RAId: User Guide

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

National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, MD 20894.

(Dated: January 31, 2018)

Contents

* to whom correspondence should be addressed: alves@ncbi.nlm.nih.gov,ogurtsov@ncbi.nlm.mih.gov,yyu@ncbi.nlm.nih.gov

I. RAId

RAId is a software designed to analyze MS/MS spectra and it provides peptide and protein identification with accurate statistical significance assignment[4]. The software is writing in C++ and implemented to execute in parallel in a single operating system taking advantage of multi-logical-cores when available. RAId has a graphical-user-interface (GUI) written in Java making it a user-friendly tool.

One of the advances offered by RAId is that for each identified peptide it reports an *E-value* which is subsequently used to assign statistical significance assignment during protein identification. RAId customized protein databases have an unique structure incorporating scientific information from already observed post-translational modifications (PTMs), single amino acid polymorphisms (SAPs) and their associated diseases when available[2]. Annotated protein databases from various species that can be used by RAId are available for download from **ftp://ftp.ncbi.nlm.nih.gov/pub/qmbp/qmbp_ms/RAId/RAId_Databases/**. RAId's unique database structure also allow users to construct specialized databases based on the user own knowledge related to PTMs, SAPs and diseases.

RAId offers the the following four scoring functions to score peptides: RAId_DbS's scoring function (RAId score)[1], Hyperscore[6], XCorr[5] and K-score[?] . Statistical significance to identified peptides can be computed by three different methods depending on the scoring function used. For RAId score scoring function statistical significance is computed using a theoretically derived parametric distribution based on the central limit theorem (CLT)[1]. For RAId score, XCorr, Hyperscore, and K-score peptide statistical significance can be assigned via extreme-value-distribution (EVD) or from all possible peptides statistics (APPS)[3].

A. Software Package

Software Site:
http://www.ncbi.nlm.nih.gov/CBBresearch/Yu/downloads/raid.html

Installation:
To install unzip and untar the file RAId.tar.gz.

```
$ gunzip RAId.tar.gz
$ tar -xvf RAId.tar
```

Compiling:

```
$ make -f RAId.mak
```

Executable:

```
$RAId
```

RAId GUI:

```
$java -jar RAId_GUI.jar
```

Databases:
RAId can perform searches in specially formatted protein databases. Special annotated and formatted protein databases for different organisms which are searchable by RAId are available for downloaded from the above ftp site. User can also generate their own database using the $-fp$ option available or they can create their own enhanced databases using the perl UserDb.pl provided.

The FASTA file used by RAId has to have the following format. **Fasta file** example:
> |*key*|Id_Seq1(sequence identifier)| sequence description.
MLLATLLLLLLGGALAHPDRIIFPNHACEDPPAVLLEVQGTLQRPLVRDSRTSPANCTWLILGSKEQTVT
. . .
> |*key*|Id_Seq2(sequence identifier)| sequence description.
MTGSERLTLRSPLQPLISLCEAPPSPLQLPGGNVTITYSYAGARAPMGQGFLLSYSQDWLMC

,**where the allowed values for *key* are: gi, sp, tr, ref, pdb**.

Example file:
The bash file raid_example.sh is an example of how to execute RAId. It contains someone of the syntax (parameters) that can be used to customize searches. One can also execute the bash file to verify that RAId is properly installed.

```
$./raid_example.sh
```

B. Syntax

$[-cg]$,
$[-daa]$, $[-db]$, $[-dsv]$, $[-dt]$,
$[-ect]$, $[-ed]$, $[-evc]$, $[-ex]$, $[-exf]$, $[-ez]$,
$[-fl]$, $[-fp]$, $[-fps]$,
$[-ip]$,
$[-mc]$, $[-mw]$,
$[-nc]$, $[-nd]$, $[-ng]$, $[-nmcs]$,
$[-of]$, $[-op]$, $[-opim]$
$[-pie]$, $[-pt]$, $[-pfd]$, $[-pfc]$,
$[-qf]$,
$[-rap]$, $[-ras]$, $[-rnp]$, $[-rtr]$,
$[-sm]$, $[-ssr]$, $[-ssk]$, $[-ssh]$, $[-ssx]$,
$[-v]$.

Executing mode option:
$[-ex]$

Enzyme option:
$[-ect]$,$[-ez]$

Cysteine modification option:
$[-mc]$

Molecular mass options:
$[-cg]$, $[-dt]$, $[-mw]$, $[-ng]$, $[-pie]$, $[-pt]$

Amino acid residue (PTMs,SAPs) options:
$[-daa]$, $[-rap]$, $[-ras]$, $[-rnp]$

Quantification options:
$[-qf]$, $[-rtr]$

Database options:
$[-db]$, $[-fp]$,

Scoring options:
$[-dsv]$, $[-evc]$, $[-sm]$,

Scoring series for different scoring functions:
$[-ssr]$, $[-ssk]$, $[-ssh]$, $[-ssx]$,

Output and Input file options:
$[-exf]$, $[-ip]$, $[-op]$

Number of logical cores:
$[-nc]$

MiCId option
$[-opim]$

C. Options

− cg

Chemical group attached to peptide *C-terminal*.
Default value: $-cg$ 17.002739
$-cg$ 17.002739 = Free Acid
$-cg$ 16.01872 = Amide
user can specify any molecular mass after the option $-cg$.

−daa

This option is used with RAId_aPS.
A list of the allowed residues to be used with RAId_aPS to generate the score histogram.
Default value:
$-daa\,[A00, G00, V00, L00, I00, P00, F00, Y00, W00, S00, T00, C00, M00, N00, Q00, D00, E00, K00, R00, H00]$.
Any of the amino acids and modifications presented in the file RAId_PTM_file are allowed choices .
The example below includes 2 PTMs G01 and G02
$-daa\,[A00, G00, G01, G02, V00, L00, I00, P00, F00, Y00, W00, S00, T00, C00, M00, N00, Q00, D00, E00, K00, R00, H00]$

−db

This option is used to specified the protein database to be searched.
$-db$ /path/database_name

−dsv

- Scoring function to score peptides using RAId_DbS or RAId_aPS algorithm.
Any combination of the different scoring functions separated by comma are allowed parameters for RAId_aPS. While for RAId_DbS selecting of only one single scoring function is allowed.
Default value: $-dsv$ 1.
Allowed options
$-dsv$ 0 = RAId score.
$-dsv$ 1 = K-score.
$-dsv$ 2 = Hyperscore.
$-dsv$ 3 = XCorr.

−dt

Product fragment ions mass accuracy δm (ppm).
Default value: $-dt$ 50.

−ect

Enzyme cleavage type. Peptide is fully-cleavaged or partially-cleavaged.
Default value: $-ect$ 0
$-ect$ 0 = Fully-cleavaged
$-ect$ 1 = Partially-cleavaged

−ed

User can used the $-ed$ to specify any experimental details to be included in the final output file.
Users must use quotation marks to specify the experimental details.
Example:
$-ed$ "Human liver cancer cell line study using ETD"

−evc

Maximum allowed peptide *E-value*.
Default value: $-evc$ 10.

−ex

RAId operation mode.
Default value: $-ex$ 1

−ex 0 = RAId_aPS mode. Generates the total number of possible peptides for a given precursor ion.
−ex 1 = RAId_DbS database search mode. Score statistics using saddle point approximation.
−ex 2 = RAId_aPS database search mode. Score statistics using all possible peptides.
−ex 3 = RAId_aPS mode. Generates the score distribution by scoring all possible possible peptides for a given precursor ion.
−ex 4 = RAId_DbS database search mode. Score statistics using extreme-value-theory.
−ex 5 = RAId protein identification mode.

−exf

Extracting MS/MS spectrum file option.
The *−exf* option will generate single MS/MS spectrum from a file containing multiple MS/MS spectra.
−exf file_name

−ez

Enzyme option.
Default value: *−ez* 1
−ez 1 = Trypsin (K,R)
−ez 2 = Lys-C (K)
−ez 3 = Arg-C (R)
−ez 4 = GluC-Phosphate (E,D)
−ez 5 = GluC-Bicarbonate (E)
−ez 6 = PepsinA (L,F)
−ez 7 = Chymotrypsin (F,Y,W,L)
−ez 8 = Cyanogen bromide (M)
−ez 9 = Cyanogen bromide + Trypsin (M,K,R)
−ez 10 = Chymotrypsin + Trypsin (F,Y,W,L,K,R)
−ez 11 = V8-DE (N,D,E,Q)
−ez 12 = V8-E (E,Q)
−ez 13 = Trypsin + PepsinA (K,R,F,L)
−ez 14 = Lys-N (K)
−ez 15 = Asp-N (D,N)
−ez 16 = Asp-N_ambic (D,E)

−fl

Specifying a list of files (separated by comma) of peptide identification done by RAId. The file list is used for protein identification.
−fl file1,file2,...,filen

−fp

Formatting database option.
The option *−fp* will generate a database that can be used by RAId_DbS and RAId_aPS from a file of protein sequences in FASTA format.
−fp /path/input_database_name /path/output_database_name

−ip

Input MS/MS spectrum file name together with directory path.
−ip /path/msms_filename

−mc

Cysteine modification options.
Default value: *−mc* C00 Unmodified Cysteine (103.009186 Da.).
Chemical group attached to the side chain of cysteine.
Other cysteine modifications can be found in the file RAId_PTM_file. If the user cysteine modification is not present in RAId_PTM_file the user can add to RAId_PTM_file the modified cysteine information value.
−mc C00 = Unmodified Cysteine (103.009186 Da.).

$-mc$ C31 = Carboxymethylation (161.014649 Da.).
$-mc$ C32 = Carbamidomethylation (160.030646 Da.).
$-mc$ C33 = Pyridylethylation (208.066421 Da.).

−mw

This option is used with RAId_aPS.
Molecular mass used to compute the total number of possible peptides using RAId_aPS when a MS/MS file is not available.
$-mw$ 2354.34 . Will compute the total number of peptides for the requested molecular mass.
The allowed molecular mass range for $-mw$ is between [57,5000].

−nc

Number of logical cores to be used. For optimum performance this number should be equal to the number of logical core available in the operation system.
Default value: $-nc$ 1

−nd

Number of random/decoy peptides. This used to estimate distribution parameters when employing extreme-value-statistics (EVD). Default value is set to score 100,000 random peptides per spectra to estimate EVD parameters.

−ng

Chemical group attached to peptide *N-terminal*.
Default value: $-ng$ 1.007825
User can specify any molecular mass after *-ng*.
Example:
$-ng$ 1.007825 = Hydrogen.
$-ng$ 43.01838 = Acetyl.

−nmcs

Number of missed-cleavaged sites allowed per peptide.
Default value: $-nmcs$ 2

−op

Output search results path.
$-op$ /path_name/

−of

Suffix used for output file name.
$-of$ suffix_name

−pie

Precursor monoisotopic mass isotope error.
This option correct for mass shift error when learning the precursor ion molecular mass. It expands measured m/z's with masses greater than 1000 Da into three masses the original measured mass, the original mass -1 Da and the original mass -2 Da.
Default value: $-pie$ 1.
$-pie$ 0 = will not correct for mass shift error
$-pie$ 1 = will correct for mass shift error

−pt

Precursor ion mass accuracy δm (ppm).
Default value: $-pt$ 100.
RAI will look for all masses within ± 3×(precursor ion mass tolerance).

−qf

This option is used to pass a list of files for quantification analysis. It takes as input files produced by RAId having .PROTEIN_QUANTIFICATION for the file suffix. The accept input has the following format:
$-qf$ Xfile-1, Xfile-2, . . . , Xfile-n; Yfile-1, Yfile-2, . . . , Yfile-n; Zfile-1, Zfile-2, . . . , Zfile-n
In the format above, files separated by colon are sample duplicates and the protein ion current from these files are averaged together and compared with the average protein ion current obtained from files separated by semi-colon.

−rap
Users can specify database annotated Post-Translational Modified (PTMs) residues during database searches.
Default value: $-rap$ NONE.
$-rap\ P01@N$ = will allowed annotated PTM of proline P01 at the N-terminal.
$-rap\ P02@C, K03@A$ = will allowed annotated PTMs of proline (P01) at the C-terminal and of lysine (K03) anywhere in the peptide.

−rnp
Users can specify novel Post-Translational Modified (PTMs) (not annotated in RAId's enhanced databases) residues during database searches.
Default value: $-rnp$ NONE.
$-rap\ P01@N$ = will allowed annotated PTM of proline P01 at the N-terminal.
$-rap\ P02@C, K03@A$ = will allowed annotated PTMs of proline (P01) at the C-terminal and of lysine (K03) anywhere in the peptide.

−ras
Users can specify database annotated Single Amino Acid Polymorphisms (SAPs) residues during database searches.
Default value: $-ras$ NONE.
Any of the 20 standard amino acids are allowed as parameter for the $-ras$ field.
$-ras$] = will search only for annotated SAP of proline (P), i.e amino acids residues that are annotated in the database with proline as possible SAP.
$-ras\ P, K$ = will search for annotated SAPs of proline and lysine.

−rtr
Retention time range in minutes, ± minutes, used to look for a peptide intensity in the MS1 spectrum. An significant peptide intensity is integrated over the specified retention time range. Default value: $-rtr$ 2.

−ssr
Fragmentation series used to by RAId scoring function.
Default value: Selected scoring function default files.
Any combination of the following are possible choices:
(b,b-H,b+H,a,a-H,a+H,c,b-NH3,b-NH3+H,b-H2O-H,y,y-H,y+H,x,z,y-NH3,y-H2O)
-ssr b,y,a-H = Scoring peptides using the b,y and a-H series.
-ssr a,c,y-H2O = Scoring peptides using the a,c and y-H2O series.

−ssh
Fragmentation series used to by RAId(Hyperscore) scoring function.
Default value: Selected scoring function default files.
Any combination of the following are possible choices:
(b,b-H,b+H,a,a-H,a+H,c,b-NH3,b-NH3+H,b-H2O-H,y,y-H,y+H,x,z,y-NH3,y-H2O)
-ssh b,y,a-H = Scoring peptides using the b,y and a-H series.
-ssh a,c,y-H2O = Scoring peptides using the a,c and y-H2O series.

−ssx
Fragmentation series used to by RAId(XCorr) scoring function.
Default value: Selected scoring function default files.

Any combination of the following are possible choices:
(b,b-H,b+H,a,a-H,a+H,c,b-NH3,b-NH3+H,b-H2O-H,y,y-H,y+H,x,z,y-NH3,y-H2O)
-ssx b,y,a-H = Scoring peptides using the b,y and a-H series.
-ssx a,c,y-H2O = Scoring peptides using the a,c and y-H2O series.

