Finding Protein Sequences Using PROWL

PROWL is a collection of tools for the identification of protein sequences using input data derived from mass spectrometry. This unit presents protocols for several of the individual PROWL tools. Specifically, PepFrag allows for the analysis of a single spectrum derived from tandem mass spectrometry (see Basic Protocol 1). GPM, on the other hand, provides for the analysis of multiple MS/MS spectra (see Basic Protocol 2). An additional protocol introduces ProFound for analyzing a single spectrum of peptide mass fingerprinting data (see Basic Protocol 3).

USING PROWL WITH THE WEB INTERFACE FROM THE ROCKEFELLER SERVER WITH TANDEM MASS SPECTROMETRY (MS/MS) DATA: PepFrag

BASIC PROTOCOL 1

There are several different tools in PROWL that make it possible to input tandem mass spectra in order to match them to sequences. The simplest tool for this type of input is called PepFrag. This tool allows the user to work with one spectrum at a time, to directly compare sequences to data with very little automated assistance.

Necessary Resources

Hardware

Any Internet-connected computer

Software

Web browser

Files

Text list of MS/MS fragment peak masses, with each mass on a single line

Prepare for a PepFrag search

1. Point the Web browser to *http://prowl.rockefeller.edu/prowl/PEPFRAGch.html*. The page illustrated in Figure 13.2.1 should appear. This page offers the user several demonstration data sets, which can be used as tutorials for new users, as well as for testing/validating the system when new versions of the software are installed.

2. Select the appropriate Database and Kingdom, reflecting the source of the original sample. The Databases available are the NCBI FASTA versions of SWISS-PROT (protein), nr (protein), and dbEST (translated nucleotide).

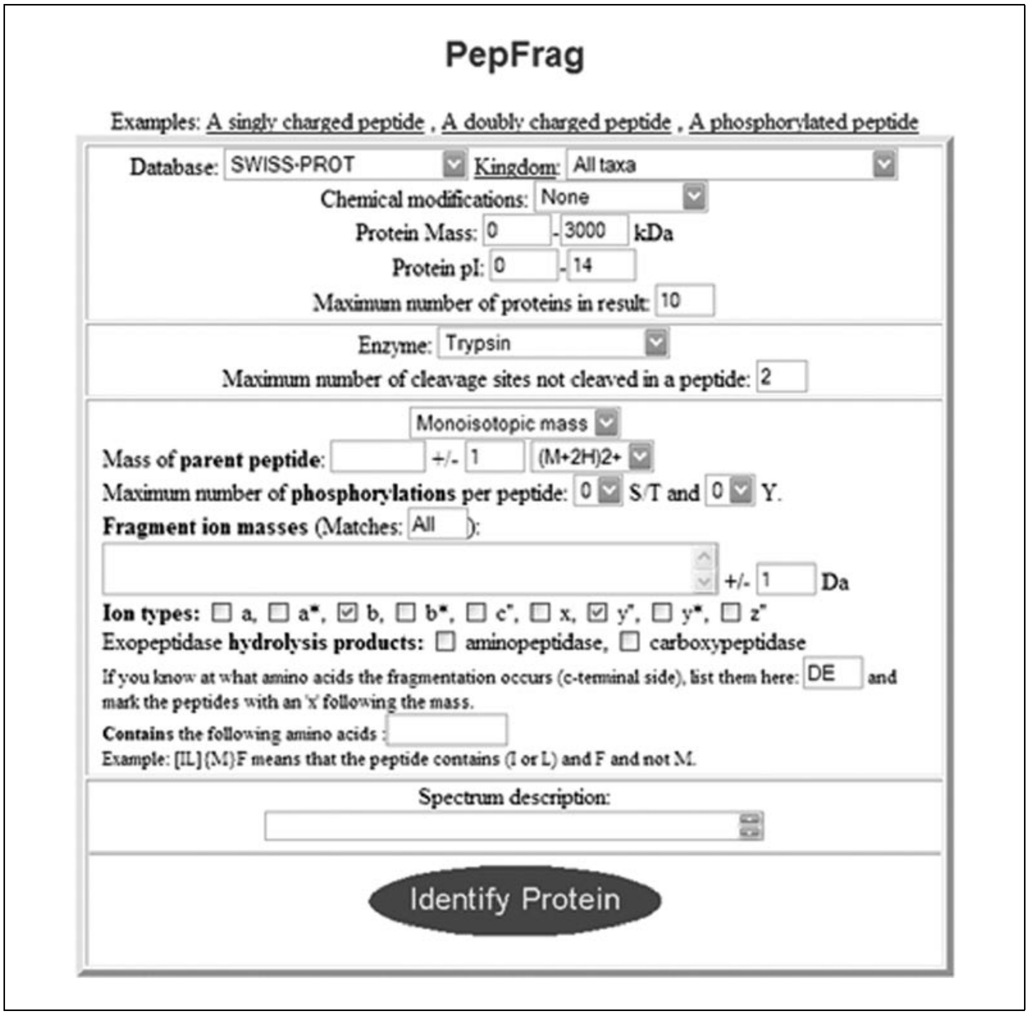
   The term Database is used here in a rather inexact way to refer to any substantial list of protein or nucleotide sequences. Formally it refers to specific data structures used to efficiently store and retrieve large amounts of information. In this context, it means a set of sequences stored in a reference file in FASTA format (APPENDIX 1B). The sequence collections used by the present version of PepFrag are periodically downloaded from ftp://ftp.ncbi.nih.gov/blast/db/FASTA. The term Kingdom refers to the taxonomic classification of the biological source of the sample.

3. Select the chemical modifications used in the experimental protocol.

   The chemical modifications allowed are for the chemical blocking of cysteine residues. This type of modification is part of most proteomic digestion protocols.

Contributed by Ronald Beavis and David Fenyö

Current Protocols in Bioinformatics (2004) 13.2.1-13.2.20



PepFrag

Examples: A singly charged peptide , A doubly charged peptide , A phosphorylated peptide

Database: SWISS-PROT Kingdom: All taxa

Chemical modifications: None

Protein Mass: 0 - 3000 kDa

Protein pI: 0 - 14

Maximum number of proteins in result: 10

Enzyme: Trypsin

Maximum number of cleavage sites not cleaved in a peptide: 2

Monoisotopic mass

Mass of **parent peptide**: +/- 1 (M+2H)2+

Maximum number of **phosphorylations** per peptide: 0 S/T and 0 Y.

Fragment ion masses (Matches: All):

+/- 1 Da

Ion types: □ a, □ a*, ☑ b, □ b*, □ c", □ x, ☑ y", □ y*, □ z"

Exopeptidase **hydrolysis products:** □ aminopeptidase, □ carboxypeptidase

If you know at what amino acids the fragmentation occurs (c-terminal side), list them here: DE and mark the peptides with an 'x' following the mass.

Contains the following amino acids :

Example: [IL]{M}F means that the peptide contains (I or L) and F and not M.

Spectrum description:

Figure 13.2.1 The start page for a PepFrag session.

4. Enter the protein intact mass and pI ranges to be searched.

   Proteins with intact masses and pI outside of these ranges will be excluded from consideration during the search. It is good practice not to limit these parameters too strictly, because the mass and pI are calculated on the full sequence as it is written in the sequence list. If the peptide of interest was derived from a proteolytic fragment of a larger protein, or if the mature form of the protein is significantly different from the sequence in the list, it may be missed.

5. Enter the maximum number of protein candidates to return for a search.

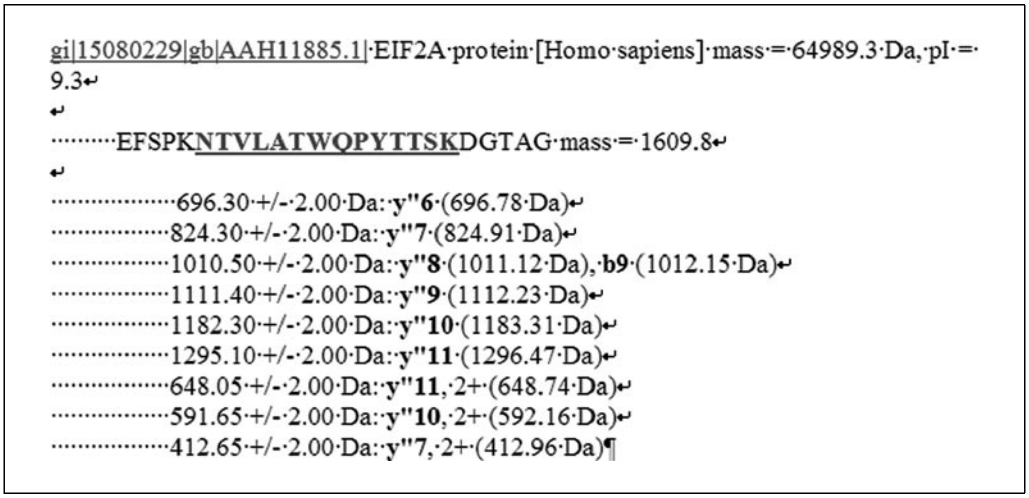
6. In the Enzyme menu, select an appropriate enzymatic cleavage chemistry for the particular protocol. In the text box on the following line, enter the number of missed cleavage sites to consider.

