NIST Library of Peptide Ion Fragmentation Spectra
July 2008

Contents

Overview

This library contains MS/MS spectra of peptide ions generated by the tryptic digestion of proteins. These spectra, also called tandem mass spectra, contain the m/z values and abundances of ion products from the collision-induced dissociation (CID) of protonated peptide ions. The principal use of these spectra is to identify or validate peptides by matching newly acquired spectra to spectra in the library. Matching a full spectrum against a full reference spectrum is more discriminating that is matching against an idealized 'theoretical' spectrum derived from sequence information alone, as is current practice. Spectrum-to-spectrum matching makes use of abundance, neutral loss peaks and other features of mass spectra not employed in current sequence search methods, enabling faster, more reliable and more sensitive determination of peptide identities. However, while identified peptide sequences are commonly used to identify the proteins from which they were derived by digestion, the library and software provided do not address this 'protein inference' step – peptide identifications and their scores are the final results.

All spectra in this library originated from peptide ions generated by electrospray ionization in LC-MS/MS experiments. Most spectra were acquired with ion trap mass spectrometers, with a smaller number from low energy collision cell instruments (qtof-class). Sources of spectra are given in the file 'references.txt' [1], which is provided with each library. Protein sequences used for initial identifications of peptides using sequence search engines are given in the compressed file also included. Three types of spectra are possible: 1) 'consensus' spectra - derived from multiple (replicate) identifications of a peptide ion, 2) best 'replicate' spectra and 3) high confidence 'single' spectrum identifications. The specific organism covered in the most recent) release are given at http://peptide.nist.gov:

This collection is intended primarily to demonstrate the utility of peptide ion fragmentation libraries by enabling the development of applications that use them for identifying peptides, although in some cases (human and yeast) it extensive enough to be employed for practical use. Emphasis has been placed on quality control, since an error in a library can be widely propagated. For example, library spectra containing significant impurity peaks can falsely identify the impurity in an analysis as the library peptide. On the other hand, to maximize completeness and minimize false negative results (peptide ion not in library), extensive extraction methods were employed. These methods combined results of several sequence search engines for initial peptide identification and improved the separation of true from false positive results. Spectra were annotated in detail to aid spectrum library scoring as well as to document the origin of the spectrum. It is important to note that none of these collections are complete. Additional measurements are needed to increase both the coverage and the quality of the library.

Background

The MS/MS spectrum of a peptide ion depends on its sequence (including modifications), its charge and its energy. Hence, for a given peptide ion, the principal source of spectrum variability is its energy, which is itself determined by collisional conditions in the instrument. In an ion trap, where ions are energized in small steps, energies of dissociating ions are largely fixed by instrument design. For "low energy" collision cells in qtof and triple quadrupole instruments, where high velocity ions are slowed through multiple collisions, energies depend primarily on initial collision energy, which in turn is controlled by instrument settings. Since spectra from these two instrument classes have fundamental differences, their spectra are separately processed, tagged (Inst=it or Inst=qtof) and for the present held in different libraries. However, spectra from ion traps are often very similar to spectra acquired at low collisional energies (low extents of dissociation) in collision cell instruments, and to some degree differences can be modeled, so that with correction by algorithm ion trap and qtof spectra can be searched against each other. Such an algorithm is implemented in the search software provided. We find that the largest differences in spectra for the two methods occur when multiply charged ions are major products, since additional energy in collision cells can dissociate these to lower charge state ions.

Since charge migration in a peptide ion is thought to be rapid relative to CID fragmentation, the sequence and charge of a peptide can be assumed to fully determine its identity; hence the origin of the peptide ion and specific electrospray conditions, should not be a source of spectrum variability. However, experimental MS/MS spectra of a given peptide ion can vary for reasons other than energy content. Different m/z ranges for ion detection is one possible cause. Another is abundance variability due to 'ion counting' or 'shot' noise and related 'centroiding' problems, which vary roughly with the square root of the number of ions detected and is most significant for low intensity peaks (low S/N, signal-to-noise ratio). Such peaks can disappear unpredictably when they fall near the detection threshold in a spectrum. Another source of variability arises from fragmentation of impurity precursor ions that are within the instrument's precursor m/z tolerance. These produce unexplained peaks in a spectrum, a problem that worsens with sample complexity and decreasing signal of the peptide ion of interest. Non-peptide contaminants are especially common, often producing major peaks from neutral losses from the parent ion.