−ssk

Fragmentation series used to by RAId(K-score) scoring function.
Default value: Selected scoring function default files.
Any combination of the following are possible choices:
(b,b-H,b+H,a,a-H,a+H,c,b-NH3,b-NH3+H,b-H2O-H,y,y-H,y+H,x,z,y-NH3,y-H2O)
-ssk b,y,a-H = Scoring peptides using the b,y and a-H series.
-ssk a,c,y-H2O = Scoring peptides using the a,c and y-H2O series.

−sm

MS/MS data collection mode.
Default value: $-sm$ 1.
$-sm$ 0 = Profile mode.
$-sm$ 1 = Centroid mode.

−v

Output RAId code current version.

D. RAId Enhanced Organism Databases Status

In the table below are some examples of enhanced databases available for download to be used with RAId. The numbers in the table have not been updated they were taken from our 2008 publication^{2} and does not reflect the information content of the most up to date enhanced databases. Databases available for download are constantly updated with newly documented SAPs, PTMs and their associated disease when available.

Organism	DB_name	Protein	NP	NM	SP	SAPs	PTMs	DB_size (byte)
Homo sapiens	hsa	29284	35059	35031	15030	116073	84406	16,265,018
Anopheles gambiae	angam	12388	12719	12706	112	350	50	6,042,277
Arabidopsis thaliana	artha	29651	31740	31711	5527	5207	11977	12,318,213
Bos taurus	botau	23796	26504	26491	3979	3295	15810	11,188,490
Caenorhabditis elegans	caele	22563	23097	23097	2890	1045	7756	10,050,609
Canis familiaris	cafam	31705	33834	33821	528	2766	4196	18,458,474
Danio rerio	darer	31192	36150	36137	1552	7358	3841	14,477,794
Drosophila melanogaster	drmel	17232	20207	20207	2568	5611	9290	9,796,785
Equus caballus	eqcab	17300	17637	17624	171	485	1045	9,404,150
Gallus gallus	gagal	18154	18724	18681	1455	1109	6522	8,728,501
Macaca mulatta	mamul	32547	38141	38128	207	1370	1262	14,498,187
Mus musculus	mumus	28506	35503	35451	12170	27614	61684	14,363,491
Oryza sativa	orsat	26636	26784	26777	1205	1291	2182	10,679,924
Pan troglodytes	patro	41464	52130	52117	482	3721	3734	20,217,986
Plasmodium falciparum	plfal	5240	5267	5267	88	56	184	3,995,386
Rattus norvegicus	ranor	28914	39425	39389	5569	9297	33240	15,879,569
Saccharomyces cerevisiae	sacer	5699	5880	0	5807	5507	13220	2,927,330

Table 1. The header abbreviations in this table are explained as follows. The second column, headed by DB_name, documents the abbreviated database name for searches using standalone version of RAId. The column headed by “Protein” indicates the final number of protein clusters in the processed organismal databases. The columns headed by NP, NM, and SP summarize the break down of the total number of accession numbers included respectively from protein products, transcript products, and SwissProt protein entries. The columns headed by SAPs and PTMs indicate respectively the total number of annotated SAPs and PTMs included. The last column shows the database size in bytes.

Figure 1. - Information-preserved protein clustering example

```
       consensus seq.     ...DPR.........LQRLVADN⟨(N08)⟩GSE ...
          member seq.     ...DPR⟨{W00}⟩...LKRLVVDN⟨(N11)⟩GSE ...
updated consensus seq.  ...DPR⟨{W00}⟩...LQ⟨{K00}⟩RLVA⟨{V00}⟩DN⟨(N08,N11)⟩GSE...
```

Figure 1. Information-preserved protein clustering example. Once a consensus sequence is selected, members of a cluster are merged into the consensus one-by-one. This figure illustrates how the information of a member sequence is merged into the consensus sequence. Amino acid followed by two zeros indicates an annotated SAP. Every annotated PTM has a two-digit positive integer that is used to distinguish different modifications. The difference in the primary sequences between a member and the consensus introduces *cluster-induced* SAPs. In this example, the residues Q and A (in red) in the consensus are different from the residues K and V (in blue) in the member sequence. As a consequence, K becomes a cluster-induced SAP associated with Q and V becomes a cluster-induced SAP associated with A. The annotated SAP, ⟨{W00}⟩, associated with residue R in the member sequence is merged into the consensus sequence, see the updated consensus sequence in the figure. Note that the annotated PTM, ⟨(N11)⟩, associated with N in the member sequence is merged with a different annotated PTM, ⟨(N08)⟩, at the same site of the consensus sequence. Although, the SAPs, PTMs are merged, each annotation's origin and disease associations are kept in the processed definition file allowing for faithful information retrieval at the final reporting stage of the RAId's program.

Figure 2. - Structure of Enhanced Database.

```
...RTLVGLCKLG SAGGTD⟨{H00}⟩YGLR QFAEGSTEKL ....................................[
...IEYISYFWVI GN⟨(N08,N09,N10,N11,N12)⟩QSSMWFAT SLSIFYFLKI ANFSNYIFLW LKSRTNMVLP
   FMIVFLLISS LLNFAYIAKI LNDYKT⟨{M00}⟩KN⟨(N08,N09,N10,N11,N12)⟩DT VWDLNMYKSE ...[
```

Figure 2. Consensus protein sequences NP_775259 (first line, residues 480 – 510 shown) and NP_076410 (second and third lines, residues 81 – 170 shown) are used as examples to demonstrate the structure of our sequence file, part of the enhanced database. A "[" character is always inserted after the last amino acid of each protein to serve as a separator. Annotated SAPs and PTMs associated with an amino acid are included in a pair of angular brackets following that amino acid. SAPs are further enclosed by a pair of curly brackets while PTMs are further enclosed by a pair of round brackets. Amino acid followed by two zeros indicates an annotated SAP. Every annotated PTM has a two-digit positive integer that is used to distinguish different modifications.

Figure 3. - Illustration of Database Compression.

```
(A)
...LEVRQGTLQPLVR⟨{W00}⟩DRSPM⟨{V00}(M01)⟩CTWLILGSKEQTVTIR ...

(B)
...LEVRQGTLQPLVRDRSPMCTWLILGSKEQTVTIR ......
JQGTLQPLVRSRSPVCTWLILGSKJQGTLQPLVRSRSPmCTWLILGSKJQGTLQPLVWSRSPMCTWLILGSK
JQGTLQPLVWSRSPVCTWLILGSKJQGTLQPLVWSRSPmCTWLILGSK
```

Figure 3. In this example, the sequence has two nearby variable sites with residues R and M colored in red. Residue R may be replaced by a residue W due to a possible SAP; while residue M may be replaced by a residue V or an acetylated methionine (M01, in our notation) due to respectively a possible SAP or PTM. This information is encoded in our sequence file as shown in part (A). To encode the same information, method proposed in reference[7] would have up to five additional highly similar peptides separated by a letter "J" appended to the end of the primary sequence, see part (B). Here a lower case m is used to denote the acetylated methionine. Another key difference in the two methods shown above is on the limit of allowed number of enzymatic miscleavages. In our method, there is no limit on the number of allowed miscleavages, while in other approaches, the number of miscleavages is usually set to below a certain threshold. As an example, in our method, the variant peptides SPVCTWLILGSKEQTVTIR and SPmCTWLILGSKEQTVTIR are already included in (A). But in the approach of reference[7], in order to allow consideration of this variant peptide, one either needs to additionally append this peptide or to have much longer flanking peptides than shown in (B).

E. Database Formatting

If users want to use a different database they can do so by first formatting the database. The database to be format has to be a file in **FASTA format** and the database can be easily format by using $-fp$ option.

The FASTA file used has to have the following format.
Fasta file example:
$>$ $|key|$Id_Seq1(sequence identifier)$|$ sequence description.
MLLATLLLLLLGGALAHPDRIIFPNHACEDPPAVLLEVQGTLQRPLVRDSRTSPANCTWLILGSKEQTVT
, where the allowed values for *key* are: gi, sp, tr, ref, pdb.

Example:

```
./RAId   −fp   /path/input_database_filename   /path/output_database_filename
```

Formatting the database will produce four files:
output_database_filename.def, **output_database_filename.inf**, **output_db_filename.prs**, **output_db_filename.seq**.

F. User Enhanced Database Formatting

RAId also permits users to create their own enhanced database. To generate a user enhanced database the user need to create two files: a FASTA file containing the protein sequences of interest and a second file containing the user expertise/knowledge of these proteins PTMs, SAPs and diseases.

The FASTA file used by RAId has to have the following format.
Fasta file example:
> |*key*|Id_Seq1(sequence identifier)| sequence description.
MLLATLLLLLLGGALAHPDRIIFPNHACEDPPAVLLEVQGTLQRPLVRDSRTSPANCTWLILGSKEQTVT
IRFQKLHLACGSERLTLRSPLQPLISLCEAPPSPLQLPGGNVTITYSYAGARAPMGQGFLLSYSQDWLMC
LQEEFQCLNHRCVSAVQRCDGVDACGDGSDEAGCSSDPFPGLTPRPVPSLPCNVTLEDFYGVFSSPGYT
. . .

> |*key*|Id_Seq2(sequence identifier)| sequence description.
MTDFFFTHIIFPNHACEDPPAVLLEVQGTLQRPLVRDSRTSPANCTWLILGSKEQTVT
GSERLTLRSPLQPLISLCEAPPSPLQLPGGNVTITYSYAGARAPMGQGFLLSYSQDWLMC
AVQRCDGVDACGDGSDEAGCSSDPFPGLTPRPVPSLPCNVTLEDFYGVFSSPGYT
. . .
, where the allowed values for *key* are: gi, sp, tr, ref, pdb.

Knowledge file example:
> |*key*|Id_Seq1

48	SAP	R	W	deadly cancer
56	PTM	N	N08,N09,N10,N11,N12	
111	PTM	N	N08,N09,N10,N11,N12	
139	SAP	M	V	diabetes
193	SAP	N	L,I,V	
193	PTM	N	N08	
299	PTM	N	N08,N09,N10,N11,N12	
365	SAP	A	T	color blind
434	SAP	S	C,T,V,P	insulin dependent diabetes
558	SAP	R	H,P,W	

The **knowledge file** structure is explained below.
> |*key*|seq_identifier
First column field is the residue position.
Second column field signifies a SAP or PTM.
Third column field is the original residue present in the sequence.
Fourth column field is either a list of possible SAPs (L,I,V) or a list of possible PTMs (N08,N09,N10,N11,N12)
Fifth column field is the disease name if any at the given position.

Once the user has created the two files as described above the user can generate a knowledge database that RAId can process by executing the UserDb.pl script as shown below.

Example:
```
./UserDb.pl    fasta_file_name    knowledge_file_name output_format_database_name
```
The output_format_database_name is the database that can be processed by RAId.

G. Post-Translation Modifications (PTMs) File

RAId_PTM_file is the file that contains information related to amino acid residues and their corresponding post-translational modifications. The user can add any new post-translational modification to this file as long as one keeps with the same annotation structure shown below.

Line Code	Description
ID	Chemical Name of Amino Acid/PTM
AC	Residue Key
TG	Target Unmodified Amino Acid
RW	Unmodified Amino Acid Molecular Mass
MW	Modified Amino Acid Molecular Mass
PA	Location of the Modification in the Amino Acid Residue
PP	Position of the Amino Residue in the Peptide
CF	Chemical Modification to the Amino Acid Residue
MM	Monoisotopic Mass Difference MM=MW-RW
KY	Other Common Names Used to Identify the Same Molecule
LT	Other Terms Found In Literature not Necessary Correct Names

Some examples of the addition of new residues to the **RAId_PTM_file** file.

ID	Cholesterol glycine ester
AC	G01
TG	Glycine
RW	57.021465
MW	425.365766
PA	Amino acid backbone.
PP	C-terminal.
CF	C27 H44
MM	368.344301
KY	Lipoprotein.
LT	None

ID	N-palmitoyl cysteine
AC	C06
TG	Cysteine
RW	103.009186
MW	341.238852
PA	Amino acid backbone.
PP	N-terminal.
CF	C16 H30 O1
MM	238.229666
KY	Lipoprotein; Palmitate; Palmitoylation.
LT	Polmitoylation

II. RAID COMMAND LINE EXECUTION EXAMPLES

A. RAId_DbS Command Line Examples

Example 1: RAId in database search mode with some search options $-xxx$.

Command line:

```
>./RAId −ex 1 −ez 1 −nmcs 3 −nc 10 −dt 50 −pt 5 −ng 1.007825 −cg 17.002739 −evc
10 −mc C32 −ssr b,y,c,z −rap NONE −ras NONE −rnp S06,T10 −db /path/database_name −ip
/path/msms_filename −op /path/ −of output_file_name.
```

The example above would execute RAId in database search mode:

RAId_DbS (parametric distribution based on the central limit theorem) $-ex$ 1.

Trypsin as the enzyme $-ez$ 1.

Number of allowed missed cleavage sites 3 $-nmcs$ 3.

Number of logical cores $-nc$ 10.

Molecular error tolerance of product ion 50 ppm. $-dt$ 50.

Molecular error tolerance of precursor ion 5 ppm. $-pt$ 5.

N-terminal group hydrogen $-ng$ 1.0078.

C-terminal group free acid $-cg$ 17.0027.

Maximum *E-value* allowed for reported peptide $-evc$ 10.

Cysteine modification $-mc$ C32.

Fragmented series used to score peptide $-ssr$ b, y, c, z.

Information from annotated post-translation modifications off $-rap$ NONE.

Information from annotated single amino acid polymorphisms off $-ras$ NONE.

All amino acid residues of serine and tyrosine are considered as modified residues $-rnp$ S06,T10.

Protein database path location $-db$ /path/database_name.

Input MS/MS spectrum file path location $-ip$ /path/msms_filename.

Search result output path location $-op$ /path/.

Output file name $-of$ output_file_name.

Example 2: RAId in database search mode with some search options $-xxx$.

Command line:

```
    >./RAId −ex 1 −ez 1 −nc 4 −dt 50 −pt 5 −ng 1.007825 −cg 17.002739 −evc 1 −mc C32 −ssr b, y
−rap NONE −ras S,T −rnp S06,T10 −db /path/database_name −ip /path/msms_filename −op /path/
−of output_file_name.
```

The example above would execute RAId in database search mode:

RAId_DbS (parametric distribution based on the central limit theorem) $-ex$ 1.

Trypsin as the enzyme $-ez$ 1.

Number of logical cores $-nc$ 4.

Molecular error tolerance of product ion 50 ppm. $-dt$ 50.

Molecular error tolerance of precursor ion 5 ppm. $-pt$ 5.

N-terminal group hydrogen $-ng$ 1.0078.

C-terminal group free acid $-cg$ 17.0027.

Maximum *E-value* allowed for reported peptide $-evc$ 1.

Cysteine modification $-mc$ C32.

Fragmented series used to score peptide $-ssr$ b, y.