   The selections available are for proteolytic enzymes commonly used in proteomics. Trypsin is by far the most widely used enzyme.

7. On the "Mass of parent peptide line," enter the parent ion *m*/*z* ratio, peak type, *m*/*z* tolerance, and the measured charge state of the parent ion.

8. Enter the number of potential phosphorylations to be considered for a particular peptide.

9. Cut and paste the list of fragment ion *m*/*z* values into the "Fragment ion masses" box.



```
gi|15080229|gb|AAH11885.1|·EIF2A·protein·[Homo·sapiens]·mass·=·64989.3·Da,·pI·=·
9.3↵
↵
··········EFSPKNTVLATWQPYTTSKDGTAG·mass·=·1609.8↵
↵
··················696.30·+/-·2.00·Da:·y"6·(696.78·Da)↵
··················824.30·+/-·2.00·Da:·y"7·(824.91·Da)↵
··················1010.50·+/-·2.00·Da:·y"8·(1011.12·Da),·b9·(1012.15·Da)↵
··················1111.40·+/-·2.00·Da:·y"9·(1112.23·Da)↵
··················1182.30·+/-·2.00·Da:·y"10·(1183.31·Da)↵
··················1295.10·+/-·2.00·Da:·y"11·(1296.47·Da)↵
··················648.05·+/-·2.00·Da:·y"11,·2+·(648.74·Da)↵
··················591.65·+/-·2.00·Da:·y"10,·2+·(592.16·Da)↵
··················412.65·+/-·2.00·Da:·y"7,·2+·(412.96·Da)¶
```

Figure 13.2.2 Results obtained from a PepFrag search.

10. Select the fragmentation ion types to be considered.
11. Select exopeptidase hydrolysis products, if appropriate.
12. Enter any information already known about the peptide's sequence, e.g., at what amino acids the fragmentation occurs and what amino acids the peptide contains.
13. Enter a text description of the spectrum for inclusion in the output display, if desired.
14. Press the Identify Protein button to submit the search.

Interpreting the results of a PepFrag search

The results returned from PepFrag are illustrated in Figure 13.2.2. The result format is a simple HTML page that presents the best possible candidate sequences for a given set of input fragment *m*/*z* values and search parameters.

15. The first line contains the information available about the identified protein readily available from the FASTA sequence list file, along with a calculated protein mass and pI.

    Clicking on the hyperlinked accession number on the first line will browse to a page that has links to other information sources for that accession number.

16. The next line is the sequence of the proposed peptide sequence match to the mass spectrum, along with the mass of the peptide in daltons (Da). The underlined section of the sequence is the identified segment, with the residues N and C-terminal to the identified segment shown in normal text.

    Clicking on the hyperlinked sequence takes the browser to a helper tool, ProteinInfo, which allows the manipulation of the sequence and calculation of a number of different parameters about the sequence that may be of interest, such as its mass, predicted MS/MS fragmentation pattern, hydrophobicity plots, and a link enabling the user to submit the sequence to BLAST (UNITS 3.3 & 3.4) at NCBI.

17. The next set of lines are as follows:

    a. The observed *m*/*z* value.

    b. The fragment ion *m*/*z* error tolerance entered on the input page.

    c. The assigned ion type and position.

    d. The calculated mass for the assigned ion type.

The ion types refer to the Roepstorff-Folman convention (Aebersold and Goodlett, 2001). If more than one ion type can be assigned (within error) to the observed ion, additional ions will be listed on the same line (see the line for m/z = 1010.50 in Fig. 13.2.2).

BASIC PROTOCOL 2

USING PROWL WITH THE WEB INTERFACE FROM THE ROCKEFELLER SERVER WITH TANDEM MASS SPECTROMETRY (MS/MS) DATA: GPM

The Global Proteome Machine (GPM) is a new addition to PROWL that allows the user to input large collections of tandem mass spectra to match them to sequences (Craig and Beavis, 2003). It is a much more comprehensive tool than PepFrag, and is based on an open-source software project that allows investigators to download and use the software and user interface themselves. It also allows the user to analyze thousands of tandem mass spectra at once, with a user interface that can display genomic information and SNP data, where they are available. This protocol presents a simple GPM search. Support Protocol describes setting advanced parameters.

Necessary Resources

Hardware

Any Internet-connected computer

Software

Web browser

Files

A collection of MS/MS fragment peak masses and parent ion masses, in PKL, DTA, or Mascot Generic Format (Matrix Science)

Prepare for a GPM search

1. Point the Web browser to *http://h.thegpm.org*. The page illustrated in Figure 13.2.3 should appear.

2. Using the Browse button near the top of the page, locate the properly formatted file that contains the peak tables for the MS/MS spectra to be identified.

3. Select the appropriate "taxon," reflecting the source of the original sample. This selection will determine the list of sequences searched.

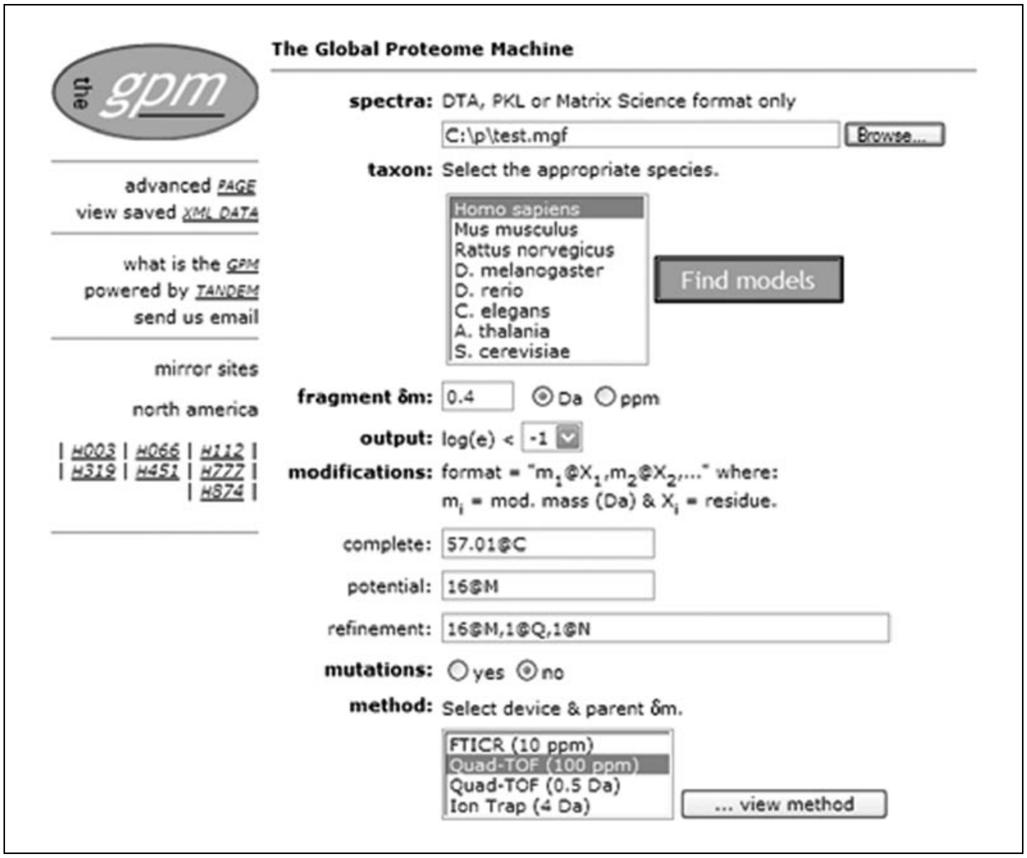
   GPM is organized around the concept of complete proteomes. Therefore, each of the taxon names represents the protein sequences corresponding to the best current translation of open reading frames of a complete genome into assembled proteins. The current public version of the GPM uses ENSEMBL sequences for most species, except S. cerevisiae (SGD) and A. thalania (ATH1). Common contaminant sequences (such as trypsin and bovine serum albumin) are included in all searches.

4. In the box labeled "fragment δm," enter the appropriate fragment ion mass tolerance, in daltons (Da) or parts per million (ppm).

   A good rule of thumb is to set this tolerance to be fairly broad, but to keep it less than 0.5 Da, if possible. If an unfamiliar instrument is being used, iteratively determine a good setting using standard proteins before using true unknowns.

5. On the line for "output," select the desired expectation value cutoff.

   GPM uses an expectation value to measure the confidence in an identification. Only proteins with expectation values less that the cutoff value will be reported on the main report page.



The Global Proteome Machine

spectra: DTA, PKL or Matrix Science format only

C:\p\test.mgf

taxon: Select the appropriate species.

Homo sapiens (selected)
Mus musculus
Rattus norvegicus
D. melanogaster
D. rerio
C. elegans
A. thalania
S. cerevisiae

fragment δm: 0.4 ● Da ○ ppm

output: log(e) < -1

modifications: format = "$m_1@X_1, m_2@X_2, \ldots$" where: m_i = mod. mass (Da) & X_i = residue.

complete: 57.01@C

potential: 16@M

refinement: 16@M,1@Q,1@N

mutations: ○ yes ● no

method: Select device & parent δm.