Other, generally less common sources of variability can be cited. Low or high instrument m/z biases or cutoffs can systematically affect abundances. Chemical reactions after ion selection, but before excitation, which may occur for highly reactive ions in ion traps, can influence spectra, though this does not appear to be significant for positive peptide ions.

In short, a mass spectrum may be viewed as an energy-dependent property of an ion. Like most measurements of such properties, it is influenced primarily by signal strength and purity and can be influenced by imperfect detection methods. The goal of this reference library of peptide MS/MS spectra is to make available as complete, fully annotated and artifact free set of these properties as possible.

Library Construction

This section outlines the methods now used for building the library. As they are refined, we expect details to change.

1) *Source Spectra*: All spectra were extracted from LC-MS/MS data files of tryptic digests, with peptide ions generated by electrospray ionization. Many experiments involved prior fractionation of digested peptides. Each experiment is assigned a 'sample tag', which provides a link between individual spectra and their source experiment. References to sample tags are given in a separate file (references.txt). The type of alkylation is given as the last segment of the sample tag, which include; none (_none), carboxyacetamide (_CAM), cleavable ICAT (_clICAT), and original, un-cleavable ICAT (_ucICAT) (ICAT = Isotope-Coded Affinity Tags).

2) *Initial Spectrum Identification*: Initial assignments of spectra to peptides were made using up to four different sequence search engines [4]. Reported scores for each of the search engines were normalized using results of searching against a combined forward (correct) and reversed (incorrect) sequence library. The best normalized score (or expectation value) among the search engines was used for further processing. In general, peptides were permitted to have to up to two missed cleavages and one non-tryptic terminus (semitryptic). Parent and fragment ion tolerances of 2 and 0.8 m/z, respectively, were generally used. Charge states of at least +1 to +3 were examined for all spectra. Variable oxidized methione and N-terminal acetylation and glutamine deamidation were used along with appropriate fixed cysteine alkylation. Deamidation of aspartic acid was also used to exclude them from contributing to unmodified peptides and are not included in the library.

3) *Consensus Spectra*: Multiple spectra assigned to a single peptide ion (replicates) were combined to form a 'consensus spectrum'. This involved the following series of steps for each identified peptide ion:

a) Identify spectra with the highest sequence search scores (maximum of 100 spectra).

b) Align m/z values in the different spectra (0.1 m/z bins are used).

c) Compute spectrum similarity (dot product [5]) for each spectrum pair.

d) Identify cluster of most similar, highest scoring spectra, reject the rest.

e) Find peaks (m/z) that were present in the majority of replicate spectra with sufficient signal-to-noise ratio to have been capable of generating the peak (spectra with signal/noise too low to have generated the peak are ignored).

f) Generate an abundance from weighted averages of peaks using the square root of computed signal-to-noise for that spectrum as the weighting factor (signal-to-noise is taken as the ratio of maximum to median abundance in a spectrum). Abundances are given as integers after base peak normalization to 10000.

g) Compute measures of overall spectrum and individual peak variation.

4) *Peak Annotation*: Product ions are labeled using conventional y,b,a type notation, along with immonium ions, internal ions (singly charged precursor only) and common neutral loss ions, including losses from the parent ion. Up to two assignments are given for each m/z, using simple rules for ordering which include proximity to the expected m/z

and ion type. A 0.6 m/z tolerance was used and up to two different ion types for a peak are allowed. For consensus spectra, additional information is given that describes the variability of the peak among the underlying replicate spectra. Assigned, but suspect peaks are labeled with an asterisk. Derived values for the entire spectrum appear in the comment field.

5) *Selection of Best Experimental (Replicate) Spectrum*: Using original sequence-search scores, parent m/z accuracy and fraction of unassigned abundance, a best replicate spectrum was identified. In cases where no good consensus spectrum could be generated due to loss of significant peaks in the averaging process, this spectrum was the only representation for a peptide ion. In cases where only highly impure replicate spectra were available, they were omitted; otherwise a replicate was generated for each consensus spectrum. High confidence peptide identifications having no replicates were also identified and are referred to as 'single' spectra.