Information from annotated post-translation modifications not used $-rap$ NONE.

Annotated single amino acid polymorphisms that mutates into serine and tyrosine are used $-ras$ S,T.

All amino acid residues of serine and tyrosine are considered as modified $-rnp$ S06,T10.

Protein database path location $-db$ /path/database_name.

Input MS/MS spectrum file path location $-ip$ /path/msms_filename.

Search result output path location $-op$ /path/.

Output file name $-of$ output_file_name.

Example 3: RAId in database search mode with some search options $-xxx$.

Command line:

```
>./RAId −ex 4 −dsv 2 −ez 1 −nc 4 −dt 50 −pt 5 −ng 1.007825 −cg 17.002739 −evc 10 −mc C32
−ssr b,y,c,z −rap NONE −ras S,T −rnp S06,T10 −db /path/database_name −ip /path/msms_filename
−op /path/ −of output_file_name.
```

The example above would execute RAId in database search mode:

RAId_DbS(statistics computed by extreme-value-theory): $-ex$ 4.

Scoring function Hyperscore $-dsv$ 2.

Trypsin as the enzyme $-ez$ 1.

Number of logical cores $-nc$ 4.

Molecular error tolerance of product ion 50 ppm. $-dt$ 50.

Molecular error tolerance of precursor ion 5 ppm. $-pt$ 5.

N-terminal group hydrogen $-ng$ 1.0078.

C-terminal group free acid $-cg$ 17.0027.

Maximum *E-value* allowed for reported peptide $-evc$ 10.

Cysteine modification $-mc$ C32.

Fragmented series used to score peptide $-ssr$ b, y, c, z.

Information from annotated post-translation modifications off $-rap$ NONE.

Annotated single amino acid polymorphisms that mutates into serine and tyrosine are used $-ras$ S,T.

All amino acid residues of serine and tyrosine are considered as modified residues $-rnp$ S06,T10.

Protein database path location $-db$ /path/database_name.

Input MS/MS spectrum file path location $-ip$ /path/msms_filename.

Search result output path location $-op$ /path/.

Output file name $-of$ output_file_name.

B. RAId_aPS Command Line Examples

Example 1: Computing the total number of possible peptides within a given molecular mass.

Command line:

```
>./RAId −ex 0 −ez 1 −dt 50 −pt 5 −ng 1.007825 −cg 17.002739 −mc C32 −ip
/path/msms_filename −op /path/ −of output_file_name.
```

RAId_aPS executing mode $-ex$ 0.

Trypsin as the enzyme $-ez$ 1.

Molecular error tolerance of product ion 50 ppm. $-dt$ 50.

Molecular error tolerance of precursor ion 5 ppm. $-pt$ 5.

N-terminal group hydrogen $-ng$ 1.0078.

C-terminal group free acid $-cg$ 17.0027.

Cysteine modification $-mc$ C32.

Input MS/MS spectrum file path location $-ip$ /path/msms_filename.

Search result output path location $-op$ /path/.

Output file name $-of$ output_file_name.

Example 2: Generating the score distribution for all possible peptides.

Command line:

```
    >./RAId −ex 3 −ez 1 −daa [A00,G00,G02,G03,V00,L00,F00,Y00,W00,S00,T00,C00,N00,Q00,D00,E00,R00]
−dt 50 −pt 5 −ng 1.007825 −cg 17.002739 −sc b,y −dsv 4 −ip /path/msms_filename −op /path/ −of
output_file_name.
```

DeNovo executing mode $-ex$ 3.

Trypsin as the enzyme $-ez$ 1.

Amino acids residues selected:
$-daa$ [A00,G00,G02,G03,V00,L00,F00,Y00,W00,S00,T00,C00,N00,Q00,D00,E00,R00].

Molecular error tolerance of product ion 50 ppm. $-dt$ 50.

Molecular error tolerance of precursor ion 5 ppm. $-pt$ 5.

N-terminal group hydrogen $-ng$ 1.0078.

C-terminal group free acid $-cg$ 17.0027.

Fragmented series used to score peptide $-sc$ b, y.

Scoring function XCorr selected to compute score histogram $-dsv$ 4.

Input MS/MS spectrum file path location $-ip$ /path/msms_filename.

Search result output path location $-op$ /path/.

Output file name $-of$ output_file_name.

Example 3: Using RAId_aPS as a database search tool.

Command line:

```
>./RAId −ex 2 −dsv 1,2,3 −ez 1 −nc 10 −dt 50 −pt 5 −ng 1.007825 −cg 17.002739 −evc
10 −mc C32 −ssr b,y,c,z −rap NONE −ras NONE −rnp S06,T10 −db /path/database_name −ip
/path/msms_filename −op /path/ −of output_file_name.
```

The example above would execute RAId in database search mode:

Database search execution mode RAId_aPS $-ex$ 2.

Scoring function used RAId score, Hyperscore and XCorr $-dsv$ 1,2,3.

Trypsin as the enzyme $-ez$ 1.

Number of logical cores $-nc$ 10.

Molecular error tolerance of product ion 50 ppm. $-dt$ 50.

Molecular error tolerance of precursor ion 5 ppm. $-pt$ 5.

N-terminal group hydrogen $-ng$ 1.0078.

C-terminal group free acid $-cg$ 17.0027.

Maximum *E-value* allowed for reported peptide $-evc$ 10.

Cysteine modification $-mc$ C32.

Fragmented series used to score peptide $-ssr$ $\;b, y, c, z$.

Information from annotated post-translation modifications off $-rap$ NONE.

Information from annotated single amino acid polymorphisms off $-ras$ NONE.

All amino acid residues of serine and tyrosine are considered as modified residues $-rnp$ S06,T10.

Protein database path location $-db$ /path/database_name.

Input MS/MS spectrum file path location $-ip$ /path/msms_filename.

Search result output path location $-op$ /path/.

Output file name $-of$ output_file_name.

C. RAId Command Line Protein Identification Example

Command line:

```
>./RAId −ex 5 −fl file1,file2,file3 −op /path/ −of output_file_name.
```

The example above would execute RAId in database search mode:

Performing protein identification using a list of files contained peptide identification performed by RAId mode $-ex$ 5.

List of files separated by comma $-fl$ file1,file2,file3.

Directory where files (file1,file2,file3) are located $-op$ /path/.

Output file name $-of$ output_file_name.

III. UTILIZING THE RAID GUI

The RAId GUI provides useful visual display tools that can be used to produce data tables and figures in a concise, easy to use, package. Data tables can be exported to TSV (tab separated values) and CSV (Comma separated values) files , while figures generated by the software can be exported to PNG (Portable networks graphics) or PDF (Portable Document Format) files. TSV, CSV, and PNG files are generated using prewritten Java utilities; however, PDF files are generated using the iText PDF generation package.

A. Testing

When the RAId GUI is downloaded there is a test data set that is included in the download file. To run this select the RAId button on the menu bar then from the drop down menu select *Test - Run Job*. This will submit a job with the sample data and generate sample results.

B. RAId

The RAId button on the menu bar has a few different options. The first is the test option that was mentioned above. The RAId menu also contains a link to the user manual and can be accessed by clicking *User Guide* from the RAId drop down menu. This opens up the manual with a default PDF viewer in Linux. The user can customize which PDF viewer to use by clicking on *Options - Configure RAId* from the RAId drop down menu. A text bar will appear where the user can choose which PDF viewer that wish to use.

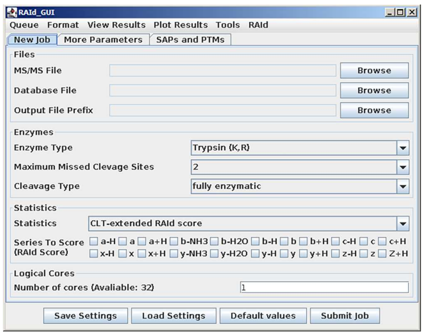
C. Submitting Jobs and Observing their progress using the RAId GUI

The RAId GUI provides a variety of in depth features that allow the user to customize and easily visualize how their specifications have an effect on the way MS/MS files are interpreted by the RAId identification software.

1. Generating a RAId Database (*.seq*) file

In order to run a Job using the RAId software protein files must be reformatted to fit the RAId standards. This can be accomplished by opening the *Format database file* option under the Format tab of the menu option. Once selected a pop up menu will appear prompting the user for a protein file fasta format and a preferred prefix.

2. Setting up and launching a RAId Job


RAId_GUI
Queue Format View Results Plot Results Tools RAId
New Job
More Parameters
SAPs and PTMs
Files
Browse
Browse
Browse
Enzymes
Statistics
Series To Score (RAId Score)
□ a-H □ a □ a+H □ b-NH3 □ b-H2O □ b-H □ b □ b+H □ c-H □ c □ c+H
□ x-H □ x □ x+H □ y-NH3 □ y-H2O □ y-H □ y □ y+H □ z-H □ z □ Z+H
Logical Cores
Save Settings
Load Settings
Default values
Submit Job

MS/MS File:
Database File:
Output File Prefix:
Enzyme Type: Trypsin (K,R)
Maximum Missed Clevage Sites: 2
Cleavage Type: fully enzymatic
Statistics: CLT-extended RAId score
Number of cores (Avaliable: 32): 1

Figure 4. Picture depicting the data specification tab of the RAId job submission software

Files:

MS/MS File: Click the *Browse* button next to select and load an experimental data file.
-Accepted file extensions: *.dta*, *.pkl*, *.mgf*, *mzData*, *.mzML*, *.mzXML*, *.raw*

Database File: Click the *Browse* button next to the Database File Field to select and load a RAId database file.
-Accepted file extensions: *.seq*

Ouput File Prefix: Click the *Browse* button to specify a prefix and the output files' destination.

Enzymes:

Enzyme Type: Select the enzyme type used to digest the protein.

Maximum Missed Cleavage Sites: Select maximum number of missed cleavage sites.

Cleavage Type: Select whether proteins are to be considered fully enzymatic or semi-enzymatic.

Statistics:

Statistics: Select a scoring function with its statistical method for the RAId software to use.

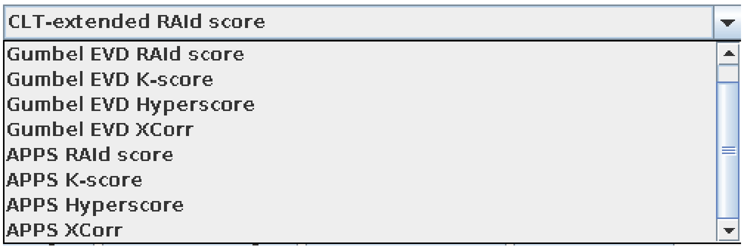


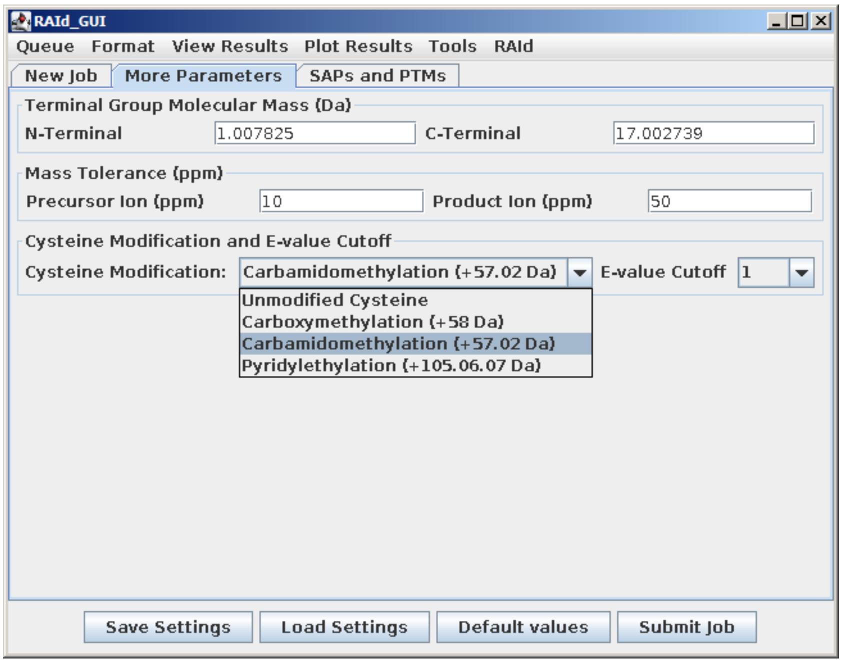
Figure 5. Picture of scoring functions along with its statistical method as they appear in the GUI

Series to score: Select peptide fragmentation series that should be used by the selected scoring function.

Logical Cores:

Number of cores: Specify the number of cores to submit the job to. (maximum varies based on device)

Specifying More Parameters:



Terminal Group Molecular Mass (Da)

N-Terminal 1.007825 C-Terminal 17.002739

Mass Tolerance (ppm)

Precursor Ion (ppm) 10 Product Ion (ppm) 50

Cysteine Modification and E-value Cutoff

Cysteine Modification: Carbamidomethylation (+57.02 Da) E-value Cutoff 1

Figure 6. Picture of the *More Parameters* tab of the Job Submission interface for the RAId GUI

Terminal Group Molecular Mass:

N − Terminal: Specify the molar mass of the N-Terminal group.

C − Terminal: Specify the molar mass of the C-Terminal group.

Mass Tolerance:

Precursor Ion (ppm): Specify the molecular error tolerance of precursor ion in parts per million (ppm).

Product Ion (ppm): Specify the molecular error tolerance of product ion (ppm).

Cysteine Modification and E-value Cutoff:

Cysteine Modification: Specify the modification of cysteine.

E − value Cutoff: Select a minimum E-value to be accepted during RAId's execution.

Selecting SAP's and PTM's:

N − Terminal: PTM occurs only if this amino acid is at the N-terminal end of the peptide.

