these recommendations include fixed and variable modifications of both chemical (*in vitro*) and biological origin and correction for erroneous monoisotopic peak assignment. Generation of optimized dataset-specific profiles of modifications serves not only for the purposes of accounting for unexpected modifications and bewaring the excessive ones, but can be also considered as a first step to a blind big data analysis without *a priori* knowledge of experimental conditions.

Although open search strategies are always limited by the requirement for matching a modified peptide to its unmodified isoform, they can demonstrate reasonable sensitivity even when over 90% of peptides in the dataset bear two or more modifications. However, the sensitivity issues of open search for identification of modifications still exist and need to be addressed. Nevertheless, open search assisted by comprehensive post-processing demonstrates a great potential for large-scale profiling of abundant and rare modifications of both biological and chemical origins.

ASSOCIATED CONTENT

Supporting Information is available.

ACKNOWLEDGEMENTS

Authors thank Dr. Mark Ivanov for helpful discussions on AA_stat development. Data processing was performed with the support from 2013 to 2020 Program of Basic Research of the Russian State Academies of Sciences (theme # AAAA-A18–118112090111-4, internal # 0068-2019-0011). Bioinformatic tool development was supported by Russian Foundation for Basic Research (project # 18-29-13015).

AUTHOR CONTRIBUTIONS

I.A.T. conceived the project, L.I.L. and J.A.B. developed the software, I.A.T. performed analyses, all authors wrote the manuscript, M.V.G. and I.A.T. supervised the entire project. L.I.L. and J.A.B. contributed equally.

REFERENCES

(1) Falkner, J. A.; Falkner, J. W.; Yocum, A. K.; Andrews, P. C. A Spectral Clustering Approach to MS/MS Identification of Post-Translational Modifications. *J. Proteome Res.* **2008**, *7* (11), 4614–4622. https://doi.org/10.1021/pr800226w.

(2) Frank, A.; Pevzner, P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Anal. Chem.* **2005**, *77* (4), 964–973. https://doi.org/10.1021/ac048788h.

(3) Frank, A.; Tanner, S.; Bafna, V.; Pevzner, P. Peptide Sequence Tags for Fast Database Search in Mass-Spectrometry. *J. Proteome Res.* **2005**, *4* (4), 1287–1295. https://doi.org/10.1021/pr050011x.

(4) Tharakan, R.; Edwards, N.; Graham, D. R. M. Data Maximization by Multipass Analysis of Protein Mass Spectra. *PROTEOMICS* **2010**, *10* (6), 1160–1171. https://doi.org/10.1002/pmic.200900433.

(5) Tsur, D.; Tanner, S.; Zandi, E.; Bafna, V.; Pevzner, P. A. Identification of Post-Translational Modifications by Blind Search of Mass Spectra. *Nat. Biotechnol.* **2005**, *23* (12), 1562–1567. https://doi.org/10.1038/nbt1168.

(6) Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. Efficiency of Database Search for Identification of Mutated and Modified Proteins via Mass Spectrometry. *Genome Res.* **2001**, *11* (2), 290–299. https://doi.org/10.1101/gr.154101.

(7) Havilio, M.; Wool, A. Large-Scale Unrestricted Identification of Post-Translation Modifications Using Tandem Mass Spectrometry. *Anal. Chem.* **2007**, *79* (4), 1362–1368. https://doi.org/10.1021/ac061515x.

(8) Baumgartner, C.; Rejtar, T.; Kullolli, M.; Akella, L. M.; Karger, B. L. SeMoP: A New

Computational Strategy for the Unrestricted Search for Modified Peptides Using LC−MS/MS Data. *J. Proteome Res.* **2008**, *7* (9), 4199–4208. https://doi.org/10.1021/pr800277y.

(9) Searle, B. C.; Dasari, S.; Wilmarth, P. A.; Turner, M.; Reddy, A. P.; David, L. L.; Nagalla, S. R. Identification of Protein Modifications Using MS/MS de Novo Sequencing and the OpenSea Alignment Algorithm. *J. Proteome Res.* **2005**, *4* (2), 546–554. https://doi.org/10.1021/pr049781j.

(10) Yonghua Han; Bin Ma; Kaizhong Zhang. SPIDER: Software for Protein Identification from Sequence Tags with de Novo Sequencing Error. In *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.*; 2004; pp 206–215. https://doi.org/10.1109/CSB.2004.1332434.