FTICR (10 ppm)
Quad-TOF (100 ppm) (selected)
Quad-TOF (0.5 Da)
Ion Trap (4 Da)

Figure 13.2.3 The start page for a GPM session.

6. In the "modifications" section of the page, in the box labeled "complete," enter a list of "complete" residue modifications to be used in the search. The choice of modifications will depend on the experimental protocols used; only chemical derivatizations that result in stochiometrically complete reactions should be entered here.

   This user interface does not have a list of "fixed" modifications available; the user is expected to know the mass difference associated with the modifications required. For example, iodoacetamide derivatization of free sulfhydryl groups results in a mass increase of 57.01 Da to cysteine residue, so the entry `57.01@C` *indicates this modification. Additional modifications to other residues can be made using the same notation, separating each type of modification with a comma. Modification masses can be positive or negative. Only one modification entry is allowed per residue type—in the case where multiple entries have been made in error, only the last.*

7. In the "modifications" section of the page, in the box labeled "potential," enter a list of "potential" modifications to be used in the search. The format for entering the modifications is the same as for step 4.

   Potential modifications—often referred to as "partial" modifications—are used to allow for the possibility of any type of residue derivatization where the stochiometry is less than 100%. Several types of common potential modifications produced by normal sample handling are the oxidation of methionine (`16@M`*) and the deamidation of glutamine and asparagine (*`1@Q,1@N`*). Common potential modifications due to biological post-translational modification of a protein are phosphorylation of serine or threonine (*`79.98@S,79.98@T`*), acetylation at lysine residues (*`42.04@K`*), and proline oxidation to hydroxproline (*`16@P`*). The list of potential modifications used in this box should only include the most likely types of modifications, such as methionine oxidation. Rare modifications should be entered in the following step.*

8. Again in the "modifications" section of the page, in the box labeled "refinement," enter a list of "potential" modifications to be used in the refinement process.

   The GPM is one of the first generation of tandem spectrum search interfaces to introduce the idea of "refining" identifications. The list of potential modifications used in step 5 is used to rapidly scan through the proteome list, finding proteins that may be represented by the mass spectra submitted. In the refine process, this more extensive list of modifications is checked against proteins that were found in the first pass, greatly speeding up the sequence-assignment process. In this refinement step, cleavage at all residues and protein N-terminal modifications are also considered for the shortened list of proteins.

9. From the two radio buttons ("yes" and "no") next to "mutations," select whether point mutation analysis should be performed as the final step in the refinement process.

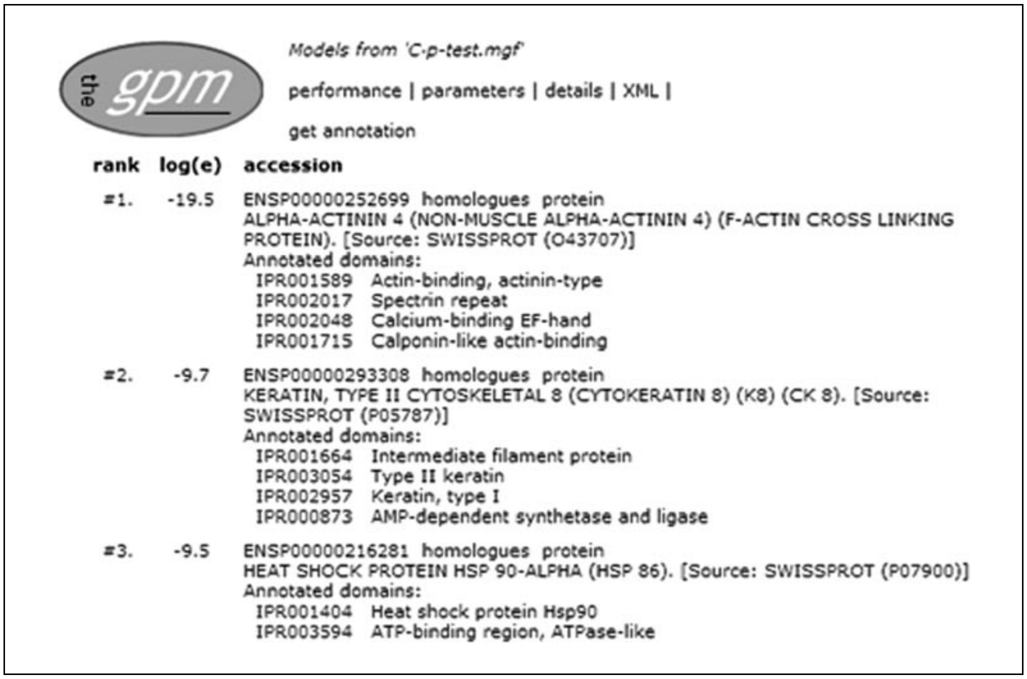
   The GPM is also the first generation of tandem-spectrum search interfaces that allows the analysis of point mutation as a part of the main search process. If "yes" is selected, all possible single-point mutations for a peptide are tested against the available set of mass spectra to find a possible match. This comparison is done at the end of the refinement process, and only the potential modifications and cleavage conditions used in the first round of the search are applied to the point-mutated sequences.

10. From the menu under "method," select a general method to be used, depending on the type of mass spectrometer used to collect the data. The full contents of a particular method can be reviewed by pressing the ". . . view method" button.

11. Press the "Find models" button (to the right of the "taxon" list) to submit the data and begin the search.

Interpreting the results of a GPM search

The results returned from the GPM are illustrated in Figure 13.2.4. The result format is a simple HTML page that presents the best possible candidate sequences for a given set of input tandem mass spectra and selected search parameters. This page is organized around the protein sequences that best fit the set of mass spectra analyzed.



Models from 'C-p-test.mgf'

```
rank  log(e)  accession
 #1.   -19.5  ENSP00000252699  homologues  protein
              ALPHA-ACTININ 4 (NON-MUSCLE ALPHA-ACTININ 4) (F-ACTIN CROSS LINKING
              PROTEIN). [Source: SWISSPROT (O43707)]
              Annotated domains:
                IPR001589   Actin-binding, actinin-type
                IPR002017   Spectrin repeat
                IPR002048   Calcium-binding EF-hand
                IPR001715   Calponin-like actin-binding
 #2.    -9.7  ENSP00000293308  homologues  protein
              KERATIN, TYPE II CYTOSKELETAL 8 (CYTOKERATIN 8) (K8) (CK 8). [Source:
              SWISSPROT (P05787)]
              Annotated domains:
                IPR001664   Intermediate filament protein
                IPR003054   Type II keratin
                IPR002957   Keratin, type I
                IPR000873   AMP-dependent synthetase and ligase
 #3.    -9.5  ENSP00000216281  homologues  protein
              HEAT SHOCK PROTEIN HSP 90-ALPHA (HSP 86). [Source: SWISSPROT (P07900)]
              Annotated domains:
                IPR001404   Heat shock protein Hsp90
                IPR003594   ATP-binding region, ATPase-like
```

Figure 13.2.4 Initial results page obtained from a GPM search.

Main ("initial results") page (Fig. 13.2.4)

12. Links that have the following behaviors are provided on the top of the page in a menu bar:
    a. performance: displays performance statistics about a particular search.
    b. parameters: displays the search parameters used for a search.
    c. details: an alternative set of views that are organized around the individual spectra, rather than identified proteins.
    d. XML: retrieves the XML file that contains all of the data about the results of a search (this file can be stored and uploaded for later viewing).

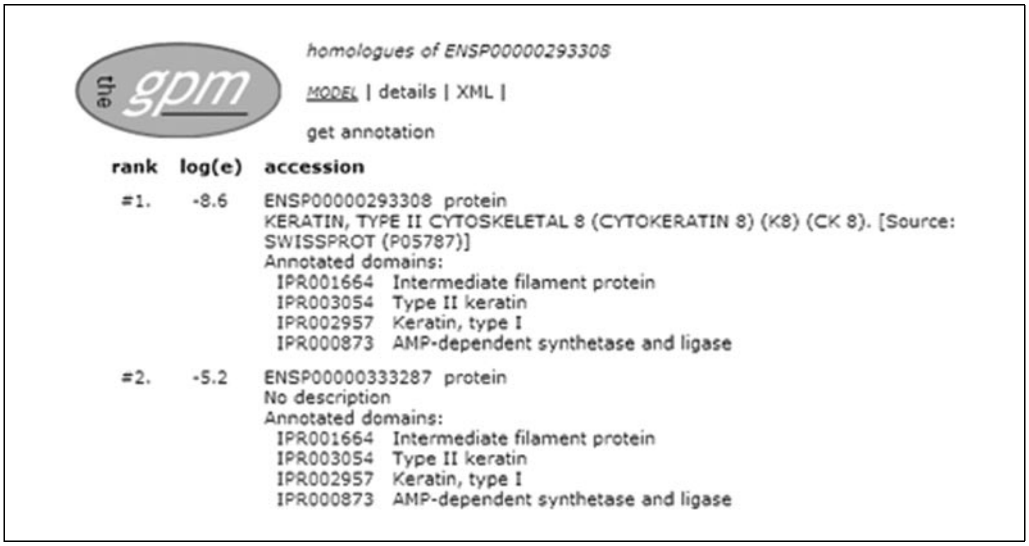
13. The main list of proteins illustrated in Figure 13.2.4 corresponds to the best-scoring proteins found during the search. Each protein listed has an associated expectation value and a short description derived from the primary information sources associated with the accession number for the protein (EBI, SCD, or TGIR). The precise form of this additional information will change from time to time, but it is meant to give the user a quick idea of what proteins have been observed.