6) *Spectrum Reliability*: A series of quality control steps was applied to refine the probability of correct identification and to limit spectra dominated by impurity peaks and from homologous peptides. Beginning with a highest probability of correct identification derived from search engines results, the probability was successively refined using the following factors:

a) Match of y/b ions with theoretical spectrum. The theoretical spectrum was derived from estimated relative cleavage rates for each pair of adjacent amino acids. These rates were derived from a collection of reliably identified peptide spectra, and were divided into two classes, one with mobile protons (numbers of charges greater than the number of basic residues in the peptide — arginine, lysine, and histidine) and the other with no mobile protons.

b) Fraction of unassigned abundance of the largest 20 peaks. This includes peaks that could not be explained as y, b or a ions, internal ions (singly charged only) or ions derived from neutral losses (unassigned peaks are marked as '?' in the peak annotation).

c) Continuity of y or b ion series. Longer series of contiguous y and b ions increase probabilities of correct identification.

d) Number of peptide ions with the same sequence, including different charge states and modifications.

Corrections were derived by comparing results for false positives identifications (from reverse library searching) to true positive results at the same sequence search engine score.

7) *Threshold Setting*

a) Spectra are included in the library when a derived quality score is above a pre-set threshold value. These threshold values are set so that likelihood of being correct for > 95% for all spectra and that quality measures indicate that it has a suitably high S/N and low impurity content for it to reliably serve as a reference mass spectrum. These thresholds depend on the class of peptide involved. In all, 12 classes of peptides, given in Table 1, were distinguished, with thresholds based on results of reversed library searching (for random match false positives) and refined by results from digests of

known proteins where false hits result primarily from peptides with similar sequences (homologous peptides). A tryptic peptide with no missed cleavages, for example, had the lowest threshold for acceptance. Peptide classes were further divided for peptides were 'confirmed' by finding their parent or product peptides as good quality identifications. For instance, if the parent ion for a semitryptic peptide (a non-tryptic cleavage at one terminus) was found to co-elute with an identified tryptic peptide that contained the sequence of that peptide, the peptide was labeled as 'in-source' (likely from fragmentation in the electrospray source). If a missed cleavage was found, and a tryptic subsequence matched a reliably identified peptide, the peptide with the missed cleavage was classified as 'confirmed'. Peptides with missed cleavages not near a terminus or acid residues [6] were rejected. For all but simple tryptic peptides with no missed cleavages and confirmed in source semi-tryptic peptides, higher thresholds were employed to minimize homology false positives. Extra penalties were also applied to non-tryptic peptides that had higher levels of unexplained peaks. These settings were based primarily on results from digests of known proteins.

b) All spectra of peptides having similar precursor m/z were compared and if sufficiently similar (dot product > 0.7) one was selected and the other represented in a homology field (Hom=…). Tryptics were preferred over semitryptics, high scores preferred over low scores and high numbers of underlying replicate spectra over lower numbers. For I/L equivalence, the form with the largest number of replicate spectra was selected and others placed in the homology field.

c) When the signal-to-noise in source spectra were low, important sequence peaks could be missing in derived consensus spectra. To avoid creating such low quality spectra, consensus spectra were rejected when major sequence peaks would be lost. However, to avoid losing spectra for significant peptides, the best spectrum for each peptide ion is included even when a reliable consensus spectrum could not be created. Similarly, when no single high quality replicate spectrum was found, but a good quality consensus spectrum could be created, no replicate spectrum was included.

Library Searching

Matching an acquired (query) spectrum against a library spectrum is a routine way of identifying volatile compounds in the well-established method of gas chromatography/mass spectrometry (GC/MS). These search methods are generally tightly integrated in data systems of those instruments. While these methods have been optimized for electron ionization (EI) spectra [5], with some modifications they work equally well for MS/MS peptide spectra. Peptide-specific scoring methods have been developed, tested and added to the NIST MS search software originally designed for electron ionization mass spectrum searching. They are available in two forms. The first is implemented as a new search type (Peptide search) in an updated version of the NIST MS Search Program for Microsoft Windows® [7]. This updated system was developed at NIST primarily for quality control and will be most useful to individuals already familiar with peptide MS/MS spectra. The second adds a 'Peptide' search type to our current 'dynamic-link library' module, also for Microsoft Windows®. This format is most suitable for integration by programmers into proteomic 'pipelines' and other automated systems for peptide identification. Source code, written in C and C++, is available on request.

NIST MS Search Program: Library spectra may be accessed from an updated version of the NIST MS search software for the NIST/EPA/NIH Mass Spectral Library (Version 2.0e). This program was originally designed for electron ionization spectra and operates under Microsoft Windows®. Many features designed for electron ionization spectra may be useful for peptide MS/MS spectra as well. Different sections of the program are selected using 'tabs' at the bottom left. Specific features available in each section may be selected using menu choices and with the right mouse button (depending on the active window and current cursor position). Each window has a properties dialog for setting display features (right button/properties). Further details are described in the on-line help system. Methods specifically designed for using the peptide ion library have been added, but are not included in the on-line help system. Illustrations of the use of these methods are described below. A key difference between peptides and EI searching is that all peaks in the latter are represented as integers.