(11) Chalkley, R. J.; Baker, P. R.; Medzihradszky, K. F.; Lynn, A. J.; Burlingame, A. L. In-Depth Analysis of Tandem Mass Spectrometry Data from Disparate Instrument Types. *Mol. Cell. Proteomics* **2008**. https://doi.org/10.1074/mcp.M800021-MCP200.

(12) Na, S.; Paek, E. Prediction of Novel Modifications by Unrestrictive Search of Tandem Mass Spectra. *J. Proteome Res.* **2009**, *8* (10), 4418–4427. https://doi.org/10.1021/pr9001146.

(13) Dasari, S.; Chambers, M. C.; Slebos, R. J.; Zimmerman, L. J.; Ham, A.-J. L.; Tabb, D. L. TagRecon: High-Throughput Mutation Identification through Sequence Tagging. *J. Proteome Res.* **2010**, *9* (4), 1716–1726. https://doi.org/10.1021/pr900850m.

(14) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–Based Proteomics. *Nat. Methods* **2017**, *14* (5), 513–520. https://doi.org/10.1038/nmeth.4256.

(15) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun Proteomics as Modified Peptides. *Nat. Biotechnol.* **2015**, *33* (7), 743–749. https://doi.org/10.1038/nbt.3267.

(16) Chi, H.; Liu, C.; Yang, H.; Zeng, W.-F.; Wu, L.; Zhou, W.-J.; Wang, R.-M.; Niu, X.-N.; Ding, Y.-H.; Zhang, Y.; Wang, Z.-W.; Chen, Z.-L.; Sun, R.-X.; Liu, T.; Tan, G.-M.; Dong, M.-Q.; Xu, P.; Zhang, P.-H.; He, S.-M. Comprehensive Identification of Peptides in Tandem Mass Spectra Using an Efficient Open Search Engine. *Nat. Biotechnol.* **2018**, *36* (11), 1059–1061. https://doi.org/10.1038/nbt.4236.

(17) Devabhaktuni, A.; Lin, S.; Zhang, L.; Swaminathan, K.; Gonzalez, C. G.; Olsson, N.; Pearlman, S. M.; Rawson, K.; Elias, J. E. TagGraph Reveals Vast Protein Modification Landscapes from Large Tandem Mass Spectrometry Datasets. *Nat. Biotechnol.* **2019**, *37* (4), 469–479. https://doi.org/10.1038/s41587-019-0067-5.

(18) An, Z.; Zhai, L.; Ying, W.; Qian, X.; Gong, F.; Tan, M.; Fu, Y. PTMiner: Localization and Quality Control of Protein Modifications Detected in an Open Search and Its Application to Comprehensive Post-Translational Modification Characterization in Human Proteome. *Mol. Cell. Proteomics* **2019**, *18* (2), 391–405. https://doi.org/10.1074/mcp.RA118.000812.

(19) Avtonomov, D. M.; Kong, A.; Nesvizhskii, A. I. DeltaMass: Automated Detection and Visualization of Mass Shifts in Proteomic Open-Search Results. *J. Proteome Res.* **2019**, *18* (2), 715–720. https://doi.org/10.1021/acs.jproteome.8b00728.

(20) Creasy, D. M.; Cottrell, J. S. Unimod: Protein Modifications for Mass Spectrometry.

