RAId: Knowledge-Integrated Proteomics Web Service with Accurate Statistical Significance Assignment

Aleksey Y. Ogurtsov, Gelio Alves, and Yi-Kuo Yu*

Mass spectrometry-based proteomics starts with identifications of peptides and proteins, which provide the bases for forming the next-level hypotheses whose “validations” are often employed for forming even higher level hypotheses and so forth. Scientifically meaningful conclusions are thus attainable only if the number of falsely identified peptides/proteins is accurately controlled. For this reason, RAId continued to be developed in the past decade. RAId employs rigorous statistics for peptides/proteins identification, hence assigning accurate P-values/E-values that can be used confidently to control the number of falsely identified peptides and proteins. The RAId web service is a versatile tool built to identify peptides and proteins from tandem mass spectrometry data. Not only recognizing various spectra file formats, the web service also allows four peptide scoring functions and choice of three statistical methods for assigning P-values/E-values to identified peptides. Users may upload their own protein database or use one of the available knowledge integrated organismal databases that contain annotated information such as single amino acid polymorphisms, post-translational modifications, and their disease associations. The web service also provides a friendly interface to display, sort using different criteria, and download the identified peptides and proteins. RAId web service is freely available at https://www.ncbi.nlm.nih.gov/CBBresearch/Yu/raid

1. Introduction

In mass spectrometry (MS)-based proteomics, both hypotheses formed and conclusion drawn are based on identified peptides and proteins, underscoring the pressing need to have analysis software that can accurately control the number of falsely identified peptides and proteins. RAId was designed to address this issue. Extending the central limit theorem (CLT), the first version of RAId[1] can compute, for each candidate peptide, the corresponding P-value and E-value accurately in a spectrum-specific manner. Having accurate E-values assigned to identified peptides allows us to estimate the proportion of false discoveries confidently, see the Supporting Information for details. To extend accurate spectrum-specific statistics to scoring functions such as XCorr,[2] Hyperscore,[3] and Kscore[4] that sum fragment scores independently, we implement in RAId a dynamic programming algorithm that can efficiently score all possible peptides to generate a discrete score histogram.[5,6] Note that in our implementation of these scoring functions, heuristics that are not described and justified in original publications are omitted.[6] To provide accurate spectrum-specific statistics for scoring functions that are not necessarily sums of independent contributions, we make use of extreme value distribution (EVD) in RAId to assign statistical significance to identified peptides.[7] Accurate P-values/E-values assigned to identified proteins are computed by using a developed formalism for combining weighted P-values of identified peptides.[8,9] More details about the statistical methods employed can be found in the Supporting Information or at https://www.ncbi.nlm.nih.gov/CBBresearch/Yu/raid.

Users of RAId web service may query MS/MS spectra in RAId’s knowledge-integrated organismal protein databases or in a customized user-provided protein database. The web service currently offers 21 organismal protein databases that contain annotated information for single amino acid polymorphisms (SAPs), post-translational modifications (PTMs), and their disease associations, when available. As an example, the current Homo sapiens knowledge-integrated protein database contains 49 893 PTMs and 23 987 921 SAPs out of which 27 401 have disease associations. The format of the databases[10] facilitates efficient yet in-depth peptide search with on-the-fly scope expansion to include annotated SAPs and PTMs. Users also have the option for including novel PTMs during searches.

2. Implementation

Written in HTML and CSS, the RAId pages use JavaScript for communications between the user and the server. Perl and C++ are used to generate dynamic web content, incorporate
Figure 1. Panel (A) shows the main dialog window of RAId when the Generate histogram tab is chosen. To the right of the scoring function heading, users may select a scoring function. Note that the symbol $R(\bullet)$ represents RAId’s implementation of scoring function $\bullet$ minus the undescribed/ unjustified heuristics. The user may select desired PTMs alongside with the 20 regular amino acids to generate the score distribution (normalized histogram) for a given MS/MS spectrum. Under the “Amino acids and PTMs” expansion tab, one can click on “Change” next to PTMs, and a pop-up window, part of which shown in panel (B), appears which allows the user to select PTMs desired. Panel (C) displays the score distribution by scoring all possible peptides made of only regular amino acids. Panel (D) also displays the score distribution by scoring all possible peptides made of 20 regular amino acids and with 31 PTMs, some of which are shown in panel (B).