The program starts with the 'Lib. Search' tab selected. The list at the upper left, Spec List, which is initially empty, shows search (query) spectra. Any spectra to be searched must first appear in this list. Spectra can be added to this list by using the file/open menu choice to select spectrum files. Spectra in a selected file can be in any of the following formats: msp (generated by the NIST search program), mgf (Mascot), pkl (peak list) or dta (Sequest). For msp spectra, the parent m/z is given as Parent=value in the comment field – other files formats are defined by the software systems that generated them.

Search details are specified in the Options/Library Search Options dialog box (also fourth button from left near the top). Peptide searching is specified by selecting in the 'Spectrum Search Type' box the 'Identity' radio button and 'Peptide' in the drop down list box. Presearch=Default limits matching spectra to library spectra with parent m/z values within the Precursor +/- box. Presearch=Off searches all spectra in the library (useful for finding homologous peptide spectra). You can also set product ion tolerances in the appropriate box. Note that the current libraries contain primarily ion trap spectra, so relatively wide tolerances are most appropriate. Libraries to be searched are selected in the 'Libraries' tab. A separate dialog box is used for setting peptide search scoring and display options (MS/MS Options) or with the Options/MSMS Options menu choice. Settings for scoring are also selected here along with various display preferences (see later).

Double clicking a spectrum in the search list (Spec List) performs a spectrum search, with the Hit List displayed at the lower left. The query spectrum and best library match are shown at the upper and lower right, respectively, with a difference window in between. The top 'Go' button or right mouse/'Library Search' can also be used to initiate a search. In the Hit List, the Score column gives a score used for ranking and Dot gives the spectrum similarity (Dot product of query and library spectra), both of which yield 999 as a perfect match. Prob gives the computed probability of each hit being correct. Other scores may be shown using the MSMS Options box discussed above. If the same peptide appears twice in a hit list, the higher probability value is shown for both. You may resort a column by clicking on its title.

Note that the 'Names' tab is helpful for browsing. Spectra can be transferred between sections of the program by drag-drop methods or by selecting Copy or Send To from the right mouse button over a highlighted spectrum in a list.

Subsets of spectra can be found using the 'Other Search' tab and performing a sequential search (Search/Sequential Method menu choice) after choosing libraries and setting appropriate constraints. These constraints are also used for spectrum searching if the checkbox "Use Constraints" in ON. The last five entries in 'Constraints' tab in the box give peptide-specific conditions:
1) Tags in Comment: An exact or substring match (or mismatch) in any of the field=value entries in the comments field may be selected. Results are case-insensitive (do not depend on letter capitalization).
2) Peptide Sequence: Select amino acids to be present/not present in various locations – these may be selected relative to the N- or C- end
3) Peptide Mobile Protons: Number of protons in excess of number of basic amino acids (R, K, H).
4) Peptide Charge
5) Peptide Number of Residues

Note that 1) allows selection of subsets of spectra type. Here are some examples:
Spec=Consensus – shows only the consensus spectra
Inst=qtof – finds only qtof spectra (ignores ion trap)
Protein=ALBU_BOVIN – shows only bovine serum albumin derived peptides
Mods=^ICAT – ignores all ICAT derivatized spectra

Dynamic-Link Library (DLL): The DLL used for the NIST/EPA/NIH Mass Spectral Library has been modified to allow peptide MS/MS searching. This code is intended for linking to external programs through a documented API (application program interface). The binary and source code (in "C") for the DLL is available on request.

Algorithms: The method of scoring library matches to search (query) spectra was derived from methods used for searching electron ionization spectra. This is a weighted dot product function [5], with the following modifications for peptide spectrum matching:

1) Parent m/z peaks are ignored
2) Isotopic peaks (based on those identified in the library spectrum) are ignored
3) Peaks near parent m/z – 18 are ignored (these are commonly due to isobaric impurity precursor ions.)
4) Neutral loss peaks from the parent ion are assigned a weight of 0.2 instead of 1.0. This is done because such peaks carry little identification information.
5) Peaks matching unidentified peaks in the library spectrum are assigned a weight of 0.2. This is done to account for the possibility that such peaks may be due to impurity ions in the library spectrum.