14. In Figure 13.2.4, the links beside the accession number for the protein are to pages that summarize additional information:
    a. homologue: this page (Fig. 13.2.5) lists all of the proteins that could be identified with the subset of the mass spectra that were used to identify the listed protein (see steps 15 and 16).
    b. protein: this page (Fig. 13.2.6) lists all of the evidence for the identification of this sequence (see steps 17 to 20).

Homologue page (Fig. 13.2.5)

15. Links that have the following behaviors are provided on the top of the page in a menu bar:
    a. model: returns to the main model page corresponding to this page.
    b. details: the same effect as the details link on the main page (see step 12).
    c. XML: the same effect as the XML link on the main page (see step 12).

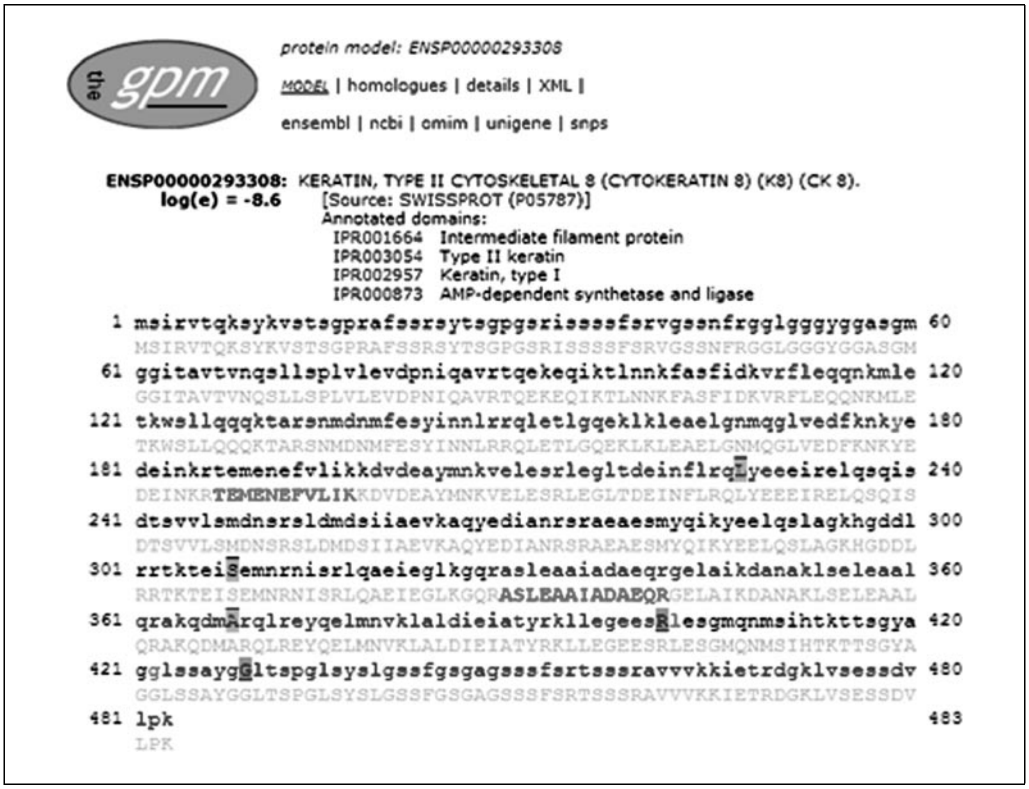
16. The format of the list of proteins is similar to the main page, starting with the best protein fitting the list of mass spectra searched.



```
homologues of ENSP00000293308

rank  log(e)  accession
=1.   -8.6    ENSP00000293308 protein
              KERATIN, TYPE II CYTOSKELETAL 8 (CYTOKERATIN 8) (K8) (CK 8). [Source:
              SWISSPROT (P05787)]
              Annotated domains:
                IPR001664  Intermediate filament protein
                IPR003054  Type II keratin
                IPR002957  Keratin, type I
                IPR000873  AMP-dependent synthetase and ligase
=2.   -5.2    ENSP00000333287 protein
              No description
              Annotated domains:
                IPR001664  Intermediate filament protein
                IPR003054  Type II keratin
                IPR002957  Keratin, type I
                IPR000873  AMP-dependent synthetase and ligase
```

Figure 13.2.5 Homologue page for one protein.


the gpm
protein model: ENSP00000293308
MODEL | homologues | details | XML |
ensembl | ncbi | omim | unigene | snps

```
ENSP00000293308: KERATIN, TYPE II CYTOSKELETAL 8 (CYTOKERATIN 8) (K8) (CK 8).
  log(e) = -8.6   [Source: SWISSPROT (P05787)]
                  Annotated domains:
                    IPR001664  Intermediate filament protein
                    IPR003054  Type II keratin
                    IPR002957  Keratin, type I
                    IPR000873  AMP-dependent synthetase and ligase

  1 msirvtqksykvstsgprafssrsytsgpgsrisssssfsrvgssnfrgglgggyggasgm 60
    MSIRVTQKSYKVSTSGPRAFSSRSYTSGPGSRISSSSSFSRVGSSNFRGGLGGGYGGASGM
 61 ggitavtvnqsllsplvlevdpniqavrtqekeqiktlnnkfasfidkvrfleqqnkmle 120
    GGITAVTVNQSLLSPLVLEVDPNIQAVRTQEKEQIKTLNNKFASFIDKVRFLEQQNKMLE
121 tkwsllqqqktarsnmdnmfesyinnlrrqletlgqeklkleaelgnmqglvedfknkye 180
    TKWSLLQQQKTARSNMDNMFESYINNLRRQLETLGQEKLKLEAELGNMQGLVEDFKNKYE
181 deinkrtemenefvlikkdvdeaymnkvelesrlegltdeinflrqLyeeeirelqsqis 240
    DEINKRTEMENEFVLIKKDVDEAYMNKVELESRLEGLTDEINFLRQLYEEEIRELQSQIS
241 dtsvvlsmdnsrsldmdsiiaevkaqyedianrsraeaesmyqikyeelqslagkhgddl 300
    DTSVVLSMDNSRSLDMDSIIAEVKAQYEDIANRSRAEAESMYQIKYEELQSLAGKHGDDL
301 rrtkteiSemnrnisrlqaeieglkgqrasleaaiadaeqrgelaikdanaklseleaal 360
    RRTKTEISEMNRNISRLQAEIEGLKGQRASLEAAIADAEQRGELAIKDANAKLSELEAAL
361 qrakqdmArqlreyqelmnvklaldieiatyrkllegeesRlesgmqnmsihtkttsgya 420
    QRAKQDMARQLREYQELMNVKLALDIEIATYRKLLEGEESRLESGMQNMSIHTKTTSGYA
421 gglssaygGltspglsyslgssfgsgagsssfsrtsssravvvkkietrdgklvsessdv 480
    GGLSSAYGGLTSPGLSYSLGSSFGSGAGSSSFSRTSSSRAVVVKKIETRDGKLVSESSDV
481 lpk                                                          483
    LPK
```

Figure 13.2.6 Protein model page, upper section showing sequence coverage. This black and white facsimile of the figure is intended only as a placeholder; for full-color version of figure go to *http://www.interscience.wiley.com/c_p/colorfigures.htm*.

The term homologue is used here in a somewhat different sense than normal. In this case, homologue proteins share the same set of mass spectra, rather than the same set of sequences. In many cases, the mass spectra will be associated with similar sequences in each protein, but this may not always be true. A protein is listed as a homologue if it was assigned to at least one spectrum that is also assigned to the first ranked protein. If a homologue was also assigned to mass spectra not associated with the first ranked protein, it will also appear in the main page.