images/graphs on-the-fly, and maintain the job queue. To effectively communicate with the servers, the NCBI C++ toolkit (https://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/) is intensively used. The web service programs and scripts employ the NetSchedule and the NetCache servers of the National Center for Biotechnological Information (NCBI). The NetSchedule queues jobs and balances web servers’ load, while the NetCache stores uploaded and generated data.
Figure 2. Panel (A) shows the main dialog window of RAId when the Database search tab is chosen. Under the “Amino acids and PTMs” expansion tab, one can enable the search program to consider annotated SAPs, novel PTMs, and annotated PTMs. All of those can be accessed via pop-up windows by clicking on the corresponding “Change” buttons on the right. In the example shown in the lower right corner of panel (A), two annotated SAPs are selected. This means that when encountering peptides whose residue A (or S) have been documented to have SAPs, the search program will automatically consider all such variant peptides during the search.

Panel (B) shows the pop-up window associated with annotated PTMs. In this example, 31 annotated PTMs were checked. Once all search parameters are entered and the button “Submit job” clicked, the main dialog window is replaced by the result window. An example of the result window is displayed in panel (C) when a Homo sapiens dataset is used. The four job status, pending, running, retrieving, and complete are self-explanatory. Once the job is complete, one may sort the results using a different criterion by clicking on any one of the remaining five open circles.
Figure 3. Panel (A) shows what the results window is like when the user sorts the search results according to “protein E-value”. With this choice, proteins are sorted in ascending order of their E-values. That is, the protein with highest identification significance is shown first and so on. If one clicks on the “plus” sign in the front of a protein row, that row expands and all peptides mappable to that protein are displayed in ascending order of their E-values. Panel (B) shows such an expansion when the fifth protein on Panel (A) is clicked. The user may also in the expanded list click on one of the peptides; this will induce a pop-up window displaying the peptide-spectrum match. Panel (C) shows such an example when clicking on the peptide AVFQANQENLILKR belonging to the fifth protein. Finally, the user may access a protein’s corresponding RefSeq page by clicking on that protein’s GI. Panel (D) shows part of the RefSeq page corresponding to the fifth protein when its accession number NP_000692 is clicked.

3. Usage

The RAId web interface contains three tabs: Database search, Generate histogram, and Compute TNPP. Within each tab, contextual help is available upon clicking on an encircled question mark. At the bottom of each tab, there is an “example” button next to the “submit job” button. The former allows the user to populate the needed parameters for that tab with a click, making it easy to explore how the web service works.

The middle tab Generate histogram, shown in panel A of Figure 1, allows users to generate the discrete score distribution from scoring all possible peptides. By clicking on the...
“Upload” button, a user may upload a MS/MS spectrum via the pop-up window; acceptable spectral data formats are: DTA, PKL, MGF(GPM), MSP(NIST or ISB), mzXML, mzML, and mzData. The “MS/MS spectra mode” drop down window allows the user to specify whether the spectral data is in the profile mode or centroid mode. Although the user may simply proceed using the default parameters, RAId web service does accommodate customization. For example, the user may select a scoring function desired and fragmentation series to score. (Currently there are four scoring functions available: XCorr, Kscore, Hyperscore, and RAId score.) In addition, expanding the “more parameters” button allows users to specify Cysteine modification, molecular masses of the C- and the N-termini groups, and precursor/daughter ion mass error tolerance; expanding the “amino acids and PTMs” button, on the other hand, enables users to limit the amino acids present and to include the PTMs desired. Panel B of Figure 1 displays the pop-up window for selecting PTMs to include. This window can be reached by clicking on the “Change” button corresponding to PTMs under the “amino acids and PTMs” button. With the parameters specified or using the default values, the web service extracts the precursor ion mass from the spectrum file and then via dynamical programming generates[5,6] the score distribution from scoring all possible peptides that satisfy the conditions (e.g., the peptide mass range determined by the precursor ion mass error tolerance) prescribed under the “more parameters” and the “amino acids and PTMs” buttons. Panels (C, D) of Figure 1 display, using the same spectrum, the score distributions when no PTM and 31 PTMs are allowed. The score distribution obtained can be used to assign P-values/E-values to identified peptides.

The usage of the Compute TNPP tab is simple. After the user specifies the peptide mass and the conditions (under the “more parameters” and “amino acids and PTMs” buttons) such as precursor peptide mass error tolerance and allowed amino acids/PTMs, the web service computes the total number of possible peptides satisfying these conditions. Knowing the total number of possible peptides for a specified molecular mass range is helpful because its inverse can be used as a lower bound for the P-value, preventing exaggerated statistics for high-scoring peptides. The output information such as peptide mass, minimum/maximum peptide length, and the total number of possible peptides of the Compute TNPP tab is also displayed above the score histogram when one uses the Generate histogram tab to produce the score histogram with an input MS/MS spectrum. For example, see the text displayed above the score histograms in panels (C, D) of Figure 1.