Additional factors are used for scoring qtof-class spectra against ion trap spectra when the Q-tof option is selected in the MSMS Search Dialog or DLL query. These are:

1) Peaks comparisons begin at the higher of the lowest m/z values for the two spectra being compared.
2) Ion fragments having no basic residues are given a weight of 0.2.
3) Low m/z peaks (< 0.7 parent m/z) are given lower weights

The final score and probabilities incorporate several features other than the dot product described above, namely:

1) OMSSA score – for each hit, the 'expectation value' (probability of a random match) is computed using the scoring system in the Open Mass Spectrometry Search Algorithm (OMSSA) [4b]. In this implementation only sequence peaks (y/b ions) identified in the library spectrum are used. The OMSSA contribution is especially useful in matching qtof with ion trap spectra and for spectra with high levels of impurities (used only when the Omssa checkbox is selected).
2) Total abundance correction – a correction is made for spectra whose signal was contained primarily in a few large peaks. Otherwise, for such spectra, the dot product would be too heavily biased toward a few large peaks.
3) For a comparable degree of spectrum matching, a match to a commonly occurring peptides is more likely to be correct than a match to a rarely observed peptide. A correction for this factor may be optionally applied. It uses the total number of replicate spectra observed for a peptide and was optimized in test studies. For analysis of mixtures likely to contain many proteins not previously observed, this correction should not be used. It is only used for Prob(ability) calculations when used when the # Replicates checkbox is selected.

To speed searching and reduce the size of the hit list, the threshold for accepting spectra for peak-by-peak comparisons may be raised by selecting High or Medium in the Score Threshold drop down list box.

All of these options may be selected from the Option/MS/MS Options dialog accessible from the main menu or the MS/MS Options button in Library Search Options dialog.

Spectrum Fields and Format

The ASCII text version of the library is composed of spectra in the .msp file format, which is an ASCII text format long used with the NIST/EPA/NIH Mass Spectral Library of electron ionization spectra (EI). Peptide spectrum annotation is contained in the 'Comment' field as field_name=field_value pairs. Individual peak annotation follows the m/z – abundance pair for each peak.

The .msp format employed is described below, where specific values should replace the descriptions given in angle brackets (<>).

1) Each spectrum begins with the line:

Name: <peptide sequence>/<charge>

where each amino acid in the <peptide sequence> is represented by the usual (upper case) letter sequence and <charge> is the positive charge on the peptide. The only modification explicitly shown is oxidized methionine as M(O).

2) The second line in a spectrum is:

MW: <exact molar mass of the peptide ion>
This is used for compatibility with electron-ionization spectra and is not used in the peptide library search; this line may be omitted.

3) The third line contains annotation information for the peptide, its origin and its spectrum. It has the format:

Comments: <field1=value field2=value ..>

Comments are composed of a series of space delimited field=value pairs, where values may be embedded within double quotes. All field names are described in Table 3. There is one mandatory field, namely Parent=<m/z>, which is the precursor ion m/z required for searching.

4) The fourth line provides the number of peaks (m/z abundance pairs) in the spectrum:

Num Peaks: <number of peaks>

5) Each peak is represented as a line divided into three tab separated fields.

<m/z><tab><relative abundance><tab>"<peak annotation(s)>"

Abundances are normalized to 10000 (base peak) and the entire peak annotation field is enclosed in quotes. Up to two annotations for a single peak can be given. Peaks appear in order of increasing m/z.

Each annotation field begins with one of the following characters: y, b, a, p, I, or ?. The first three denote ion types formed in peptide fragmentation. p denotes the parent ion, I denotes an immonium ion or internal ion and ? indicates that the peak cannot be assigned to a known cleavage type (including common neutral losses). Symbols y, b or a are followed by an integer, representing the cleaved position in the parent peptide, and then, for neutral loss ions, by –n, where n is the neutral loss in Daltons (note – this is actual mass units, water loss is always -18, for example). ^{13}C isotopic peaks are denoted with a following i (x is used in place of possible isotopic peaks that do not have preceding non-isotopic peaks). If the peak assignment is suspect, an asterisk follows. Multiply charged ions are represented by ^c, where c is the charge. A ‘/’ follows, followed by the difference between measured m/z and theoretical (exact) m/z. Internal ions are represents as Int/*seq*/ where *seq* is the sequence of the internal ion. In qtof spectra, immonium ions

are represented as IXY, where X is the parent amino acid and Y are letters A,B,C.. used to distinguish different immonium products of the amino acid X.