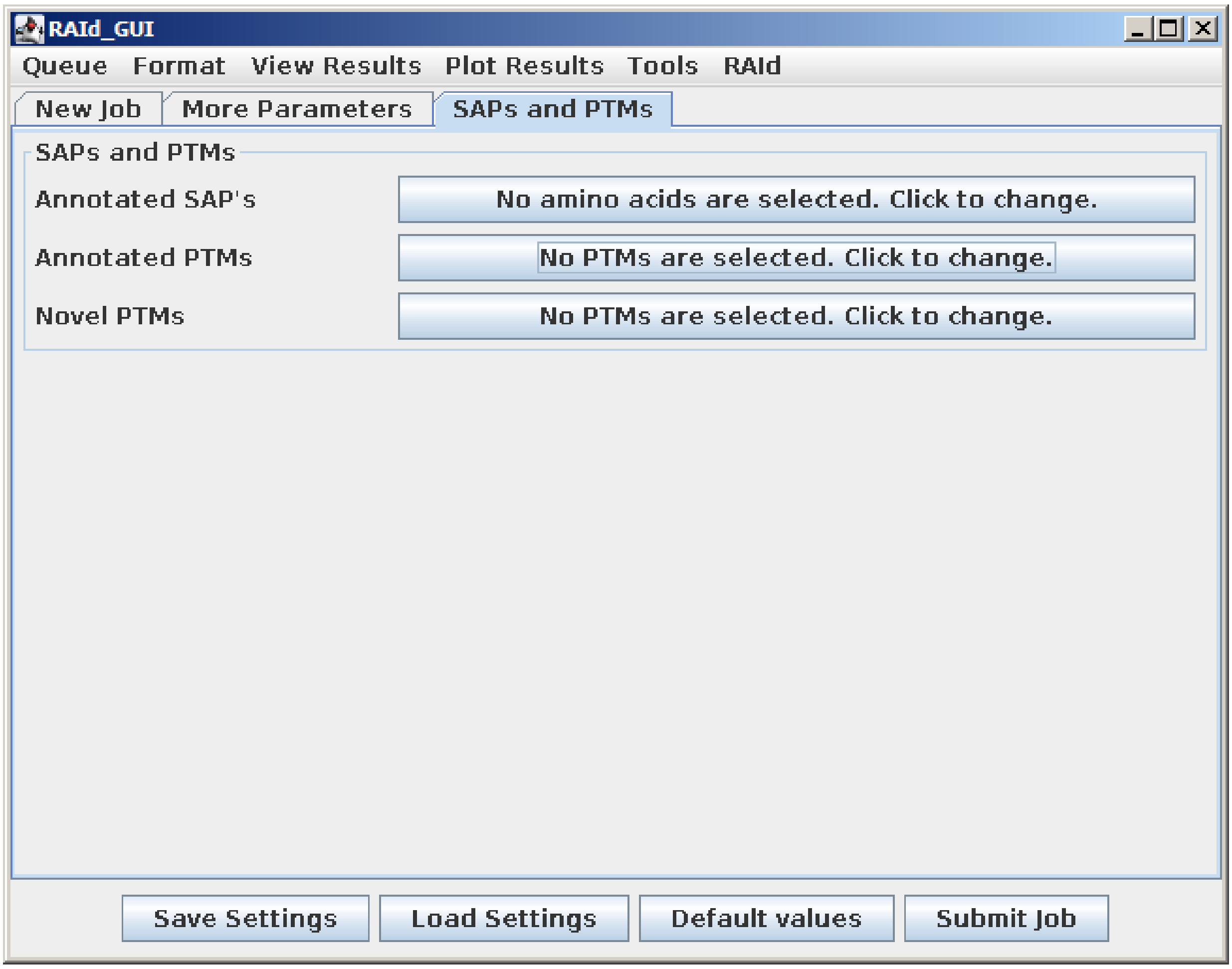
C − Terminal: PTM occurs only if this amino acid is at the C-terminal end of the peptide.

Any Position: PTM may occur at any position along the peptide.

Annotated SAP's: Click the corresponding button and check the boxes of all desired SAP's.

Annotated PTM's: Click the corresponding button and specify desired PTM's and their positions.

Novel PTM's: Click the corresponding button and specify desired PTM's and their positions.



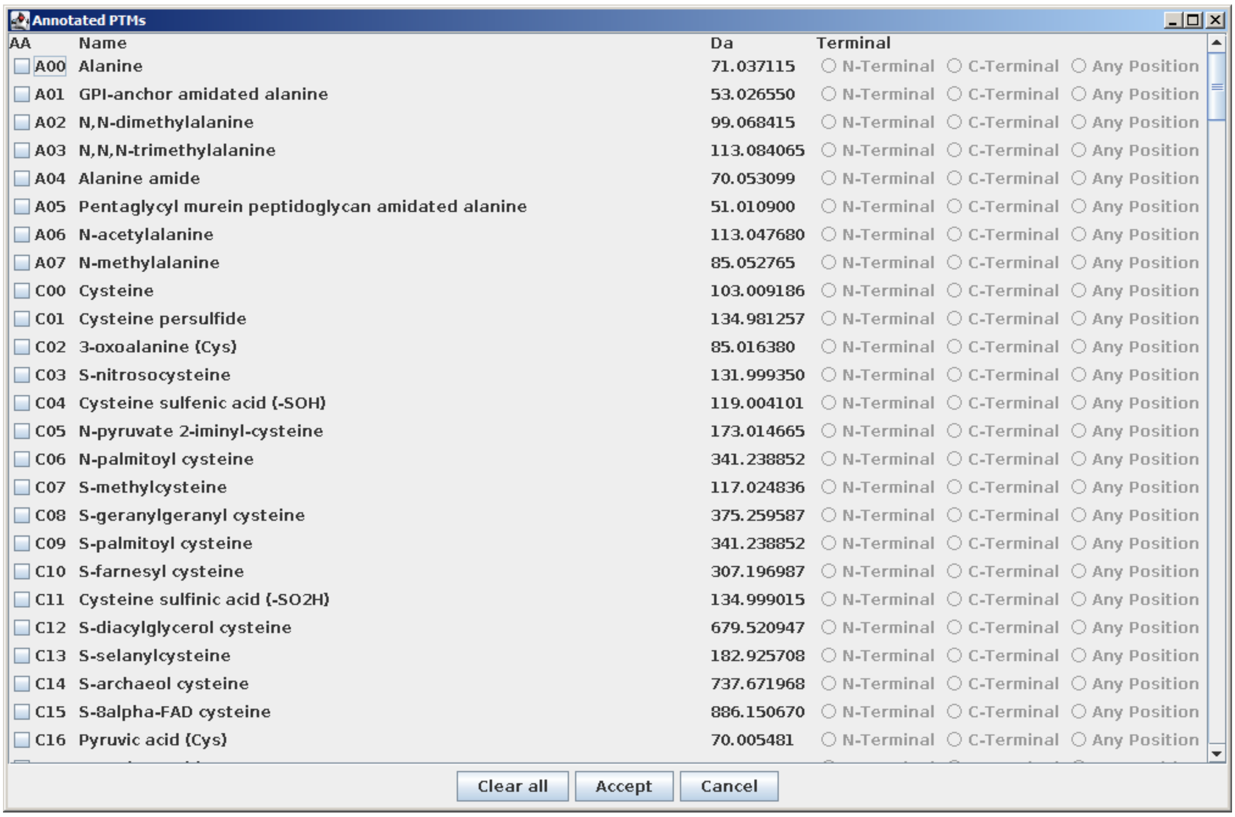
SAPs and PTMs

Annotated SAP's: No amino acids are selected. Click to change.

Annotated PTMs: No PTMs are selected. Click to change.

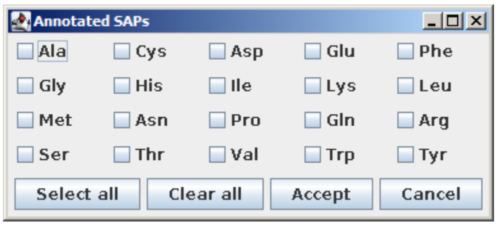
Novel PTMs: No PTMs are selected. Click to change.

Figure 7. Picture of the SAPs and PTM's tab of the Job Submission interface for the RAId GUI



AA		Name	Da	Terminal		
□	A00	Alanine	71.037115	○ N-Terminal	○ C-Terminal	○ Any Position
□	A01	GPI-anchor amidated alanine	53.026550	○ N-Terminal	○ C-Terminal	○ Any Position
□	A02	N,N-dimethylalanine	99.068415	○ N-Terminal	○ C-Terminal	○ Any Position
□	A03	N,N,N-trimethylalanine	113.084065	○ N-Terminal	○ C-Terminal	○ Any Position
□	A04	Alanine amide	70.053099	○ N-Terminal	○ C-Terminal	○ Any Position
□	A05	Pentaglycyl murein peptidoglycan amidated alanine	51.010900	○ N-Terminal	○ C-Terminal	○ Any Position
□	A06	N-acetylalanine	113.047680	○ N-Terminal	○ C-Terminal	○ Any Position
□	A07	N-methylalanine	85.052765	○ N-Terminal	○ C-Terminal	○ Any Position
□	C00	Cysteine	103.009186	○ N-Terminal	○ C-Terminal	○ Any Position
□	C01	Cysteine persulfide	134.981257	○ N-Terminal	○ C-Terminal	○ Any Position
□	C02	3-oxoalanine (Cys)	85.016380	○ N-Terminal	○ C-Terminal	○ Any Position
□	C03	S-nitrosocysteine	131.999350	○ N-Terminal	○ C-Terminal	○ Any Position
□	C04	Cysteine sulfenic acid (-SOH)	119.004101	○ N-Terminal	○ C-Terminal	○ Any Position
□	C05	N-pyruvate 2-iminyl-cysteine	173.014665	○ N-Terminal	○ C-Terminal	○ Any Position
□	C06	N-palmitoyl cysteine	341.238852	○ N-Terminal	○ C-Terminal	○ Any Position
□	C07	S-methylcysteine	117.024836	○ N-Terminal	○ C-Terminal	○ Any Position
□	C08	S-geranylgeranyl cysteine	375.259587	○ N-Terminal	○ C-Terminal	○ Any Position
□	C09	S-palmitoyl cysteine	341.238852	○ N-Terminal	○ C-Terminal	○ Any Position
□	C10	S-farnesyl cysteine	307.196987	○ N-Terminal	○ C-Terminal	○ Any Position
□	C11	Cysteine sulfinic acid (-SO2H)	134.999015	○ N-Terminal	○ C-Terminal	○ Any Position
□	C12	S-diacylglycerol cysteine	679.520947	○ N-Terminal	○ C-Terminal	○ Any Position
□	C13	S-selanylcysteine	182.925708	○ N-Terminal	○ C-Terminal	○ Any Position
□	C14	S-archaeol cysteine	737.671968	○ N-Terminal	○ C-Terminal	○ Any Position
□	C15	S-8alpha-FAD cysteine	886.150670	○ N-Terminal	○ C-Terminal	○ Any Position
□	C16	Pyruvic acid (Cys)	70.005481	○ N-Terminal	○ C-Terminal	○ Any Position

Figure 8. Annotated or Novel PTM's selection interface



☐ Ala	☐ Cys	☐ Asp	☐ Glu	☐ Phe
☐ Gly	☐ His	☐ Ile	☐ Lys	☐ Leu
☐ Met	☐ Asn	☐ Pro	☐ Gln	☐ Arg
☐ Ser	☐ Thr	☐ Val	☐ Trp	☐ Tyr

Figure 9. Picture of the Annotated SAPs selection interface

Button Options:

Save Settings: Saves all settings selected by the user to a properties file.

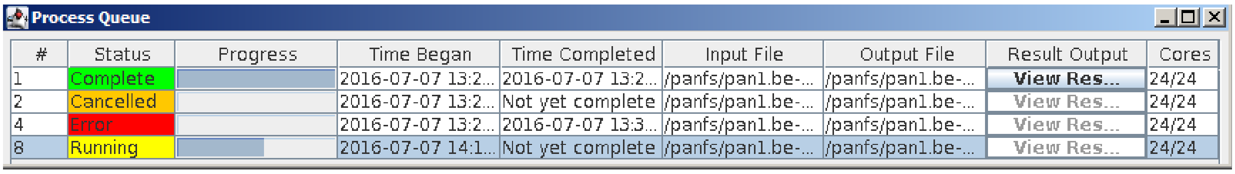
Load Settings: Loads a user selected properties file to display settings on the GUI.

Default values: Restores all fields to the values contained in them when the GUI was first loaded.

Submit Job: Initiates a job for the RAId tandem Identification software with the user specified options.

3. Viewing the progress of submitted jobs

Process Queue:



#	Status	Progress	Time Began	Time Completed	Input File	Output File	Result Output	Cores
1	Complete		2016-07-07 13:2...	2016-07-07 13:2...	/panfs/pan1.be-...	/panfs/pan1.be-...	View Res...	24/24
2	Cancelled		2016-07-07 13:2...	Not yet complete	/panfs/pan1.be-...	/panfs/pan1.be-...	View Res...	24/24
4	Error		2016-07-07 13:2...	2016-07-07 13:3...	/panfs/pan1.be-...	/panfs/pan1.be-...	View Res...	24/24
8	Running		2016-07-07 14:1...	Not yet complete	/panfs/pan1.be-...	/panfs/pan1.be-...	View Res...	24/24

Figure 10. Picture of the Process Queue Window for the RAId GUI

The process queue is opened after submitting a job, or by selecting the *Process Queue* option from the Queue Menu Item in the menu bar on the main interface. The process queue window displays The status, Progress, start and end time, input file, output file, cores being used, and a button to open a result viewer (If the job completed successfully). There will never be more than one process queue open at a time.

Right click a process to open a menu with the following options:

Restart Task: Select a process that has completed, was canceled, or failed, and restart it.

Force Start: Immediately allocates resources to start a process currently waiting.

Cancel: Cancels the selected process.

Cancel and Remove Task: Cancels and removes the selected process from the process queue.

Remove Task: Removes a task from the process queue.

A process can be in one of six states:

Waiting: The Process is currently waiting for resources to become available.

Starting: The RAId code is preparing to start analyzing the input files.

Running: The RAId code is running a process.

Finishing: The RAId code is currently saving all data to the corresponding output files.

Cancelled: The user canceled the process during it's run time.

Error: A fatal error occurred while the RAId software was running.

D. Viewing the data results of the RAId Assay

A user may view the results of a successful RAId execution by either clicking the $ViewResults$ button in the Process Queue window, after a Process has successfully completed, or by selecting the $ViewResults$ option from the Results menu item in the menu bar on the main interface. When on e of these two options are selected the user will be immediately prompted to select a file. There will also be a drop down menu with three file filter choices: *RAId_Peptide_Id*, *Protein_Id*, and *Protein_Peptide_Id*. Different filter choices will change the data that is displayed.

All Data tables contain two buttons *Export as TSV* and *Export as CSV* that will export the contents of the table to a user specified file of the corresponding file type

1. Viewing Peptide ID data

Protein ID: all proteins that could have contained the peptide sequence.

Peptide: The sequence of the peptide identified during RAId's execution.

Molecular Mass (Da): The Molecular mass of the peptide.

E-value: The RAId assigned E-value of the peptide.

P-value: The RAId assigned P-value of the peptide

PFD: The calculated PFD at a given E-value cut off (E_c) using the Soric formula.

$$PFD(E_c) = \frac{E_c \times N}{\pi(E < E_c)},$$

Where N is the total number of MS/MS spectra and $\pi(E < E_c)$ is the total number of peptides with $E < E_c$.

Beginning Position/Ending position: position of the peptide on its corresponding protein.