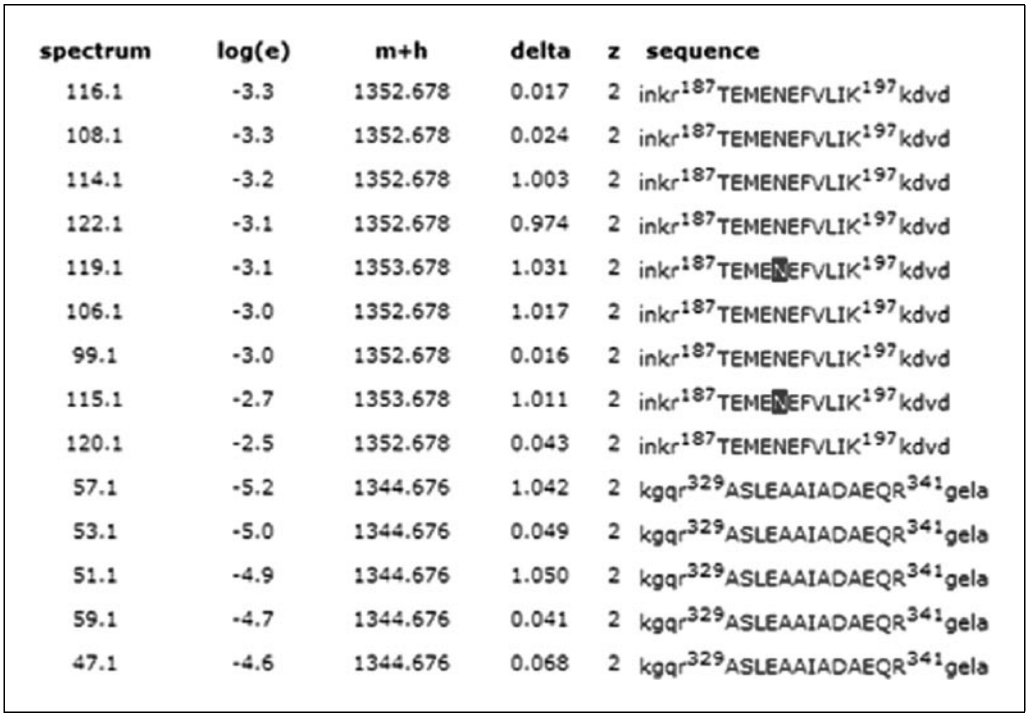
Protein page (Figs. 13.2.6 and 13.2.7)

17. Links are provided on the top of the page (Fig. 13.2.6) in a menu bar that have the same behaviors as those on the homologues page (see step 15).

18. There is an additional menu bar that provides links to external sources of information (e.g., "ensembl," "ncbi," "omim," "snps," shown in Fig. 13.2.6) that can be accessed using the accession number associated with the protein. The links in this bar differ from protein to protein and taxon to taxon, as the same types of information may not be available for different proteins.

19. The display in Figure 13.2.6 is the top section of the protein page, showing the placement of the peptides that have been associated with the protein's sequence. When available, this information is superimposed on the protein sequence's gene model, showing exon boundaries and known single nucleotide polymorphisms (SNPs).

    The upper line (in lowercase) represents the gene model, with exon boundaries indicated by a color change in the residues from black to blue. Residues shown in red are necessary for correct exon overlaps. Residues with green backgrounds are synonymous SNPs and red backgrounds are nonsynonymous SNPs. The lower line (in uppercase) shows the peptides



spectrum	log(e)	m+h	delta	z	sequence
116.1	-3.3	1352.678	0.017	2	inkr^{187}TEMENEFVLIK^{197}kdvd
108.1	-3.3	1352.678	0.024	2	inkr^{187}TEMENEFVLIK^{197}kdvd
114.1	-3.2	1352.678	1.003	2	inkr^{187}TEMENEFVLIK^{197}kdvd
122.1	-3.1	1352.678	0.974	2	inkr^{187}TEMENEFVLIK^{197}kdvd
119.1	-3.1	1353.678	1.031	2	inkr^{187}TEMENEFVLIK^{197}kdvd
106.1	-3.0	1352.678	1.017	2	inkr^{187}TEMENEFVLIK^{197}kdvd
99.1	-3.0	1352.678	0.016	2	inkr^{187}TEMENEFVLIK^{197}kdvd
115.1	-2.7	1353.678	1.011	2	inkr^{187}TEMENEFVLIK^{197}kdvd
120.1	-2.5	1352.678	0.043	2	inkr^{187}TEMENEFVLIK^{197}kdvd
57.1	-5.2	1344.676	1.042	2	kgqr^{329}ASLEAAIADAEQR^{341}gela
53.1	-5.0	1344.676	0.049	2	kgqr^{329}ASLEAAIADAEQR^{341}gela
51.1	-4.9	1344.676	1.050	2	kgqr^{329}ASLEAAIADAEQR^{341}gela
59.1	-4.7	1344.676	0.041	2	kgqr^{329}ASLEAAIADAEQR^{341}gela
47.1	-4.6	1344.676	0.068	2	kgqr^{329}ASLEAAIADAEQR^{341}gela

Figure 13.2.7 Protein model page, lower section showing spectrum assignments.

that correspond to mass spectra in bold red. Residues with red backgrounds correspond to observed point mutations. Chemical modifications are not shown on this display. Clicking on observed sequence links to the evidence for that assignment, lower down on the same page (Fig. 13.2.7).

20. The display in Figure 13.2.7 is the bottom section of the protein page. It is composed of a table that indicates which spectra are associated with a particular sequence and information about that assignment. The spectra are listed in the order in which the assigned peptides are arranged in the protein, from N-terminal to C-terminal. If multiple spectra are associated with the same peptide, these spectra are arranged with the best assignment appearing first.

    a. spectrum: the number format is `xxx.zz`, where `xxx` is the position of the spectrum in the original spectrum listing file, and `zz` is the peptide assignment number. If `zz` is greater than 1, the spectrum has been equivalently assigned to more than one sequence in the same protein.

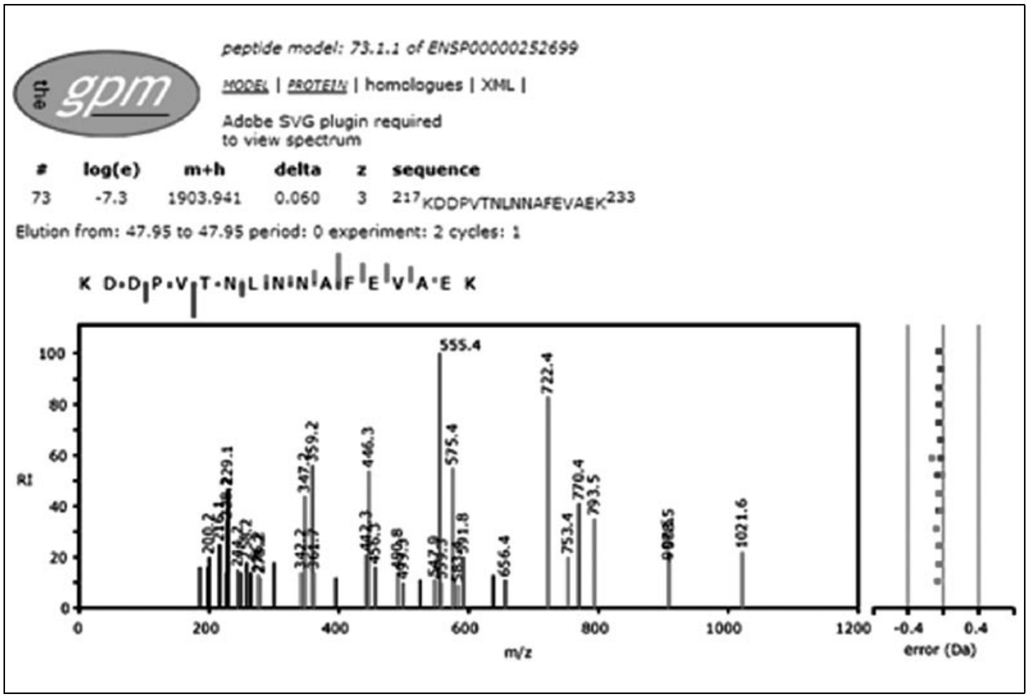
    b. log(e): the base-10 logarithm corresponding to the confidence of the assignment.

    c. m+h: the calculated mass of the singly charged peptide.

    d. delta: the mass difference between the calculated and measured mass of the assigned peptide.

    e. z: the charge state of the peptide ion.

    f. sequence: the assigned sequence (uppercase, blue on screen), with the sequence region N- and C-terminal included (lowercase, black on screen). The residue numbers correspond to the position of the peptide in the original sequence. Clicking on the uppercase (blue) sequence opens the peptide model page (Fig. 13.2.8), which allows for the detailed inspection of the spectrum assignment data quality.



peptide model: 73.1.1 of ENSP00000252699

MODEL | PROTEIN | homologues | XML |

Adobe SVG plugin required
to view spectrum

#	log(e)	m+h	delta	z	sequence
73	-7.3	1903.941	0.060	3	^{217}KDDPVTNLNNAFEVAEK^{233}

Elution from: 47.95 to 47.95 period: 0 experiment: 2 cycles: 1

Figure 13.2.8 The peptide model display page. This black and white facsimile of the figure is intended only as a placeholder; for full-color version of figure go to *http://www.interscience.wiley.com/c_p/colorfigures.htm*.

Peptide page (Fig. 13.2.8)

21. Links are provided on the top of the page illustrated in Figure 13.2.8 in a menu bar that have the same behaviors as those on the Homologues page (see step 15).

    The model number at the top of the page is in the format `xxx.yy.zz`*, where* `xxx` *is the position of the spectrum in the original spectrum listing file,* `yy` *is the homologue rank number and* `zz` *is the peptide assignment number.*

22. Details similar to the table on the Protein page (see step 20 and Fig. 13.2.7) are shown in a table at the top of the page. If available, a line describing the experimental details associated with the mass spectrum is also displayed.