For these assignments, a tolerance of 0.8 m/z was used. In cases where more than one assignment is given peaks are ordered with the following preference: neutral loss from parent (X-n), y, b or a with no neutral loss, y, b or a with -18 or -17 neutral loss, other losses. If a second assignment is possible, a second, comma-separated full assignment is given. For two assignments of the same class described above, the m/z closer to theoretical is preferred. Also, if multiple losses from the parent are possible, only the one closest to theoretical m/z is kept.

For consensus spectra, a space follows the assignment(s), after which is given the number of replicate spectra having that peak and the minimum number of spectra required for that peak to have been reported – these numbers are separated by a '/'. A space follows, after which the median deviation of m/z (in 1/100th of an m/z) of the peaks in the original spectra used to create that peak.

Copyright and Redistribution

Disclaimer

Appendices

Peptide Ion Identity

A peptide ion is uniquely specified by its sequence, charge state and modifications (below).

Modifications

Peptide modifications are given in the following format:

Mods=#/*n*,*aa*,*tag*/*n*,*aa*,*tag*…

Where # is the number of modifications, with modifications separated by a forward slash '/' . Hence, Mods=0 denotes no modifications. Modifications are arranged in order of amino acid position and, if multiple modifications occur at a single position, they are arranged alphabetically by *tag*.

n is the position of the substituted amino acid, starting from 0
aa is the modified amino acid symbol
tag is the name of the modifications as given by Unimod (unimod.org). Except for the protein standard library, only the top 7 modifications were employed (for each modification, the tag is in quotes, followed by mass correction in Daltons):

{ "Oxidation", 15.994915 },
{ "Carbamidomethyl ", 57.02146 },
{ "ICAT_light", 227.12 },
{ "ICAT_heavy", 236.12 },
{ "AB_old_ICATd0", 442.20 },
{ "AB_old_ICATd8", 450.20 },
{ "Acetyl", 42.0106 },
{ "Deamidation", 0.9840 },
{ "Pyro-cmC", -17.026549 },
{ "Pyro-glu", -17.026549 },
{ "Pyro_glu", -18.010565 },
{ "Amide", -0.984016 },
{ "Phospho", 79.9663},
{ "Methyl", 14.0157 },
{ "Carbamyl", 43.00581 },

Table 1
Peptide Classes

Each peptide is assigned a class based on its termini and missed cleavages. These classes were used for setting reliability thresholds for inclusion in the library.

Format: Pep=<terminus description>/<missed cleavage description>

- Tryptic
Tryptic
Tryptic/Miss_good_confirmed
Tryptic/Miss_good_unconfirmed
Tryptic/Miss_bad_confirmed
Tryptic/Miss_bad_unconfirmed (all were rejected)

- Semitryptic – no missed cleavages
N-Semitryp_insource_inotherpep_confirmed
C-Semitryp_insource_inotherpep_confirmed

N-Semitryp_insource_inotherpep_unconfirmed
C-Semitryp_insample_inotherpep_unconfirmed

N-Semitryp_insample_unconfirmed
C-Semitryp_insample_unconfirmed

N-Semitryp_insource_unconfirmed
C-Semitryp_insource_unconfirmed

- Semitryptic – with missed cleavage
(Semitryp_ok, in source and/or confirmed, Semitryp_irreg = not in source and not confirmed)
N-Semitryp_ok/miss_good
C-Semitryp_ok/miss_good

N-Semitryp_irreg/miss_good
C-Semitryp_irreg/miss_good

N-Semitryp_ok/miss_bad
C-Semitryp_ok/miss_bad

N-Semitryp_irreg/miss_bad
C-Semitryp_irreg/miss_bad

I. Terminus:

A. Tryptic – Cleavage after arginine or lysine except when followed by proline

B. Semitryptic – One non-tryptic terminus

N-Semitryptic – non-tryptic N-terminus
C-Semitryptic – non tryptic C-terminus

_insource – left of proline, right of aspartic acid, or parent m/z is a major peak in a parent peptide
_irreg – not insource

_confirmed – parent m/z found as a major peak in parent tryptic peptide
_unconfirmed – no major peak in parent tryptic peptide

-inotherpep – found parent tryptic peptide

II. Missed Cleavage

miss_good – nearby acidic residues or near terminus
miss_bad – not near acidic residues or terminus

_confirmed –internal tryptic peptide found
_confirmed – internal tryptic peptide not found