2. Viewing Protein ID data

This data table contains the data on specific proteins identified during RAId's execution:

Protein ID: A protein's identification code.
(Note: these ID's are hyper linked to take you to their corresponding NCBI web pages)

ln(E-value): The natural log of the RAId assigned E-value of that protein.

ln(P-value): The natural log of the RAId assigned P-value of that protein.

PFD: The calculated PFD of that protein

$$PFD(E_c) = \frac{E_c \times \Pi}{\Pi(E < E_c)},$$

Where Π is the total number of proteins and $\Pi(E < E_c)$ is the total number of proteins with $E < E_c$.

Number of peptides: Number of peptides corresponding to that proteins sequence

Cluster number: Number representing which cluster a protein belongs to

3. Viewing Protein Peptide ID data

This data table contains information on each protein that was found in this assay and all peptides that may be found in that protein's sequence. The rows are colored based on the E-Value contained in each row. All protein sequences were collected and referenced from the database. In this table there are three separate interfaces, the *protein list* (top), the *peptide list* (middle), and the *protein sequence display*.

By selecting a protein from the protein list, the peptide list is filled with all identified peptides that correspond to that sequence, and the protein sequence display is filled with that proteins sequence. By selecting a peptide from the peptide list, that peptide sequence will be highlighted in the protein sequence viewer. Multiple peptides can be selected at a time by pressing Ctrl-click on each individual peptide, or Shift-click to select all peptides between the clicked peptide an the last peptide that was clicked. This table also allows the user to copy a value from any cell by left clicking the cell. The user can then paste it into another document or into other features of the RAId GUI.

Protein table:

Rank: Protein rank base on assigned E-value.

ProteinID: Protein identification number according to the database used.

ln(E-value): Natural log of the RAId assigned E-value (E).

ln(P-value): Natural log of the RAId assigned P-value (P).

PFD: The calculated PFD at a given E-value cut off (E_c).

Number of Peptides: The number of RAId identified peptides corresponding to the protein.

Peptide table:

RN: Record number.

Peptide: The sequence of the peptide.

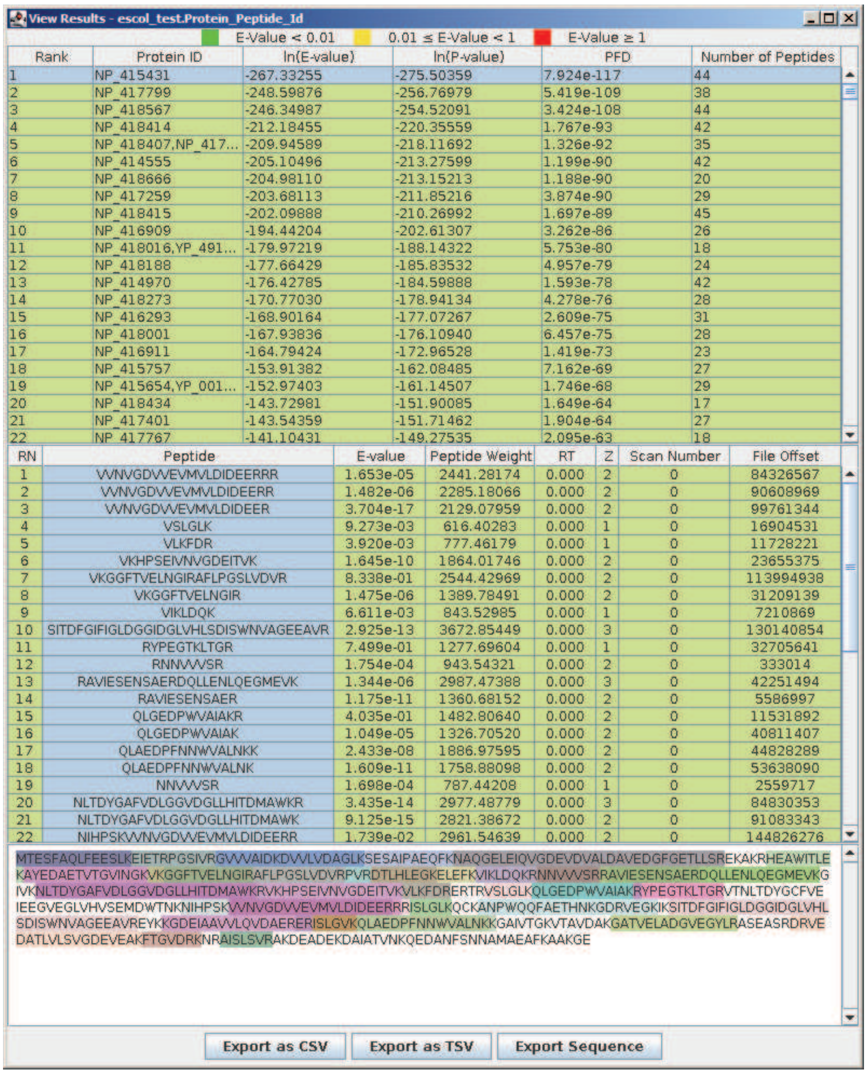
E-value: The RAId assigned E-value.

RT: Retention time.

Z: Charge.

Scan Number: The scan number of the MS/MS spectra.

File Offset: The location in the file for the MS/MS spectrum.



🟩 E-Value < 0.01 🟨 0.01 ≤ E-Value < 1 🟥 E-Value ≥ 1

Rank	Protein ID	ln(E-value)	ln(P-value)	PFD	Number of Peptides
1	NP_415431	-267.33255	-275.50359	7.924e-117	44
2	NP_417799	-248.59876	-256.76979	5.419e-109	38
3	NP_418567	-246.34987	-254.52091	3.424e-108	44
4	NP_418414	-212.18455	-220.35559	1.767e-93	42
5	NP_418407,NP_417...	-209.94589	-218.11692	1.326e-92	35
6	NP_414555	-205.10496	-213.27599	1.199e-90	42
7	NP_418666	-204.98110	-213.15213	1.188e-90	20
8	NP_417259	-203.68113	-211.85216	3.874e-90	29
9	NP_418415	-202.09888	-210.26992	1.697e-89	45
10	NP_416909	-194.44204	-202.61307	3.262e-86	26
11	NP_418016,YP_491...	-179.97219	-188.14322	5.753e-80	18
12	NP_418188	-177.66429	-185.83532	4.957e-79	24
13	NP_414970	-176.42785	-184.59888	1.593e-78	42
14	NP_418273	-170.77030	-178.94134	4.278e-76	28
15	NP_416293	-168.90164	-177.07267	2.609e-75	31
16	NP_418001	-167.93836	-176.10940	6.457e-75	28
17	NP_416911	-164.79424	-172.96528	1.419e-73	23
18	NP_415757	-153.91382	-162.08485	7.162e-69	27
19	NP_415654,YP_001...	-152.97403	-161.14507	1.746e-68	29
20	NP_418434	-143.72981	-151.90085	1.649e-64	17
21	NP_417401	-143.54359	-151.71462	1.904e-64	27
22	NP_417767	-141.10431	-149.27535	2.095e-63	18

RN	Peptide	E-value	Peptide Weight	RT	Z	Scan Number	File Offset
1	VVNVGDVVEVMVLDIDEERRR	1.653e-05	2441.28174	0.000	2	0	84326567
2	VVNVGDVVEVMVLDIDEERR	1.482e-06	2285.18066	0.000	2	0	90608969
3	VVNVGDVVEVMVLDIDEER	3.704e-17	2129.07959	0.000	2	0	99761344
4	VSLGLK	9.273e-03	616.40283	0.000	1	0	16904531
5	VLKFDR	3.920e-03	777.46179	0.000	1	0	11728221
6	VKHPSEIVNVGDEITVK	1.645e-10	1864.01746	0.000	2	0	23655375
7	VKGGFTVELNGIRAFLPGSLVDVR	8.338e-01	2544.42969	0.000	2	0	113994938
8	VKGGFTVELNGIR	1.475e-06	1389.78491	0.000	2	0	31209139
9	VIKLDQK	6.611e-03	843.52985	0.000	1	0	7210869
10	SITDFGIFIGLDGGIDGLVHLSDISWNVAGEEAVR	2.925e-13	3672.85449	0.000	3	0	130140854
11	RYPEGTKLTGR	7.499e-01	1277.69604	0.000	1	0	32705641
12	RNNVVVSR	1.754e-04	943.54321	0.000	2	0	333014
13	RAVIESENSAERDQLLENLQEGMEVK	1.344e-06	2987.47388	0.000	3	0	42251494
14	RAVIESENSAER	1.175e-11	1360.68152	0.000	2	0	5586997
15	QLGEDPWVAIAKR	4.035e-01	1482.80640	0.000	2	0	11531892
16	QLGEDPWVAIAK	1.049e-05	1326.70520	0.000	2	0	40811407
17	QLAEDPFNNWVALNKK	2.433e-08	1886.97595	0.000	2	0	44828289
18	QLAEDPFNNWVALNK	1.609e-11	1758.88098	0.000	2	0	53638090
19	NNVVVSR	1.698e-04	787.44208	0.000	1	0	2559717
20	NLTDYGAFVDLGGVDGLLHITDMAWKR	3.435e-14	2977.48779	0.000	3	0	84830353
21	NLTDYGAFVDLGGVDGLLHITDMAWK	9.125e-15	2821.38672	0.000	2	0	91083343
22	NIHPSKVVNVGDVVEVMVLDIDEERR	1.739e-02	2961.54639	0.000	2	0	144826276

```
MTESFAQLFEESLKEIETRPGSIVRGVVVAIDKDVVLVDAGLKSESAIPAEQFKNAQGELEIQVGDEVDVALDAVEDGFGETLLSREKAKRHEAWITLE
KAYEDAETVTGVINGKVKGGFTVELNGIRAFLPGSLVDVRPVRDTLHLEGKELEFKVIKLDQKRNNVVVSRRAVIESENSAERDQLLENLQEGMEVKG
IVKNLTDYGAFVDLGGVDGLLHITDMAWKRVKHPSEIVNVGDEITVKVLKFDRERTRVSLGLKQLGEDPWVAIAKRYPEGTKLTGRVTNLTDYGCFVE
IEEGVEGLVHVSEMDWTNKNIHPSKVVNVGDVVEVMVLDIDEERRRISLGLKQCKANPWQQFAETHNKGDRVEGKIKSITDFGIFIGLDGGIDGLVHL
SDISWNVAGEEAVREYKKGDEIAAVVLQVDAERERISLGVKQLAEDPFNNWVALNKKGAIVTGKVTAVDAKGATVELADGVEGYLRASEASRDRVE
DATLVLSVGDEVEAKFTGVDRKNRAISLSVRAKDEADEKDAIATVNKQEDANFSNNAMAEAFKAAKGE
```

Figure 11. Picture of the Protein and peptide data tables and the protein sequence display

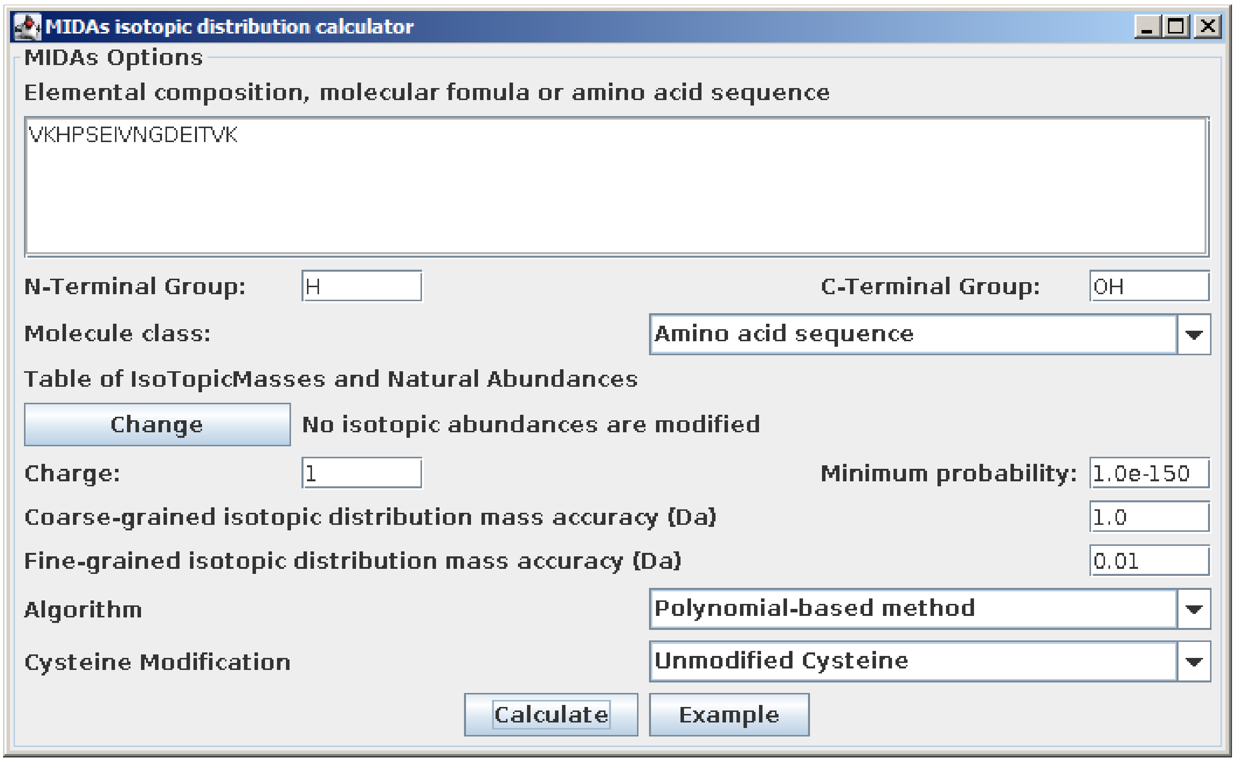
Unique Buttons:

Export Sequence: Exports the highlighted sequence as an image to a PDF or PNG file.

E. Using the MIDAs Isotopic distribution calculator

The MIDAs Isotopic Distribution Calculator can be accessed by clicking the *Tools* options on the menu bar of the main panel and selecting *MIDAs (Isotopic Distribution)*. This tool calculates the coarse and fine grained isotopic distributions. Isotopic masses and frequencies may be modified by the user.

1. Calculating an isotopic distribution



MIDAs isotopic distribution calculator

MIDAs Options

Elemental composition, molecular fomula or amino acid sequence

VKHPSEIVNGDEITVK

N-Terminal Group: H

C-Terminal Group: OH

Molecule class: Amino acid sequence

Table of IsoTopicMasses and Natural Abundances

Change

No isotopic abundances are modified

Charge: 1

Minimum probability: 1.0e-150

Coarse-grained isotopic distribution mass accuracy (Da) 1.0

Fine-grained isotopic distribution mass accuracy (Da) 0.01

Algorithm: Polynomial-based method

Cysteine Modification: Unmodified Cysteine

Figure 12. Picture of the MIDAs Window, and the isotopic modification window

MIDAs Options:

Input the elemental composition, molecular formula, or amino acid sequence in the designated field. The two ways the user can input into the field is either by typing the entry or using the pasting an entry by right clicking the text area. This will paste the most recently copied text, as well as a copied cell value from any of the peptide tables from the *View Results* tab.