PROTEOMICS **2004**, *4* (6), 1534–1536. https://doi.org/10.1002/pmic.200300744.
(21) Geiszler, D. J.; Kong, A. T.; Avtonomov, D. M.; Yu, F.; Leprevost, F. V.; Nesvizhski, A. I. PTM-Shepherd: Analysis and Summarization of Post-Translational and Chemical Modifications from Open Search Results. *bioRxiv* **2020**, 2020.07.08.192583. https://doi.org/10.1101/2020.07.08.192583.
(22) Bubis, J. A.; Levitsky, L. I.; Ivanov, M. V.; Gorshkov, M. V. Validation of Peptide Identification Results in Proteomics Using Amino Acid Counting. *PROTEOMICS* **2018**, *18* (23), 1800117. https://doi.org/10.1002/pmic.201800117.
(23) Tarasova, I. A.; Chumakov, P. M.; Moshkovskii, S. A.; Gorshkov, M. V. Profiling Modifications for Glioblastoma Proteome Using Ultra-Tolerant Database Search: Are the Peptide Mass Shifts Biologically Relevant or Chemically Induced? *J. Proteomics* **2019**, *191*, 16–21. https://doi.org/10.1016/j.jprot.2018.05.010.
(24) Stepath, M.; Zülch, B.; Maghnouj, A.; Schork, K.; Turewicz, M.; Eisenacher, M.; Hahn, S.; Sitek, B.; Bracht, T. Systematic Comparison of Label-Free, SILAC, and TMT Techniques to Study Early Adaption toward Inhibition of EGFR Signaling in the Colorectal Cancer Cell Line DiFi. *J. Proteome Res.* **2020**, *19* (2), 926–937. https://doi.org/10.1021/acs.jproteome.9b00701.
(25) Saei, A. A.; Beusch, C. M.; Chernobrovkin, A.; Sabatier, P.; Zhang, B.; Tokat, Ü. G.; Stergiou, E.; Gaetani, M.; Végvári, Á.; Zubarev, R. A. ProTargetMiner as a Proteome Signature Library of Anticancer Molecules for Functional Discovery. *Nat. Commun.* **2019**, *10* (1), 5715. https://doi.org/10.1038/s41467-019-13582-8.
(26) Narimatsu, Y.; Joshi, H.; Schjoldager, K. T.; Hintze, J.; Halim, A.; Steentoft, C.; Nasson, R.; Mandel, U.; Bennett, E. P.; Clausen, H.; Vakhrushev, S. Y. Exploring Regulation of Protein O-Glycosylation in Isogenic Human HEK293 Cells by Differential O-Glycoproteomics. *Mol. Cell. Proteomics* **2019**. https://doi.org/10.1074/mcp.RA118.001121.
(27) Yang, W.; Jackson, B.; Zhang, H. Identification of Glycoproteins Associated with HIV Latently Infected Cells Using Quantitative Glycoproteomics. *PROTEOMICS* **2016**, *16* (13), 1872–1880. https://doi.org/10.1002/pmic.201500215.
(28) Qin, H.; Cheng, K.; Zhu, J.; Mao, J.; Wang, F.; Dong, M.; Chen, R.; Guo, Z.; Liang, X.; Ye, M.; Zou, H. Proteomics Analysis of O-GalNAc Glycosylation in Human Serum by an Integrated Strategy. *Anal. Chem.* **2017**, *89* (3), 1469–1476. https://doi.org/10.1021/acs.analchem.6b02887.
(29) Hansen, B. K.; Gupta, R.; Baldus, L.; Lyon, D.; Narita, T.; Lammers, M.; Choudhary, C.; Weinert, B. T. Analysis of Human Acetylation Stoichiometry Defines Mechanistic Constraints on Protein Regulation. *Nat. Commun.* **2019**, *10* (1), 1055. https://doi.org/10.1038/s41467-019-09024-0.
(30) Musiani, D.; Bok, J.; Massignani, E.; Wu, L.; Tabaglio, T.; Ippolito, M. R.; Cuomo, A.; Ozbek, U.; Zorgati, H.; Ghoshdastider, U.; Robinson, R. C.; Guccione, E.; Bonaldi, T. Proteomics Profiling of Arginine Methylation Defines PRMT5 Substrate Specificity. *Sci. Signal.* **2019**, *12* (575). https://doi.org/10.1126/scisignal.aat8388.
(31) Vasaikar, S.; Huang, C.; Wang, X.; Petyuk, V. A.; Savage, S. R.; Wen, B.; Dou, Y.; Zhang, Y.; Shi, Z.; Arshad, O. A.; Gritsenko, M. A.; Zimmerman, L. J.; McDermott, J. E.; Clauss, T. R.; Moore, R. J.; Zhao, R.; Monroe, M. E.; Wang, Y.-T.; Chambers, M. C.; Slebos, R. J. C.; Lau, K. S.; Mo, Q.; Ding, L.; Ellis, M.; Thiagarajan, M.; Kinsinger, C. R.; Rodriguez, H.; Smith, R.

D.; Rodland, K. D.; Liebler, D. C.; Liu, T.; Zhang, B.; Pandey, A.; Paulovich, A.; Hoofnagle, A.; Mani, D. R.; Chan, D. W.; Ransohoff, D. F.; Fenyo, D.; Tabb, D. L.; Levine, D. A.; Boja, E. S.; Kuhn, E.; White, F. M.; Whiteley, G. A.; Zhu, H.; Zhang, H.; Shih, I.-M.; Bavarva, J.; Whiteaker, J.; Ketchum, K. A.; Clauser, K. R.; Ruggles, K.; Elburn, K.; Hannick, L.; Watson, M.; Oberti, M.; Mesri, M.; Sanders, M. E.; Borucki, M.; Gillette, M. A.; Snyder, M.; Edwards, N. J.; Vatanian, N.; Rudnick, P. A.; McGarvey, P. B.; Mertins, P.; Townsend, R. R.; Thangudu, R. R.; Rivers, R. C.; Payne, S. H.; Davies, S. R.; Cai, S.; Stein, S. E.; Carr, S. A.; Skates, S. J.; Madhavan, S.; Hiltke, T.; Chen, X.; Zhao, Y.; Wang, Y.; Zhang, Z. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell* **2019**, *177* (4), 1035-1049.e19. https://doi.org/10.1016/j.cell.2019.03.030.