Table 2. Representing Search Engine Results

Results of search spectra for a peptide are given in the format

Se=#/^x*n*tag1=value, tag2=value, … /…

Where # is the number of search engines that identified the peptide, *n* is the number of spectra for the peptide ion found by the search engine and ^<letter> denotes the search engine, where

^M = Mascot
^O = OMSSA
^S = Sequest
^X = X!Ttandem
^P = PeptideProspector/Batch-Tag

Tags depend on the specific search engine, possible values for spectra are:

sc =	^M, ^S	score (^M=ion score, ^S=xcorr)
td=	^M, ^O, ^X	score difference between top hit and best lower rank tryptic
sr=	^M	ion score – homology score (used as ranking score)
sd=	^M, ^O	score difference between top two hits
hs=	^M	homology score
pr=	^X, ^S	reported probability
pb=	^S	derived probability (peptide prophet)
dc=	^S	Sequest dcn
ps=	^S	Sequest pscore
ex=	^O, ^X	expectation value

Tags for consensus spectra only:

bs=	^M, ^O, ^X, ^S	best score or lowest expectation value of replicates
b2=	^M, ^O, ^X, ^S	second best score or lowest expectation value of replicates
bd=	^M, ^O, ^X	score or expectation value difference between top two hits

For consensus spectra, median and median deviation are given in the format <median>/<deviation>, for individual spectra (replicate or singular), original values are given. For deriving in consensus spectra, the best value among all ‘bs’ (best score) values is used. These values are converted to probabilities prior to comparison.

Table 3. Fields in Comments:

Spec	Consensus, (Best) Replicate or Singular
Pep	Peptide Class (see Table 1)
Fullname	In format preAA.sequence.postAA/n where preAA and postAA are flanking amino acids and n is the charge state. Oxidized methionine oxidation is represented by M(O)
Mods	Modifications, see separate section
Parent	Observed parent m/z
Inst	Instrument class – qtof, it (ion trap, default)
Mz_diff	Observed – Theoretical m/z
Mz_exact	Exact theoretical m/z for parent ion
Mz_av	Average mass m/z for parent ion
Protein	Identifier for first protein matched that contained the peptide sequence, in quotes (fasta file provided separately)
Pseq	Number of sequences for protein group/proteins in group
Organism	Yeast, Radiodurans, Human, Protein (single) …
Se	Search engine results, see Table 2.
Sample	Sample code(s) for source spectra, each code is followed by number of spectra used for building consensus spectrum and total number of good quality replicate spectra found (comma separated variables)
Nreps	Number of replicate spectra used/total number identified
Missing	Fraction of total abundance in peaks not present in consensus spectrum
Parent_med	Median parent m/z of replicates
Max2med_orig	Maximum-to-median abundance in original replicate spectra, exact or median
Dotfull	Median dot product of all pairs of replicate spectra used in creating consensus spectrum
Dot_cons	Median dot product of all spectra used for creating consensus spectrum with consensus spectrum
Flags	Internal Use
Dotbest	Dot product of consensus with selected best spectrum
Unassign_all	Fraction of all abundance not assigned to known fragmentation paths
Unassigned	Fraction of abundance of top 20 peaks not assigned to known fragmentation paths
Naa	Number of amino acids in peptide
DUScorr	Probability corrections for match to theoretical spectrum (dot), unassigned abundance and sequential y/b ions, comma separated
Dottheory	Match to theoretical spectrum

Pfin	Final computed relative probability for spectrum being correct relative to random spectrum (inverse of expectation value)
Probcorr	Correction to probability due to peptide type and other non-spectral factors
Tfratio	Relative probability that the peptide identification is correct
Pfract	Median relative abundance of sequence in identified runs
Hom	Rejected homologous peptide match (spectrum matched more than one peptide)
OverflowRec	Omitted Homology Lines
	[Replicate / Singular Spectrum Only]
Npksorig	Number of peaks in original spectrum
Dotvscon	Dot product of replicate with consensus spectrum
Scan	Scan number in original spectrum file
Origfile	Name of original peak list file

References

[1] Sources of spectra are included in the text file references.txt, which is provided along with the library. Each line in the file contains up to seven tab-delimited fields, namely, Dataset, Contributor, Number of Files, Source, Reference, Title, Authors. Internet links are in curly braces {}. Sources of spectra are given for each spectrum retrieved by the DLL and user interface versions of the search software.