Molecule Class: Designate the molecular class of the input molecule (Chemical, Amino acid sequence, DNA, or RNA).

Charge: Provide the charge of the input chemical.

MinimumProbability: Provide the minimum probability to accept a calculated molar mass.

Coarse − Grained Mass Accuracy: Provide the Coarse-grained mass accuracy in daltons.

Fine − Grained Mass Accuracy: Provide the Fine-grained mass accuracy in daltons.

Algorithm: Select an algorithm for performing calculations.
(Polynomial-based, FFT-based (mass domain), and FFT-based (nucleon domain) methods).

Example: Fill each field with some example data sets.

Calculate: Generate an isotopic distribution.

Changing isotopic masses and frequencies:

Click the change button to open up the isotope window. Within this window, any of the values contained in text boxes may be modified.

Reset: Changes all modified fields back to the default, and clears the memory of committed changes.

Accept: Commits all modified values to memory.

Close: Closes the window.

2. Displaying results of the Isotopic Distributions

After pressing calculate, the *IsotopicDistributionResults* window will pop up. This window has three tabs *Statistics*, *CoarseGrainedDistribution*, and *FineGrainedDistribution*.

The Statistics tab:

This tab contains information about the accuracy of the MIDAs calculations.

Export: Click to save the statistics, and data for the coarse and fine grained distribution to a user specified TSV or CSV file.

The Coarse Grained Distribution tab:

This tab displays the results of the Coarse grained distribution as a curve.

The exact mass of a point on the graph is shown when the cursor hovers over it.

Clicking export on this tab will save the displayed image to a user specified PDF or PNG file.

The Fine Grained Distribution tab:

This tab displays the results of the Fine grained distribution as a thin line bar graph.

The user can zoom in on specific ranges clicking and dragging a box across the specified range.

The exact mass of a point on the graph is shown when the cursor hovers over it.

Clicking export on this tab will save the displayed image to a user specified PDF or PNG file.

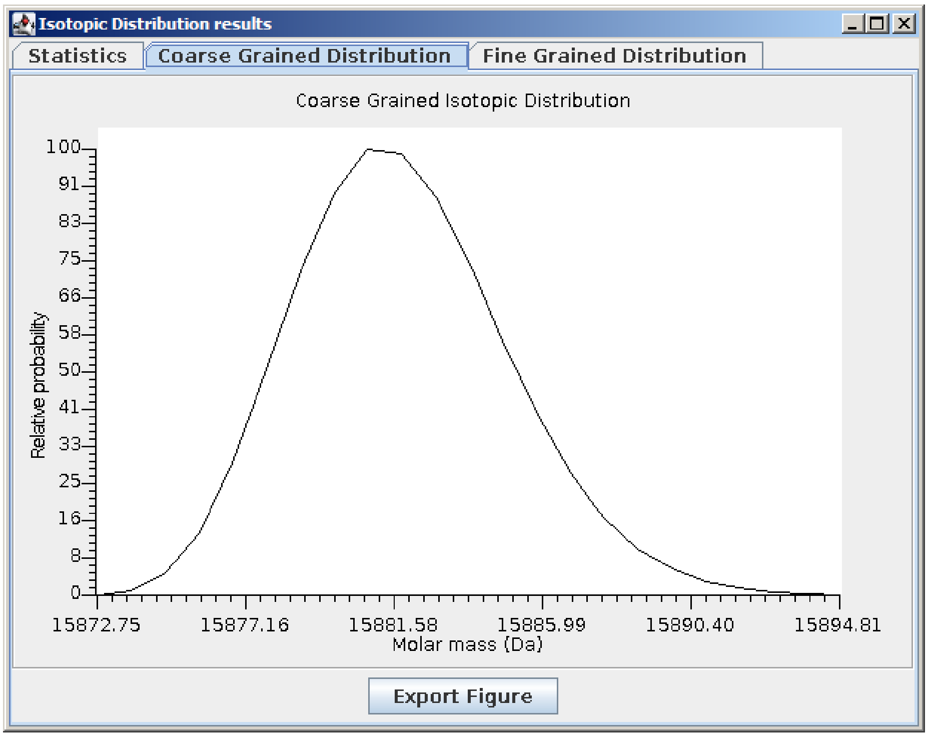


Figure 13. Example of a Coarse Grained Isotopic Distribution

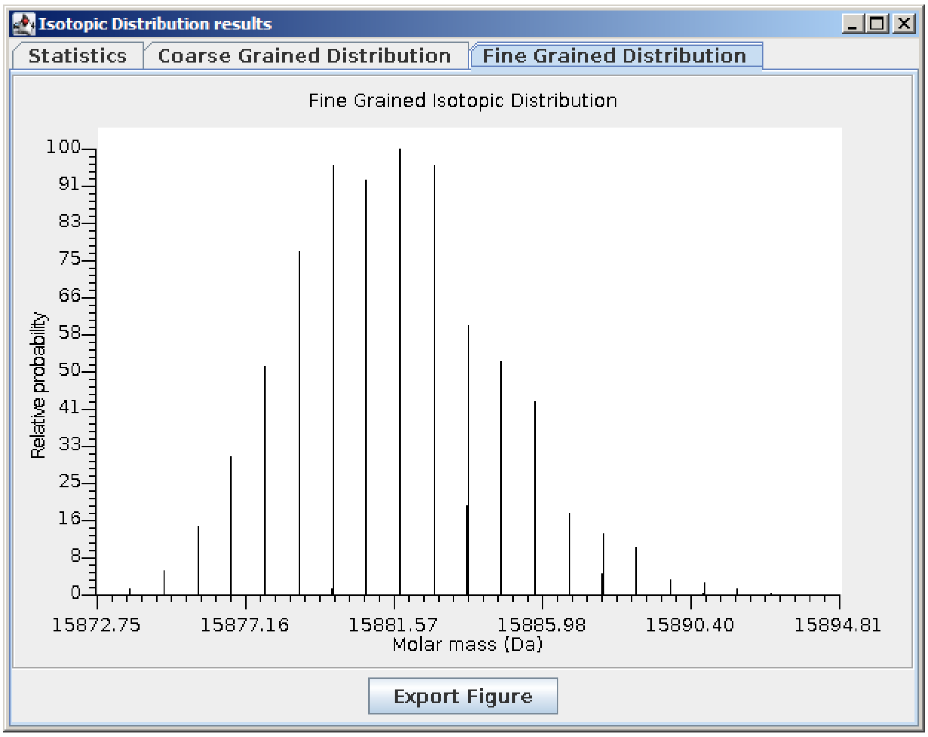


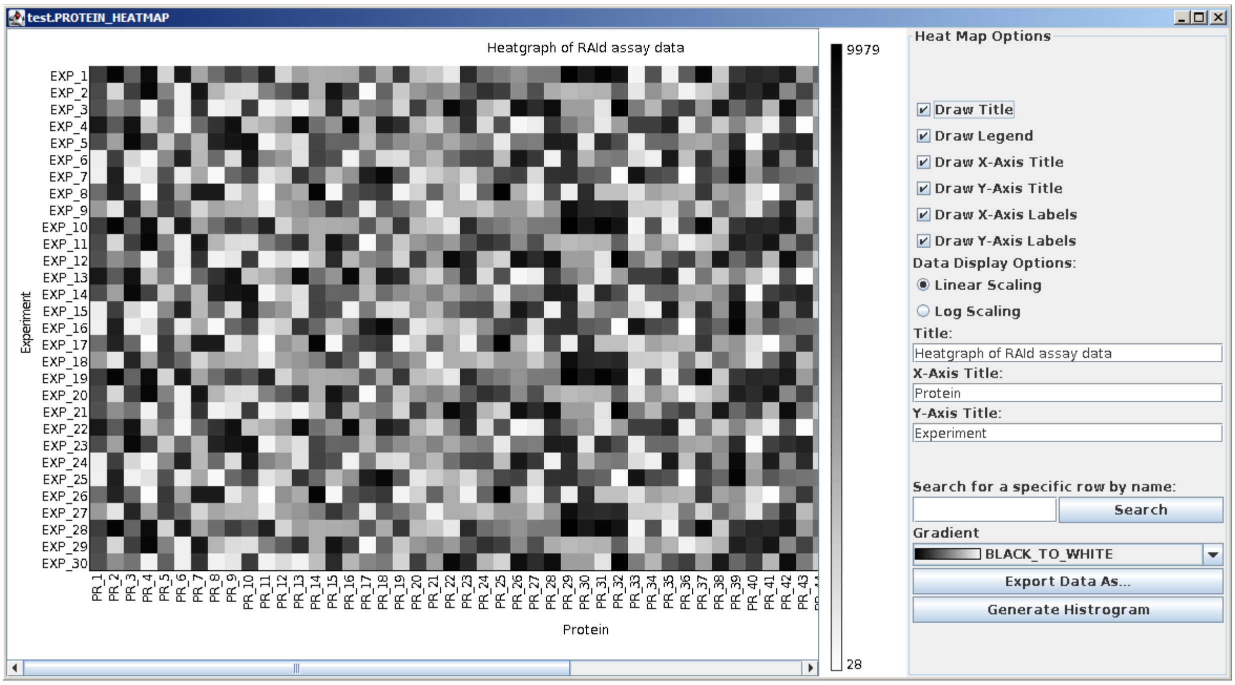
Figure 14. Picture of an example of a Fine Grained Isotopic Distribution

F. Heatmap Display of Experimental Data

The Heatmap is opened by going to the *Plot Results* tab of the main panel's menu bar and selecting the *Plot Heatmap* option. Once selected a file chooser will appear then select a file.

Valid file extensions: *.Protein_Heatmap*

1. Heatmap options


test.PROTEIN_HEATMAP

Heat Map Options

☑ Draw Title

☑ Draw Legend

☑ Draw X-Axis Title

☑ Draw Y-Axis Title

☑ Draw X-Axis Labels

☑ Draw Y-Axis Labels

Data Display Options:

● Linear Scaling

○ Log Scaling

Title: Heatgraph of RAld assay data

X-Axis Title: Protein

Y-Axis Title: Experiment

Search for a specific row by name:

Gradient: BLACK_TO_WHITE

Heatgraph of RAld assay data
Experiment
EXP_1
EXP_2
EXP_3
EXP_4
EXP_5
EXP_6
EXP_7
EXP_8
EXP_9
EXP_10
EXP_11
EXP_12
EXP_13
EXP_14
EXP_15
EXP_16
EXP_17
EXP_18
EXP_19
EXP_20
EXP_21
EXP_22
EXP_23
EXP_24
EXP_25
EXP_26
EXP_27
EXP_28
EXP_29
EXP_30
PR_1 PR_2 PR_3 PR_4 PR_5 PR_6 PR_7 PR_8 PR_9 PR_10 PR_11 PR_12 PR_13 PR_14 PR_15 PR_16 PR_17 PR_18 PR_19 PR_20 PR_21 PR_22 PR_23 PR_24 PR_25 PR_26 PR_27 PR_28 PR_29 PR_30 PR_31 PR_32 PR_33 PR_34 PR_35 PR_36 PR_37 PR_38 PR_39 PR_40 PR_41 PR_42 PR_43
Protein
9979
28
Search
Export Data As...
Generate Histrogram

Figure 15. Picture of a heatmap distribution

Draw Title: If checked then the title will be displayed.

Draw Legend: If checked then the legend will be displayed.

DrawX − Axis title: If checked then the X-axis title will be displayed.

DrawY − Axis title: If checked then the Y-axis title will be displayed.

DrawX − Axis Labels: If checked then the experiment names will be displayed along the X-axis.

DrawY − Axis Labels: If checked then the Protein names will be displayed along the Y-axis.

The scaling of the heatmap's cells can be changed by selecting *Linearscaling* or *Logscaling*. The heatmap is initialized with a linear scale. In the event that the user tries to log scale data with values less than or equal to zero a prompt will appear addressing the issue and the option will be deactivated.

Title: Alter the title of the graph.

X − Axis title: Alter the X-Axis title.

Y − Axis title: Alter the Y-Axis title.

To search for a specific column enter a column name in the corresponding text box and click *search*.

The gradient for cell intensity can be changed using the drop down menu.

ExportDataAs...: Click to save the data as a PDF or PNG.

GenerateHistogram: Click and then specify a row or column name to generate a histogram of the data from that data set

2. Histogram options

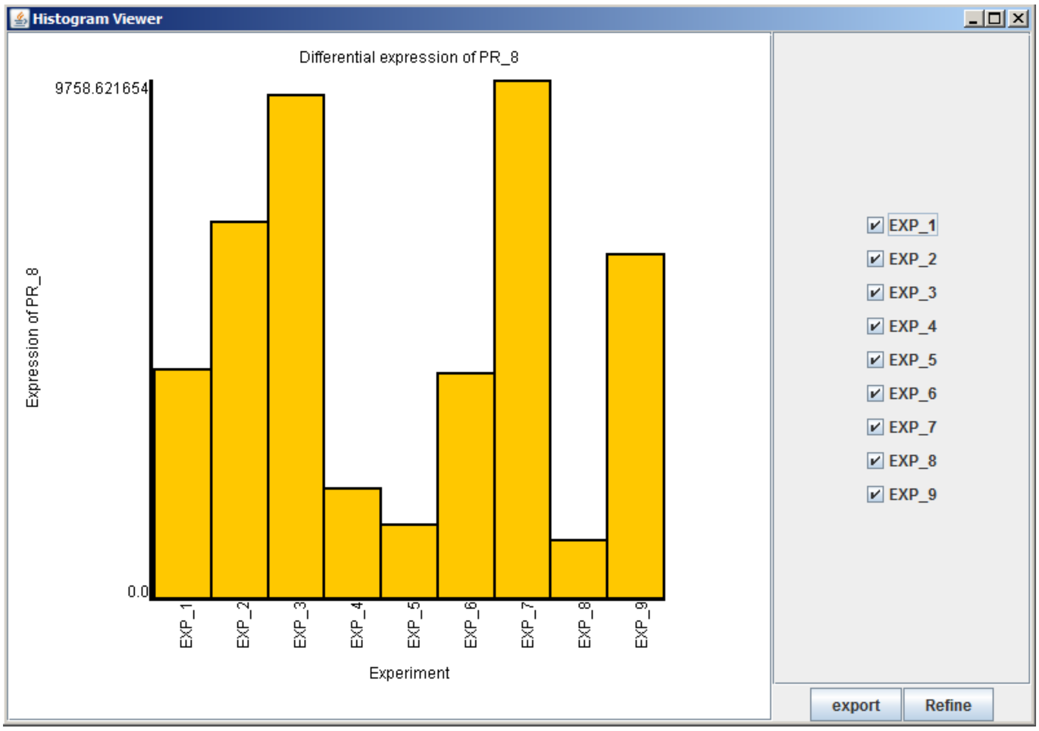


Figure 16. Picture of a heatmap distribution

The X-axis contains the names of the categories while the Y-Axis represents the range of the data. Minimum automatically set to zero if all intensities are greater than 0.