(32) Emadali, A.; Gallagher-Gambarelli, M. La protéomique quantitative par la méthode SILAC - Technique et perspectives. *médecine/sciences* **2009**, *25* (10), 835–842. https://doi.org/10.1051/medsci/20092510835.

(33) Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Anal. Chem.* **2003**, *75* (8), 1895–1904. https://doi.org/10.1021/ac0262560.

(34) Goloborodko, A. A.; Levitsky, L. I.; Ivanov, M. V.; Gorshkov, M. V. Pyteomics—a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics. *J. Am. Soc. Mass Spectrom.* **2013**, *24* (2), 301–304. https://doi.org/10.1021/jasms.8b04453.

(35) Levitsky, L. I.; Klein, J. A.; Ivanov, M. V.; Gorshkov, M. V. Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *J. Proteome Res.* **2019**, *18* (2), 709–714. https://doi.org/10.1021/acs.jproteome.8b00717.

(36) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat. Methods* **2007**, *4* (3), 207–214. https://doi.org/10.1038/nmeth1019.

(37) Ivanov, M. V.; Levitsky, L. I.; Bubis, J. A.; Gorshkov, M. V. Scavager: A Versatile Postsearch Validation Algorithm for Shotgun Proteomics Based on Gradient Boosting. *PROTEOMICS* **2019**, *19* (3), 1800280. https://doi.org/10.1002/pmic.201800280.

(38) Szklarczyk, D.; Gable, A. L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N. T.; Morris, J. H.; Bork, P.; Jensen, L. J.; Mering, C. von. STRING V11: Protein–Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* **2019**, *47* (D1), D607–D613. https://doi.org/10.1093/nar/gky1131.

(39) Levitsky, L. I.; Ivanov, M. V.; Lobas, A. A.; Bubis, J. A.; Tarasova, I. A.; Solovyeva, E. M.; Pridatchenko, M. L.; Gorshkov, M. V. IdentiPy: An Extensible Search Engine for Protein Identification in Shotgun Proteomics. *J. Proteome Res.* **2018**, *17* (7), 2249–2255. https://doi.org/10.1021/acs.jproteome.7b00640.

(40) Xiao, H.; Sun, F.; Suttapitugsakul, S.; Wu, R. Global and Site-Specific Analysis of Protein Glycosylation in Complex Biological Systems with Mass Spectrometry. *Mass Spectrom. Rev.* **2019**, *38* (4–5), 356–379. https://doi.org/10.1002/mas.21586.

(41) Polasky, D. A.; Yu, F.; Teo, G. C.; Nesvizhskii, A. I. Fast and Comprehensive N- and O-Glycoproteomics Analysis with MSFragger-Glyco. *bioRxiv* **2020**, 2020.05.18.102665.

https://doi.org/10.1101/2020.05.18.102665.
(42) Izaham, A. R. A.; Scott, N. E. Open Database Searching Enables the Identification and Comparison of Bacterial Glycoproteomes without Defining Glycan Compositions Prior to Searching. *Mol. Cell. Proteomics* **2020**. https://doi.org/10.1074/mcp.TIR120.002100.
(43) Onisko, B. C. The Hydroxyproline Proteome of HeLa Cells with Emphasis on the Active Sites of Protein Disulfide Isomerases. *J. Proteome Res.* **2020**, *19* (2), 756–768. https://doi.org/10.1021/acs.jproteome.9b00625.
(44) Shoulders, M. D.; Raines, R. T. Collagen Structure and Stability. *Annu. Rev. Biochem.* **2009**, *78* (1), 929–958. https://doi.org/10.1146/annurev.biochem.77.032207.120833.
(45) Gorres, K. L.; Raines, R. T. Prolyl 4-Hydroxylase. *Crit. Rev. Biochem. Mol. Biol.* **2010**, *45* (2), 106–124. https://doi.org/10.3109/10409231003627991.