23. The main display is a graphical representation of the quality of the mass spectrum assignment in three sections.
    a. At the top of the display is a fragmentation diagram, showing the relative strengths of signals corresponding to each bond in the peptide.
    b. A display of the MS/MS spectrum, with peaks assigned to b ions shown in blue and those assigned to y ions in red. Unassigned ions are shown in black and ions that can be assigned to trivial neutral loss reactions are shown in mauve.
    c. To the right of the spectrum display is a scatter plot, showing the difference between the calculated and observed mass for each of the assigned ions.

SUPPORT PROTOCOL

USING ADVANCED FEATURES FOR GPM SEARCHES

On the top left side of the simple search page described in Basic Protocol 2 and illustrated in Figure 13.2.3, there is a link ("advanced PAGE") to an advanced search page that allows the user access to more parameters that can be adjusted for a search session. The values of these additional parameters are fixed by the selection of a "method" in Basic Protocol 2 (see step 10 in that protocol); the values fixed in a "method" are appropriate for many common experiments, but may need to be varied for a specific circumstance. This list of parameters is useful if the experiment being performed involves the use of customized reagents or protocols. Parameters that were not discussed above are described below.

1. *isotope error:* If the radio button "no" is chosen, then the parent ion mass tolerance is simply applied and any peptide with a calculated mass outside of the allowed range is not scored. If "yes" is chosen, then multiple parent ion mass tolerance windows are used to attempt to detect errors made by the mass spectrum data system in assigning the correct all-^{12}C peak associated with the mass spectrum.

2. *protein N-terminus:* The mass value (in Da) entered into the text box here is added to the N-terminal residue of any peptide that includes the protein's N-terminus.

3. *protein C-terminus:* The mass value (in Da) entered into the text box here is added to the C-terminal residue of any peptide that includes the protein's C-terminus.

4. *cleavage C-terminal change:* When a peptide is cleaved by an enzyme or reagent, the C-terminus revealed by the cleavage is modified by the addition (or subtraction) of some moiety whose mass is entered into the text box here. For example, hydrolysis results in the addition of a hydroxyl group (~17 Da). Nonhydrolytic reactions will add different moieties.

5. *cleavage N-terminal change:* When a peptide is cleaved by an enzyme or reagent, the N-terminus revealed by the cleavage is modified by the addition (or subtraction) of some moiety whose mass is entered into the text box here. For example, hydrolysis results in the addition of a hydrogen (~1 Da). Nonhydrolytic reactions will add different moieties.

6. *cleavage site:* The cleavage chemistry used in the initial round of scoring is set using a simple expression, written into the text box here using the sequence pattern format originally used for PROSITE (a slight modification of the more general "regular expression" format). The expression is in the form of two subexpressions specifying the N-terminal and C-terminal residues adjacent to the cleaved bond, separated by a pipe (|) that represents the position of the bond to be cleaved. The subexpressions can be in one of two forms.

   a. `[ABC]` indicating that the residue in that position can be any of the set of residues with the square brackets.

   b. {`ABC`} indicating that the residue in that position can be any residue except the residues in the curly brackets.

   The special residue character "X" represents all residues, so that the expression `[X]|[X]` *represents cleavage at all residues. The expression* {`X`} *is not useful, as it means that bond cannot represent cleavage for any residue combination.*

   For example, the commonly used enzyme trypsin can be specified by the expression `[KR]|{P}`*, V8 proteinase can be specified by* `[ED]|[X]` *and the enzyme Asp-N can be specified by* `[X]|[D]`.

7. *missed sites:* The entry typed in the text box here represents the number of cleavable sites that may be present in a peptide.

   No protein-cleavage reagent will cleave all accessible peptide bonds at the same rate. Some bonds will be preferred, either because of sequence-specific effects or secondary/tertiary structural constraints. Therefore, in any protein digest mixture there will be peptides that contain sites that satisfy the cleavage specification mentioned in step 6. In order to speed up the calculation, it is often advantageous to limit the number of potentially missed sites to a small number.

   IMPORTANT NOTE: *in the case of cleavage at any bond (* `[X]|[X]` *) it is necessary to set this value to allow for many missed cleavages, e.g., 50, which will be sufficient to consider peptides up to a parent ion mass of up to ~5000 Da.*

8. *refine model:* If the radio button "no" is chosen in steps 1 to 7, then the sequence-to-spectrum comparison process ends after the parameter set in steps 1 to 7 has been considered. If "yes" is selected, then further rounds of comparison are performed, based on the following set of parameters. This analysis is performed only on protein sequences that contain at least one peptide that was assigned to a spectrum with an expectation value of less than the "valid expectation" value specified in below in point "f."

   a. refinement potential modifications (`mass@X`): This entry is a list of potential modifications, specified using the same format used for modifications on the simple search page. These modifications are used, along with a "missed sites" value of 5, to more thoroughly check for modified peptides with larger numbers of missed cleavage sites than specified in the first round.

   b. use these modifications throughout: If the radio button "no" is selected, then the partial modification list specified above is not used for further rounds of analysis, and the program reverts to the list specified in the first round. If "yes" is selected, then this list is carried forward through subsequent rounds.

   c. unanticipated cleaves: If the radio button "yes" is selected, then the sequences from the first round are re-examined with cleavage at all bonds (`[X]|[X]`) and missed sites = `50`.

   d. potential N-terminus modifications (`mass@[`): If this entry is not left blank, then the sequences from the first round are re-examined with cleavage at all bonds (`[X]|[X]`) and missed sites = `50`, with each of the masses represented by the entry applied to the N-terminus of each peptide.

   e. potential C-terminus modifications (`mass@]`): If this entry is not left blank, then the sequences from the first round are re-examined with cleavage at all bonds (`[X]|[X]`) and missed sites = `50`, with each of the masses represented by the entry applied to the C-terminus of each peptide.

   f. valid expectation (< `value`) This value is the maximum expectation value cutoff used to determine if a sequence should be used in the refinement rounds of analysis.

   It is best to keep this value near 1 to include sequences that may have only weak identifications in the first round, but which may identify more strongly following refinement (the default value is 0.1).

9. *spectrum synthesis:* If the radio button "no" is selected, then all peptide bonds are considered to be equally likely to generate fragment ions. If "yes" is selected, then a set of rules is applied that biases the score for a spectrum-peptide pair, based on the observed pairwise propensity of specific residue combinations to form fragment ions.

10. *noise suppression:* If the radio button "yes" is selected, then the following parameters are used to limit the number of spectra and fragment ions that are considered.

    a. minimum parent M+H: Mass spectra in an input list with parent ion masses less than the value in Da entered into the text box here will not be scored.

    b. minimum fragment m/z: Fragment-ion peaks with *m/z* values less than the value entered into the text box here will not be scored.

    c. total peaks: The number entered into the text box here represents the maximum number of peaks in a single spectrum that will be considered. The peaks are ordered from the most intense to the least intense and the total peak count is used to cut off the low intensity peaks, if necessary.

    d. minimum peaks: Any spectrum with fewer peaks than the number entered into the text box here will not be scored.

11. Press the “create models” button (immediately below the “taxon” list drop-down menu) to submit the data and begin the search.

USING PROWL WITH THE WEB INTERFACE FROM THE ROCKEFELLER SERVER WITH PEPTIDE FINGERPRINTING MASS SPECTROMETRY (MS) DATA: ProFound

BASIC PROTOCOL 3

ProFound makes it possible to input peptide mass fingerprinting spectra and match them to sequences. This tool allows the user to work with one spectrum at a time, comparing sequences to data (Zhang and Chait, 2000).

Necessary Resources

Hardware

Any Internet-connected computer

Software

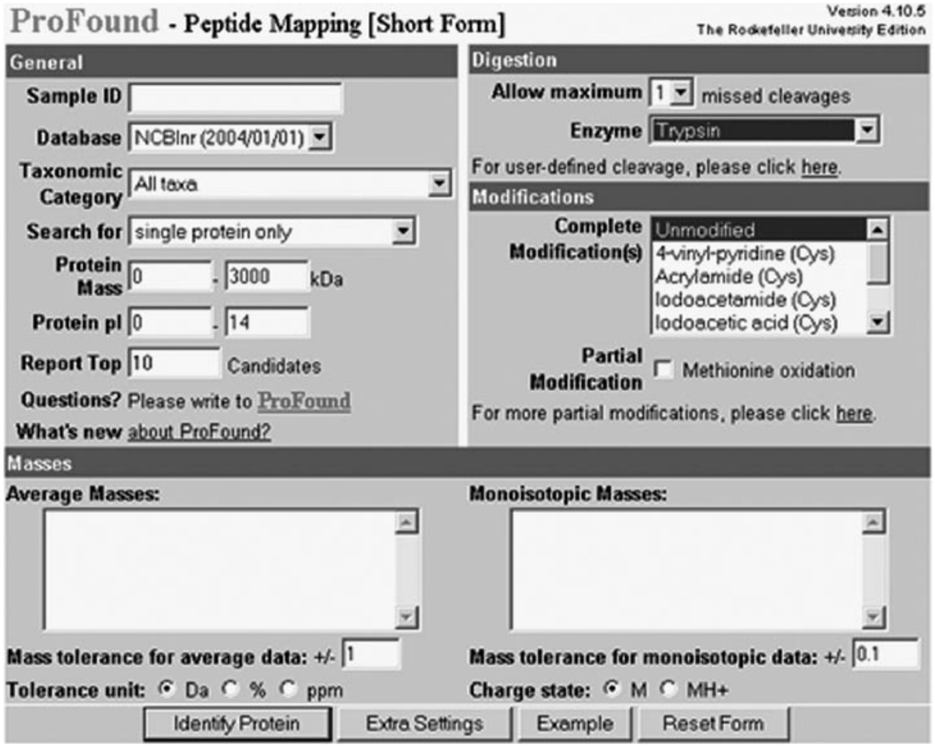
Web browser

Files

Text list of MS peak masses, with each mass on a single line

Enter general search requirements

1. Point the Web browser to *http://prowl.rockefeller.edu/profound_bin/WebProFound.exe*. The page illustrated in Figure 13.2.9 should appear. This page offers the user a demonstration data set (click on the “Example” button), which can be used as a tutorial for new users, as well as for testing/validating the system when new versions of the software and databases are installed.