Under the Database search tab, see panel (A) of Figure 2, users can submit either a single MS/MS spectrum or a set of MS/MS spectra resulting from an experiment for analysis after specifying at least the following parameters: MS/MS spectra file, database, protein digestion enzyme used, maximum number of missed cleavage sites allowed, peptide scoring function, fragmentation series to score. Note that aside from using one of our knowledge-integrated databases, the user may also upload and hence use a customized protein database (in fasta format), see the second entry under the Database search tab. Additional parameters that can be changed under the “more parameters” and the “amino acids and PTMs” buttons include molecular masses of chemical groups attached to the C- and the N-termini, precursor and product ions mass error tolerances, and in particular inclusion into search the annotated SAPs/PTMs, and/or novel PTMs. Panel (B) of Figure 2 shows the pop-up window that allows the user to select annotated PTMs to be considered during the searches. The feature of incorporating the annotated SAPs/PTMs into database searches evidently requires knowledge-integrated databases such as ours. The format of our knowledge-integrated databases[10] facilitates efficient yet in-depth peptide search with on-the-fly scope expansion to include annotated SAPs and PTMs. Users also have the option for including novel PTMs during searches.

In addition to the aforementioned spectral data format, RAId web service also recognizes the RAW data type which often results from the raw profile data of an experiment. This is very useful as it allows the user to upload the whole experimental data in one go with specified parameters. The results window, see panel (C) of Figure 2, will appear upon job submission. Within the results window, the four status indicators, pending, running, retrieving, and complete, are self-explanatory. Once the analysis is completed, the user can interactively filter and sort the obtained results (see panel (C) of the Figure 2) using the following criteria: MS/MS spectrum number, peptide E-value, protein’s best peptide E-value, the total number of hits per protein, the total number of unique peptides per protein, and the protein’s E-value.

Panel (A) of Figure 3 shows an example run whose results are sorted in ascending order of E-values of the proteins identified. If one clicks on the plus sign at the very beginning of a row, all identified peptides (with redundancy) mappable to the protein in that row are shown in ascending order of their E-values. Panel (B) of Figure 3 displays the expanded results when clicking on the fifth protein in panel (A). For each reported peptide the overlap between the theoretical fragments and the fragments in the filtered MS/MS spectrum is readily accessible by clicking on the reported peptide. Panel (C) of Figure 3 displays such an example when clicking on the peptide AVFQANQENLPILKR in panel (B), the expanded view of the fifth protein in panel (A). Proteins reported also have their accession numbers linked to the corresponding NCBI’s Entrez web pages. Panel (D) of Figure 3 displays the RefSeq page when the accession number of the fifth protein in panel (A) is clicked.

Finally, the RAId output files and the resulting tables can be downloaded as text files. Users can also retrieve recently processed jobs by clicking on the “Retrieve old runs” link on the right menu column of the web page.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This research was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health.
Conflict of Interest

The authors declare no conflict of interest.

Keywords

knowledge-integrated database, MS/MS data analyses, peptide/protein identifications, proportion of false discoveries, statistical significance assignment

Received: October 12, 2018
Revised: March 13, 2019

[1] G. Alves, A. Y. Ogurtsov, Y. K. Yu, Biology Direct 2007, 2, 25. [Online].
[2] C. Y. Park, A. A. Klammer, L. Kall, M. J. MacCoss, W. S. Noble, J. Proteome Res. 2008, 7, 3022.
[3] R. Craig, R. C. Beavis, Bioinformatics 2004, 20, 1466.
[4] B. MacLean, J. K. Eng, R. C. Beavis, M. McIntosh, Bioinformatics 2006, 22, 2830.
[5] G. Alves, Y. K. Yu, Physica A 2008, 387, 6538.
[6] G. Alves, A. Y. Ogurtsov, Y. K. Yu, PLoS ONE 2010, 5, e15438.
[7] G. Alves, Y. K. Yu, Bioinformatics 2016, 32, 2642.
[8] G. Alves, Y. K. Yu, PLoS ONE 2011, 6, e22647.
[9] G. Alves, Y. K. Yu, Bioinformatics 2015, 31, 699.
[10] G. Alves, A. Y. Ogurtsov, Y. K. Yu, BMC Genomics 2008, 9, 505.