The data displayed can be changed by selecting the desired data from the list on the right.

Refine: Click to confirm changes to the displayed data made.

Export: Click to export the current image as either a PNG or PDF file.

G. Protein and Peptide PFD Plot Viewer

The Protein and Peptide PFD plot viewers can be accessed by going to the *Plot Results* section of the main panel's menu bar and selecting the *Plot PFD* option. Once selected a file chooser will appear, select any of the three options *Protein PFD plot*, *Protein Cluster PFD plot*, *Peptide PFD plot*, and then select a file with the corresponding file extension. Instead of using only utilities offered by the Java Swing API, this GUI uses the JFreeChart API to generate and render the graph.

Valid file extension for *Protein PFD plot* and *Protein Cluster PFD plot*: *.Protein_Id*

Valid file extension for *Peptide PFD plot*: *.RAId_Peptide_Id*

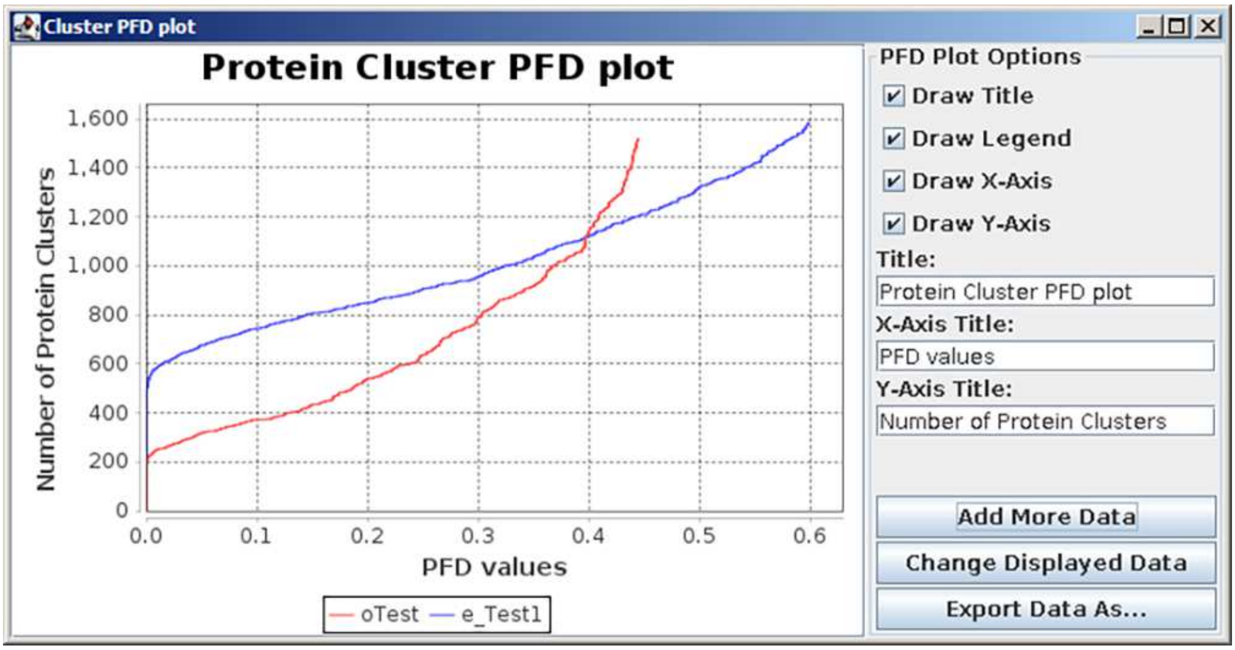
Proteins and peptide are ranked 1 to n (where n is the number of proteins) by their PFD values, in ascending order.

The protein PFD plots plot PFD as a function of protein rank.

The protein cluster PFD plots plot PFD of the lowest ranking protein in a cluster as a function of cluster rank.

The peptide PFD plots plot PFD as a function of peptide rank.

1. PFD plot options:



PFD Plot Options

☑ Draw Title

☑ Draw Legend

☑ Draw X-Axis

☑ Draw Y-Axis

Title: Protein Cluster PFD plot

X-Axis Title: PFD values

Y-Axis Title: Number of Protein Clusters

Figure 17. Picture of the PFD plot interface

Draw Title: If checked then the title will be displayed.

Draw Legend: If checked then the legend will be displayed.

Draw X − Axis: If checked then the X-axis title and the values written beneath it will be displayed.

Draw Y − Axis: If checked then the Y-axis title and the values written alongside it will be displayed.

Title: Alter the title of the graph.

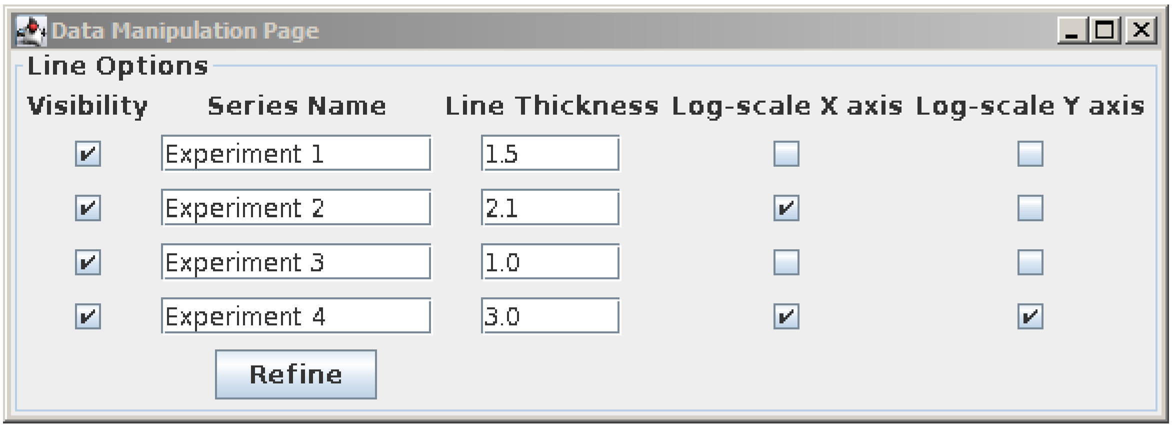
X − Axis title: Alter the X-Axis title.

Y − Axis title: Alter the Y-Axis title.

Click and drag from left to right the graph to zoom in on a specific range and domain.

Add More Data: Clicking this button will prompt the user for another file, the data from the new file will be added to the graph, provided the file does not share a name with any of the series existing on the graph.

Change Displayed Data: Clicking this button will open the data modification window. In this window the user will be able to set the visibility, name, and line thickness, as well as log scale the X and/or Y axis.



Line Options

Visibility	Series Name	Line Thickness	Log-scale X axis	Log-scale Y axis
☑	Experiment 1	1.5	□	□
☑	Experiment 2	2.1	☑	□
☑	Experiment 3	1.0	□	□
☑	Experiment 4	3.0	☑	☑

Figure 18. Picture of the PFD plot interface

Export Data As...: Click this button to save the current PFD plot as a PDF or PNG file.

2. Right Click Menu Options available from JFreeChart

Copy: Copies the plot.

Save As: Saves the current plot as a .png file.

Print: Opens up standard printing interface for the current plot.

Zoom In / Zoom out: User can zoom in or out just the domain or range axis, or both axis.

Auto Range: Re-scales the domain, range, or both axis to fit all series.

Properties: Opens the JFreeChart Properties window.

Title Tab

Show Title: Sets the visibility of the title.

Text: Allows for altercation of the title.

Font: Configure the current font style of the title.

Color: Change the color of the title.

Plot Tab - Domain and Range Axis

Label: Set the title of the domain and range axis.

Font: Configures font style of the axes.

Paint: Set the color of the axes.

Show Tick Labels: Toggles display of the incremental tick values.

Tick Label Font: Configures the font style size and attributes.

Show Tick Marks: By checking this option the tick marks are visible on the plot.

Auto-Adjust Range: When this operation is toggled on, adjustment will occur to fit new data points.

Maximum Range Value: Allows user to custom select a maximum range value.

Minimum Range Value: Allows user to custom select a minimum range value.

Plot Tab - Appearance

Outline Stroke: Configures the borderline style to the user's choice.

Outline Paint: Sets the color of the outline stroke.

Background Paint: Sets the background color to the user's choice.

Orientation: Adjusts plot orientation. (Vertical puts Domain on X-axis, horizontal puts Domain on Y-axis).

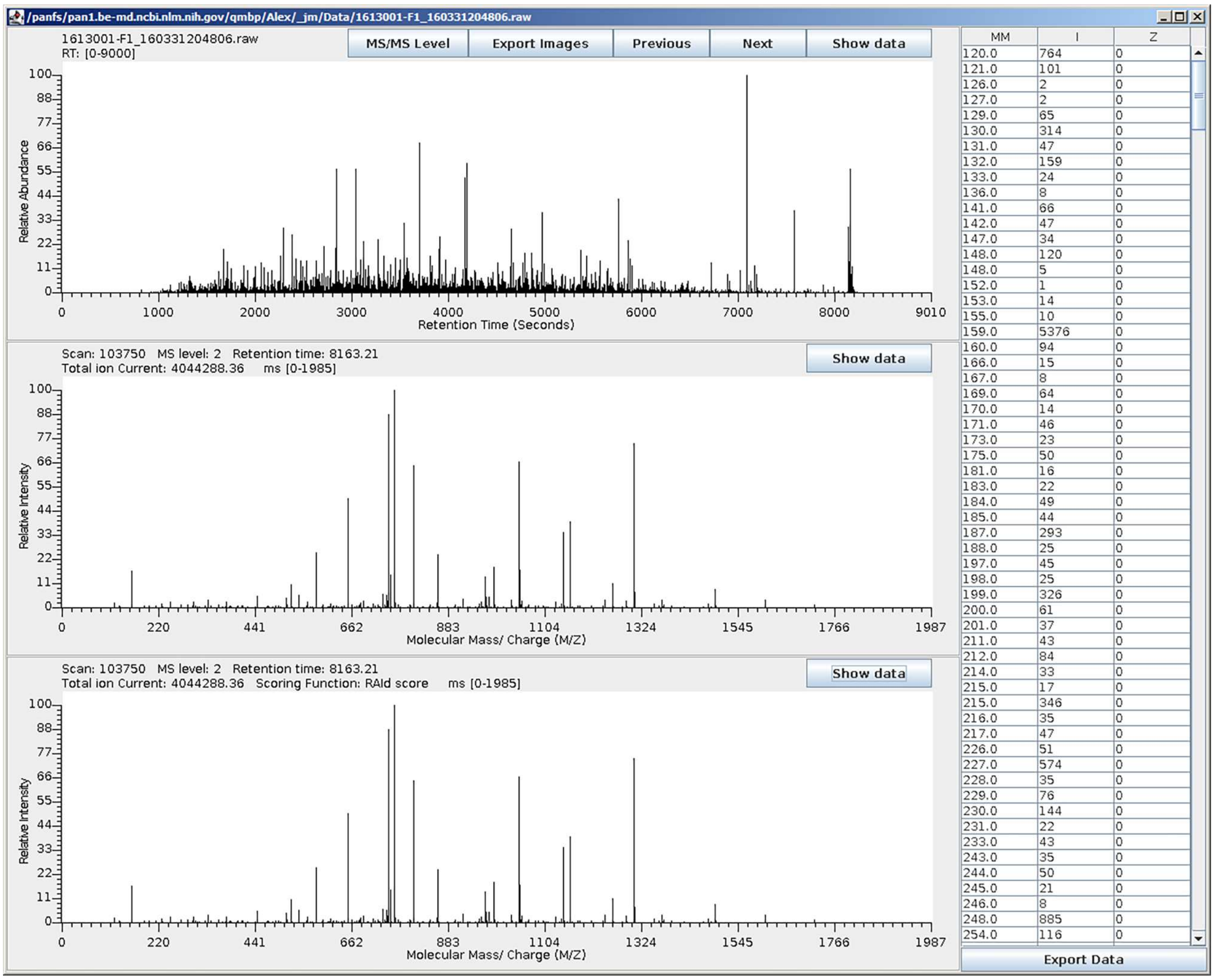
Other Tab

Draw Anti-Aliased: By checking this option the lines on the plot are more visible.

Background Paint: Adjusts the color of the window's background.

H. Mass Spectrometry data display

The mass spectrometry data viewer allows the user to view the results of chromatography and mass spectrometry in three separate graphs. This mass spectrometry viewer provides tools to analyze both chromatography data and individual spectrum data. To open this display click the click the *Tools* option from the menu bar on the main panel and select *View MS Spectrum*.



1613001-F1_160331204806.raw
RT: [0-9000]

Scan: 103750 MS level: 2 Retention time: 8163.21
Total ion Current: 4044288.36 ms [0-1985]

Scan: 103750 MS level: 2 Retention time: 8163.21
Total ion Current: 4044288.36 Scoring Function: RAId score ms [0-1985]

MM	I	Z
120.0	764	0
121.0	101	0
126.0	2	0
127.0	2	0
129.0	65	0
130.0	314	0
131.0	47	0
132.0	159	0
133.0	24	0
136.0	8	0
141.0	66	0
142.0	47	0
147.0	34	0
148.0	120	0
148.0	5	0
152.0	1	0
153.0	14	0
155.0	10	0
159.0	5376	0
160.0	94	0
166.0	15	0
167.0	8	0
169.0	64	0
170.0	14	0
171.0	46	0
173.0	23	0
175.0	50	0
181.0	16	0
183.0	22	0
184.0	49	0
185.0	44	0
187.0	293	0
188.0	25	0
197.0	45	0
198.0	25	0
199.0	326	0
200.0	61	0
201.0	37	0
211.0	43	0
212.0	84	0
214.0	33	0
215.0	17	0
215.0	346	0
216.0	35	0
217.0	47	0
226.0	51	0
227.0	574	0
228.0	35	0
229.0	76	0
230.0	144	0
231.0	22	0
233.0	43	0
243.0	35	0
244.0	50	0
245.0	21	0
246.0	8	0
248.0	885	0
254.0	116	0

Figure 19. Picture of Mass Spectrometry data display with a spectrum and data table displayed

1. Opening a Mass Spectrometry data file

Once View spectrum has been selected a window will pop up prompting the user for a MS file (*.dta*, *.pkl*, *.mgf*, *mzData*, *.mzML*, *.mzXML*, *.raw*
), C and N terminal molecular masses, product and precursor ion mass tolerance and the scoring function.