[2] Peptides were derived from tryptic digests of the following proteins: bovine-actin, bovine-alpha-casein, bovine-alpha-lactalbumin, bovine-beta-casein, bovine-beta-lactoglobulin, bovine-carbonic-anhydrase, bovine-serotransferrin, bovine-ubiquitin, bovine-serum-albumin, chicken-lysozyme, chicken-ovalbumin, ecoli-beta-galactosidase, ecoli-isomerase, equine-myoglobin, horseradish-peroxidase, human-keratins (mixture), porcine-trypsin, rabbit-gapdh, rabbit-phosphorylaseb

[3]
a) Omenn, G.S.; States, D.J.; Adamski, M.; Blackwell, T.W.; Menon, R.; Hermjakob, H.; Apweiler, R.; Haab, B.B.; Simpson, R.J.; Eddes, J.S.; Kapp, E.A.; Moritz, R.L.; Chan, D.W.; Rai, A.J.; Admon, A.; Aebersold, R.; Eng, J.; Hancock, W.S.; Hefta, S.A.; Meyer, H.; Paik, Y.-K.; Yoo, J.-S.; Ping, P.; Pounds, J.; Adkins, J.; Qian, X.; Wang, R.; Wasinger, V.; Wu, C.Y.; Zhao, X.; Zeng, R.; Archakov, A.; Tsugita, A.; Beer, I.; Pandey, A.; Pisano, M.; Andrews, P.; Tammen, H.; Speicher, D.W.; Hanash, S.M.; "Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database", Proteomics 2005, 5 (13), 3226-3245.
b) HUPO Plasma Proteome Project Website: http://www.hupo.org/research/hppp/

[4] The four search engines were, Mascot, Omssa, Sequest and X!tandem.

a) Mascot: Perkins, D.N.; Pappin, D.J.C.; Creasy, D.M.; Cottrell, J.S.; "Probability-based protein identification by searching sequence databases using mass spectrometry data"; Electrophoresis; 1999, 20(18): 3551-3567;
Ion scores relative to homology scores were converted to probabilities based on reverse library search scores.

b) OMSSA : Geer, L.Y.; Markey, S.P.; Kowalak, J.A.; Wagner, L.; Xu, M.; Maynard, D.M.; Yang, X.; Shi, W.; Bryant, S.H.; "Open Mass Spectrometry Search Algorithm", J. Proteome Res.; 2004; 3(5); 958-964.

c) Sequest: Eng, J.K.; McCormack, A.L.; Yates, J.R. III; "An Approach to Correlate Tandem Mass Spectral Data with Amino Acid Sequences in a Proteins Database", J. Am. Soc. Mass Spectrom.; 1994, 5(11), 976-989. Searches were done by the Institute of Systems Biology for Yeast digests as part of its Peptide Atlas project (http://www.peptideatlas.org/). Probabilities from PeptideProphet [Keller, A.; Nesvizhskii, A.I.; Kolker, E.; Aebersold, R.; "Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search"; Anal. Chem.; 2002, 74(20), 5383-5392.] were used as input scores.

d) X!tandem: Craig, R; Beavis, R.C.; “TANDEM: matching proteins with tandem mass spectra”, Bioinformatics. 2004; 20(9), 1466-1467; Craig, R.; Cortens, J.P.; Beavis, R.C.; “Open Source System for Analyzing, Validating, and Storing Protein Identification Data”, J. Proteome Res.; 2004; 3(6); 1234-1242; Expectation values served as original scores.

[5] Stein, S.E.; Scott, D.R.; “Optimization and Testing of Library Search Algorithms for Compound Identification”, J. Am. Soc. Mass Spectrom, 1994, 5(9), 859-866.

[6]
a) Parker, K.C.; “Scoring Methods in MALDI Peptide Mass Fingerprinting: ChemScore and the ChemApplex Program”, J. Am.Soc. Mass Spectrom. 2002, 13(1), 22-39.
b) Yen, C.-Y.; Russell, S.; Mendoza, A.M.; Meyer-Arendt, K.; Sun, S.; Cios, K.J.; Ahn, N.G.; Resing, K.A.; “Improving Sensitivity in Shotgun Proteomics Using a Peptide-Centric Database with Reduced Complexity: Protease Cleavage and SCX Elution Rules from Data Mining of MS/MS Spectra”; Anal. Chem. 2006; 78(4); 1071-1084.

[7]
a) NIST/EPA/NIH Mass Spectral Library: http://www.nist.gov/srd/nist1a.htm
b) http://chemdata.nist.gov/