ProFound - Peptide Mapping [Short Form]

Version 4.10.5
The Rockefeller University Edition

General

Sample ID
Database: NCBInr (2004/01/01)
Taxonomic Category: All taxa
Search for: single protein only
Protein Mass: 0 - 3000 kDa
Protein pI: 0 - 14
Report Top: 10 Candidates
Questions? Please write to ProFound
What's new about ProFound?

Digestion

Allow maximum 1 missed cleavages
Enzyme: Trypsin
For user-defined cleavage, please click here.

Modifications

Complete Modification(s): Unmodified, 4-vinyl-pyridine (Cys), Acrylamide (Cys), Iodoacetamide (Cys), Iodoacetic acid (Cys)
Partial Modification: □ Methionine oxidation
For more partial modifications, please click here.

Masses

Average Masses:
Monoisotopic Masses:
Mass tolerance for average data: +/- 1
Mass tolerance for monoisotopic data: +/- 0.1
Tolerance unit: ● Da ○ % ○ ppm
Charge state: ● M ○ MH+

Figure 13.2.9 The start page for a ProFound session.

2. *Optional:* Enter a Sample ID for a personal reference to the search.

3. Select the appropriate Database; also select the appropriate organism from the Taxonomic Category menu, reflecting the source of the original sample. The Databases available (see discussion of the usage of this term in Basic Protocol 1) are the NCBI FASTA versions of protein sequences from SWISS-PROT and NR.

4. Under "Search for," select how complex the protein mixture of interest is suspected to be.

   It is usually best to start with "single protein only." An alternative way of searching protein mixtures iteratively is described below.

5. Enter the protein intact mass and pI ranges to be searched.

   Proteins with intact masses and pI outside of these ranges will be excluded from consideration during the search. It is good practice not to limit these parameters too strictly, because the mass and pI are calculated on the full sequence as it is written in the sequence list. If the peptide of interest was derived from a proteolytic fragment of a larger protein, or if the mature form of the protein is significantly different from the sequence in the list, it may be missed.

6. In the line labeled "Report Top," select the maximum number of protein candidates to return for a search.

Specify digestion conditions and chemical modifications

7. From the Enzyme menu on the right side of the page, select an appropriate enzymatic cleavage chemistry for the particular protocol. From the menu on the line above the Enzyme menu, enter the number of missed cleavage sites to consider. An enzymatic cleavage chemistry can be user-defined by following the hyperlink provided.

   The selections available are for proteolytic enzymes commonly used in proteomics. Trypsin is by far the most widely used enzyme. Commonly, one missed cleavage site is allowed in the proteolytic peptides.

8. In the Modifications section of the page, select appropriate chemical modifications used in the experimental protocol from the list. User-defined chemical modifications can be defined by following the hyperlink provided.

Enter the mass and charge data (Masses section of the ProFound page)

9. Cut and paste the list of *m*/*z* values into the Average Masses and Monoisotopic Masses boxes.

10. In the "Mass tolerance for average data box," enter the *m*/*z* tolerance for average and monoisotopic masses.

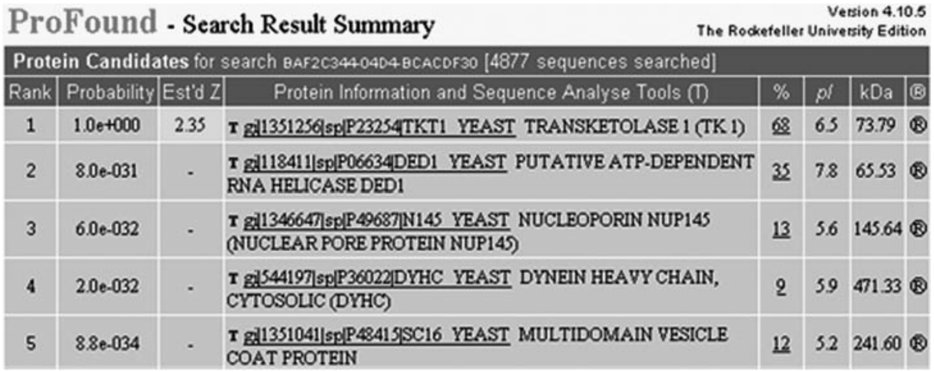
11. Enter the units for the *m*/z tolerance for average and monoisotopic masses (Da, %, or ppm).

12. Select the charge state (M, mass of the neutral molecule, or MH$^+$, mass of the singly protonated molecule).

13. Press the Identify Protein button to submit the search.

Interpreting the results of a ProFound search

The results returned from ProFound are illustrated in Figure 13.2.10. The result format is a simple HTML page that presents the best possible candidate sequences for a given set of input *m*/*z* values and search parameters.



ProFound - Search Result Summary

Version 4.10.5
The Rockefeller University Edition

Protein Candidates for search BAF2C344-04D4-BCACDF30 [4877 sequences searched]

Rank	Probability	Est'd Z	Protein Information and Sequence Analyse Tools (T)	%	pI	kDa	®
1	1.0e+000	2.35	T gi|1351256|sp|P23254|TKT1_YEAST TRANSKETOLASE 1 (TK 1)	68	6.5	73.79	®
2	8.0e-031	-	T gi|118411|sp|P06634|DED1_YEAST PUTATIVE ATP-DEPENDENT RNA HELICASE DED1	35	7.8	65.53	®
3	6.0e-032	-	T gi|1346647|sp|P49687|N145_YEAST NUCLEOPORIN NUP145 (NUCLEAR PORE PROTEIN NUP145)	13	5.6	145.64	®
4	2.0e-032	-	T gi|544197|sp|P36022|DYHC_YEAST DYNEIN HEAVY CHAIN, CYTOSOLIC (DYHC)	9	5.9	471.33	®
5	8.8e-034	-	T gi|1351041|sp|P48415|SC16_YEAST MULTIDOMAIN VESICLE COAT PROTEIN	12	5.2	241.60	®

Figure 13.2.10 Results obtained from a ProFound search.

14. A list of protein candidates is shown in the results, with the protein candidate best matching the data shown on top.

15. The following information is shown for each candidate protein (online help is available by clicking on the text in the table heading):

    a. Rank.

    b. Probability: the Bayesian probability calculated by ProFound.

    c. Est'd Z: The estimated Z-score is as an indicator of the quality of the search result.

    d. Protein Information and Sequence Analyse Tools: The protein identifier and a short description of the function of the protein.

    e. %: The percentage coverage.

    f. pI: The calculated pI of the protein.

    g. kDa: The calculated mass of the protein in kilodaltons.

    The probability for a protein candidate being a false positive can be calculated from the estimated Z-score. Z-scores of 2.33 and 1.64 correspond to a false-positive rate of 1% and 5%, respectively.

16. More information can be viewed for each candidate protein by following the following hyperlinks:

    a. Clicking on the "T" in front of the protein identifier leads to the ProteinInfo helper tool.

    b. Clicking on the protein identifier leads to the Entrez Web site (also see *UNIT 1.3*) for more information on the protein.