Once all specifications have been finalized press *DisplayData*.

2. Interpreting the displayed results

When the display first starts up there will only be one graph. This graph contains the results of the protein chromatography with retention time along the X-axis and the normalized total ion current along the Y-axis. Once the other two graphs are opened they will contain mass per charge along the X-axis and the normalized intensity along the Y-axis.

Each graph contains a *Show data* button, which causes the data of that graph to be displayed on the right side of the three graphs. (changing any graph with it's data displayed will change the displayed data table as well)

The user can zoom in on specific ranges clicking and dragging a box across the specified range.

Unique functionality of the top most graph:

The exact retention time of a point on the graph is shown when the cursor hovers over it.

The *MSLevelX* (where X is an integer greater than zero) button will advance the MSLevel on the button by one (going back to one once the max is reached), remove the spectrum graphs and data table if they are active, and change the data displayed on the topmost graph to consist of only data of the new MS level.

Select a spectrum by clicking on any individual peak, this will open, or replace if already opened, two new graphs directly below the original depicting the mass spectrums taken at that instant in time.

The *Next* and *Previous* buttons are only active when a spectrum is selected. These buttons will show the next and previous spectrums of the same MS level, respectively.

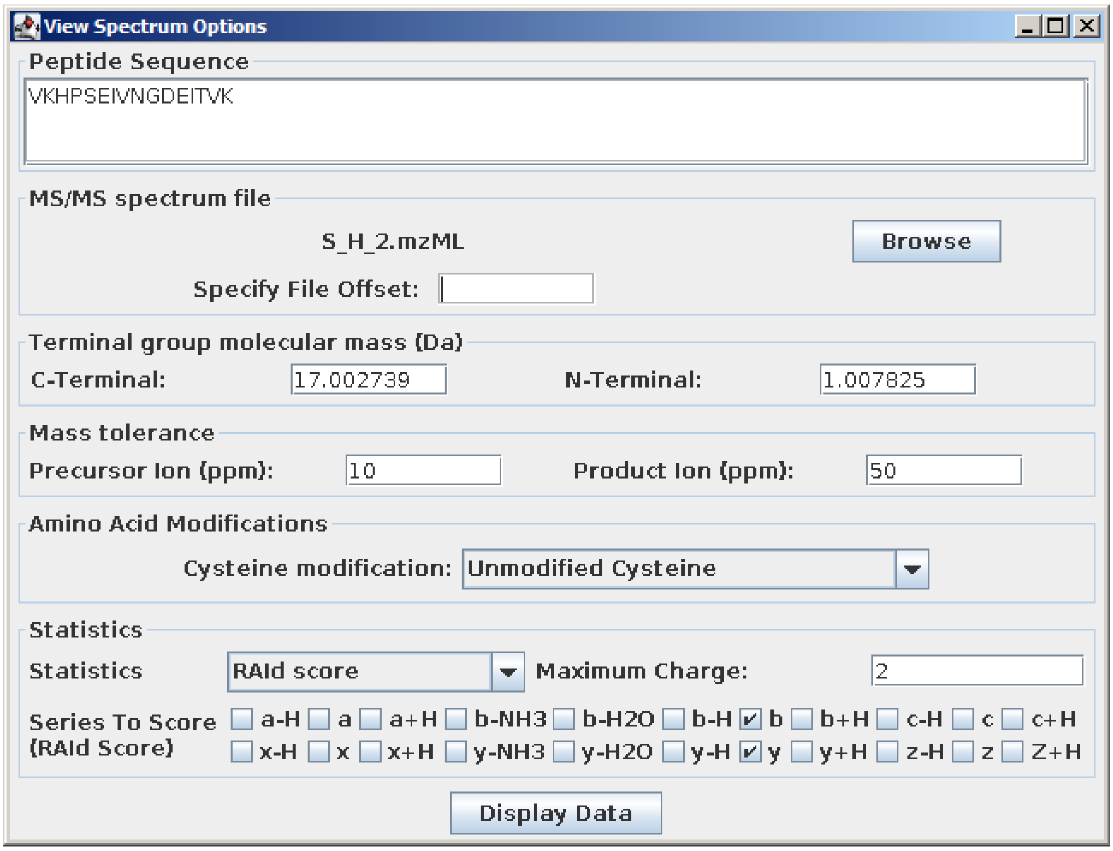
The *Export* button asks the user if they want to save all of the active graphs as a PDF or PNG then ask if they want to save the data table as a TSV or CSV

Unique functionality of the middle and bottom graphs:

The exact mass per charge and charge of a point on the graph is shown when the cursor hovers over it.

I. Peptide Fragmentation Tool

The peptide fragmentation tool generates a table displaying the ideal mass of peptide fragment based on a user inputted peptide sequence. The user may also supply an MS/MS spectrum file and the file offset for a specific spectrum to display that spectrum with all the matching peaks in red. To open this display click the *Tools* option from the menu bar on the main panel and select *Peptide Product-ions*.



Peptide Sequence
VKHPSEIVNGDEITVK

MS/MS spectrum file
S_H_2.mzML
Specify File Offset:

Terminal group molecular mass (Da)
C-Terminal: 17.002739
N-Terminal: 1.007825

Mass tolerance
Precursor Ion (ppm): 10
Product Ion (ppm): 50

Amino Acid Modifications
Cysteine modification: Unmodified Cysteine

Statistics
Statistics: RAId score
Maximum Charge: 2

Series To Score (RAId Score)
□ a-H □ a □ a+H □ b-NH3 □ b-H2O □ b-H ☑ b □ b+H □ c-H □ c □ c+H
□ x-H □ x □ x+H □ y-NH3 □ y-H2O □ y-H ☑ y □ y+H □ z-H □ z □ Z+H

Figure 20. Picture of the peptide specification window where user defines parameters for a peptide fragmentation

1. Preparing a Peptide Sequence for Fragmentation

Peptide Sequence

Enter a peptide sequence by either typing or pasting with the right click button on the mouse to be fragmented in the text field. Post-Translation Modifications (PTM's) can be specified by specifying the PTM id number to the right of the modified amino acid.

MS/MS Spectrum file

Browse: Opens a file chooser allowing the user to search for and upload their MS/MS spectrum file.

Valid file extensions: *.raw*, *.mzml*, *.mzXML*, and *.mgf*

Specify File Offset: Provide the offset of the mass spectrum.

Terminal Group Molecular Mass (Da)

C-Terminal: Specify the mass of the Carboxy-terminal group.

N-Terminal: Specify the mass of the Amino-terminal group.

Mass Tolerance

Precurssor Ion (ppm): Specify the Precursor Ion's mass tolerance.

Product Ion (ppm): Specify the Product Ion's mass tolernace.

Amino Acid Modifications

Cysteine Modification: Sets the default state of cysteine during the peptide fragmentation.

Statistics

Statitics: Select a scoring function to analyze the spectrum.

Available Scoring functions: RAId score, K-score, HyperScore, and XCorr.

Maximum Charge: Specify the maximum fragmentation of the peptide sequence.

Series To Score: Select all fragments to be generated during the peptide's fragmentation.

2. Viewing a Peptide's Fragmentation

The fragmentation Data Table

Regardless of whether or not a spectrum is specified by the user, a peptide fragmentation data table will be displayed. If no spectrum data has been supplied then peptide fragment masses will be calculated[8] and the table will display all possible A,B,C,X,Y, and Z fragments of the peptide sequence chosen by the user.

If a spectrum has been provided then the table will display all ideal fragments, the calculated mass of those fragments, and if there is an experimental peak that could be matched to the ideal peak, the mass of the experimental peak and it's intensity. Matched Ideal fragments will also appear in red text.

Export Data Table: Exports the current data table to a user specified CSV or TSV file.

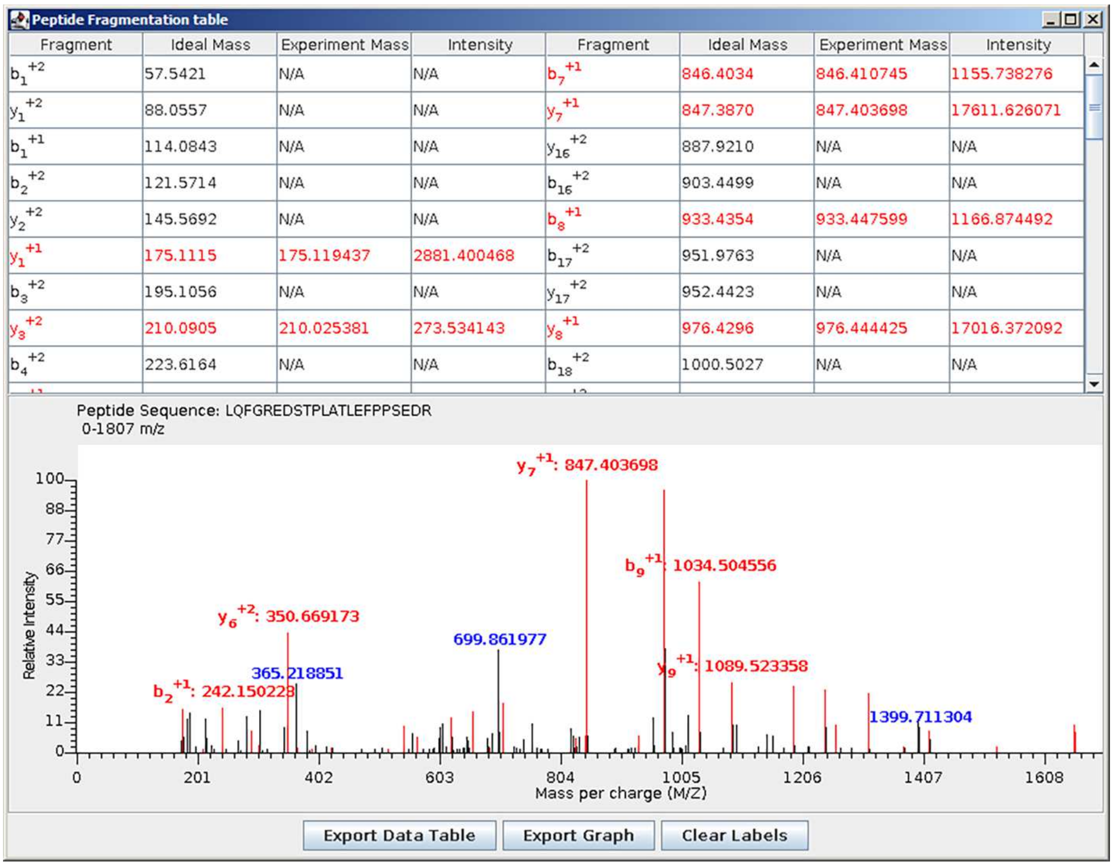
Scoring Experimental Peaks in a User Provided Spectrum

The following scoring function is used to determine which peak's most likely represent the data provided by the user. Any experimental peaks with a score greater than 0 are considered potential matches to a theoretical mass, and an experimental peak may have more than one matching fragment.

$$\varepsilon(T_{MW}) = T_{MW} * 10^{-6} * PI$$

$$S(E_{MW}) = I(E_{MW}) * e^{\frac{T_{MW}-E_{MW}}{\varepsilon(T_{MW})}} * \theta(1-(T_{MW}-E_{MW}))$$

Where T$_{MW}$ is the theoretical molecular weight of a fragment, E_{MW} is an experimentally identified molecular weight, PI is the Product ion mass tolerance (in ppm), and I(E_{MW}) is the intensity at an



Fragment	Ideal Mass	Experiment Mass	Intensity	Fragment	Ideal Mass	Experiment Mass	Intensity
b_1^{+2}	57.5421	N/A	N/A	b_7^{+1}	846.4034	846.410745	1155.738276
y_1^{+2}	88.0557	N/A	N/A	y_7^{+1}	847.3870	847.403698	17611.626071
b_1^{+1}	114.0843	N/A	N/A	y_{16}^{+2}	887.9210	N/A	N/A
b_2^{+2}	121.5714	N/A	N/A	b_{16}^{+2}	903.4499	N/A	N/A
y_2^{+2}	145.5692	N/A	N/A	b_8^{+1}	933.4354	933.447599	1166.874492
y_1^{+1}	175.1115	175.119437	2881.400468	b_{17}^{+2}	951.9763	N/A	N/A
b_3^{+2}	195.1056	N/A	N/A	y_{17}^{+2}	952.4423	N/A	N/A
y_3^{+2}	210.0905	210.025381	273.534143	y_8^{+1}	976.4296	976.444425	17016.372092
b_4^{+2}	223.6164	N/A	N/A	b_{18}^{+2}	1000.5027	N/A	N/A

Peptide Sequence: LQFGREDSTPLATLEFPPSEDR
0-1807 m/z

Figure 21. Picture of a sample peptide fragmentation. A few significant matching peaks and non matching peaks have had their labels fixed in place.

Experimentally determined molecular weight.

Spectrum Matches

When a user supplies spectrum data, that data will be displayed as a graph directly below the data table. All matching peaks will be highlighted red.

When the mouse hovers over a peak a label will be displayed right above it. If the peak was successfully matched to one or more ideal fragments then the label will be red and show all matched ideal fragments and the mass per charge of the peak; otherwise the label will be blue and show only the mass per charge.

If a user clicks on any peak then a label will be fixed to that position until the peak is clicked again, or the *Clear Labels* button is clicked.

The user can zoom in on specific ranges clicking and dragging a box across the specified range.

Export Graph: Exports the current displayed graph to a user specified PNG or PDF.

Clear Labels: Clears all labels saved on the graph.

References

Alves, G., A. Y. Ogurtsov, and Y. K. Yu, 2007, Biol. Direct **2**, 25.

Alves, G., A. Y. Ogurtsov, and Y. K. Yu, 2008, BMC Genomics **9**, 505.

Alves, G., and Y. K. Yu, 2008, Physica A **387**, 6538.

Alves, G., and Y. K. Yu, 2015, Bioinformatics **31**(5), 699.

Eng, J. K., A. L. McCormack, and J. R. Yates III, 1994, J. Amer. Soc. Mass Spectrom. **5**, 976.

Feny, D., and R. C. Beavis, 2003, Anal. Chem. **75**, 768.

Schandorff, S., J. V. Olsen, J. Bunkenborg, B. Blagoev, Y. Zhang, J. S. Andersen, and M. Mann, 2007, Nat. Methods **4**, 465.

Steen, H., and M. Mann, 2004, Nature reviews Molecular cell biology **5**(9), 699.