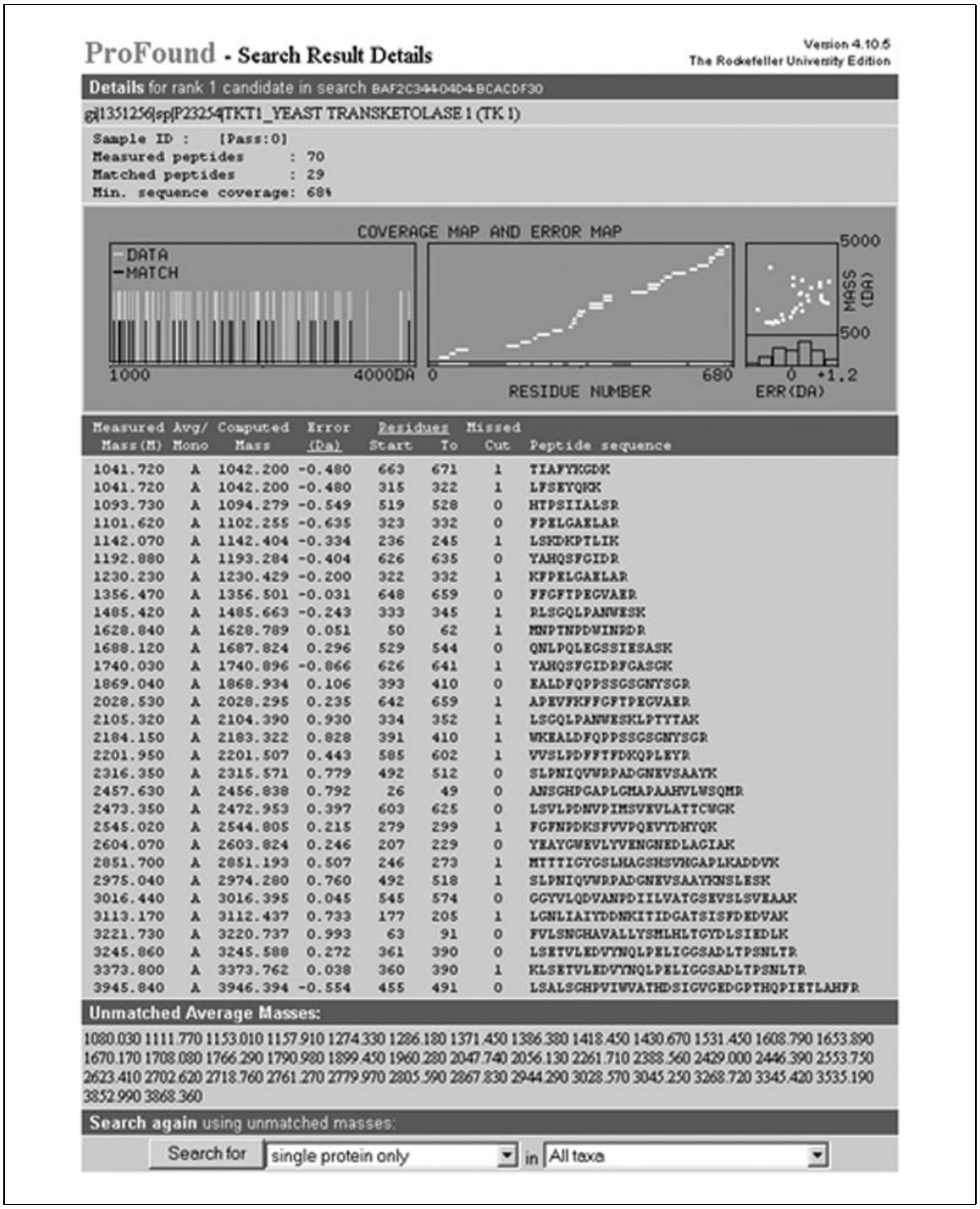
    c. The details of the match between the data and the protein sequence can be viewed by clicking on the number showing the % coverage.

17. The details of the match between the data and a protein sequence are summarized in three graphs and a table (Fig. 13.2.11).

18. The first graph illustrates what fraction of the data is explained by the protein sequence.

19. The second graph shows what parts of the protein sequence are observed in the experiment (coverage map).

20. The third graph shows the difference between the measured and the calculated masses as a function of peptide mass (error map).



ProFound - Search Result Details

Version 4.10.5
The Rockefeller University Edition

Details for rank 1 candidate in search BAF2C3440404BCACDF30

gi|1351256|sp|P23254|TKT1_YEAST TRANSKETOLASE 1 (TK 1)

```
Sample ID :    [Pass:0]
Measured peptides      : 70
Matched peptides       : 29
Min. sequence coverage: 68%
```

Measured	Avg/	Computed	Error	Residues		Missed	
Mass(M)	Mono	Mass	(Da)	Start	To	Cut	Peptide sequence
1041.720	A	1042.200	-0.480	663	671	1	TIAFYKGDK
1041.720	A	1042.200	-0.480	315	322	1	LFSEYQKK
1093.730	A	1094.279	-0.549	519	528	0	HTPSIIALSR
1101.620	A	1102.255	-0.635	323	332	0	FPELGAELAR
1142.070	A	1142.404	-0.334	236	245	1	LSKDKPTLIK
1192.880	A	1193.284	-0.404	626	635	0	YAHQSFGIDR
1230.230	A	1230.429	-0.200	322	332	1	KFPELGAELAR
1356.470	A	1356.501	-0.031	648	659	0	FFGFTPEGVAER
1485.420	A	1485.663	-0.243	333	345	1	RLSGQLPANWESK
1628.840	A	1628.789	0.051	50	62	1	MNPTNPDWINRDR
1688.120	A	1687.824	0.296	529	544	0	QNLPQLEGSSIESASK
1740.030	A	1740.896	-0.866	626	641	1	YAHQSFGIDRFGASGK
1869.040	A	1868.934	0.106	393	410	0	EALDFQPPSSGSGNYSGR
2028.530	A	2028.295	0.235	642	659	1	APEVFKFFGFTPEGVAER
2105.320	A	2104.390	0.930	334	352	1	LSGQLPANWESKLPTYTAK
2184.150	A	2183.322	0.828	391	410	1	WKEALDFQPPSSGSGNYSGR
2201.950	A	2201.507	0.443	585	602	1	VVSLPDFFTFDKQPLEYR
2316.350	A	2315.571	0.779	492	512	0	SLPNIQVWRPADGNEVSAAYK
2457.630	A	2456.838	0.792	26	49	0	ANSGHPGAPLGMAPAAHVLWSQMR
2473.350	A	2472.953	0.397	603	625	0	LSVLPDNVPIMSVEVLATTCWGK
2545.020	A	2544.805	0.215	279	299	1	FGFNPDKSFVVPQEVYDHYQK
2604.070	A	2603.824	0.246	207	229	0	YEAYGWEVLYVENGNEDLAGIAK
2851.700	A	2851.193	0.507	246	273	1	MTTTIGYGSLHAGSHSVHGAPLKADDVK
2975.040	A	2974.280	0.760	492	518	1	SLPNIQVWRPADGNEVSAAYKNSLESK
3016.440	A	3016.395	0.045	545	574	0	GGYVLQDVANPDIILVATGSEVSLSVEAAK
3113.170	A	3112.437	0.733	177	205	1	LGNLIAIYDDNKITIDGATSISFDEDVAK
3221.730	A	3220.737	0.993	63	91	0	FVLSNGHAVALLYSMLHLTGYDLSIEDLK
3245.860	A	3245.588	0.272	361	390	0	LSETVLEDVYNQLPELIGGSADLTPSNLTR
3373.800	A	3373.762	0.038	360	390	1	KLSETVLEDVYNQLPELIGGSADLTPSNLTR
3945.840	A	3946.394	-0.554	455	491	0	LSALSGHPVIWVATHDSIGVGEDGPTHQPIETLAHFR

Unmatched Average Masses:

1080.030 1111.770 1153.010 1157.910 1274.330 1286.180 1371.450 1386.380 1418.450 1430.670 1531.450 1608.790 1653.890 1670.170 1708.080 1766.290 1790.980 1899.450 1960.280 2047.740 2056.130 2261.710 2388.560 2429.000 2446.390 2553.750 2623.410 2702.620 2718.760 2761.270 2779.970 2805.590 2867.830 2944.290 3028.570 3045.250 3268.720 3345.420 3535.190 3852.990 3868.360

Search again using unmatched masses:

Figure 13.2.11 Details of a match between the data and a protein sequence.

21. The table lists information about the matching peptides:
    a. Measured Mass (M): Measured mass.
    b. Avg/Mono: Indicates if the mass is average (A) or monoisotopic (M).
    c. Computed Mass: Calculated mass.
    d. Error (Da): The difference between calculated and measured mass in daltons.
    e. Residues: The location of the peptide in the protein sequence, Start and To (end) residue.
    f. Missed: The number of missed cleavage sites in the peptide.
    g. Peptide sequence.

GUIDELINES FOR UNDERSTANDING RESULTS

The final steps in each protocol discuss data interpretation.

MS/MS data do not mean a peptide was sequenced

There has been a tendency to somewhat overstate the value of MS/MS data by drawing an analogy to peptide sequencing as obtained by an Edman sequencer. Peptides are "identified" by mass spectrometry on the basis of a set of gas-phase fragments that do not completely define the sequence of the peptide (Yates et al., 1995). Instead, the set of fragments are sufficient information to allow the statistically significant assignment of a particular peptide sequence, compared to all of the other possible peptide sequences being considered (Mann and Wilm, 1994). Peptides that fragment so as to give unequivocal evidence for a particular amino acid sequence are relatively rare (Nielsen et al., 2002).

MS data do not contain an amino acid analysis

Peptide fingerprint identifications, such as those provided by ProFound, are in some ways analogous to identifying a sequence by very accurate amino acid analysis (AAA). By knowing the amino acid composition of a tryptic peptide, it is often possible to eliminate all but a few candidate peptides from even very large lists of protein sequences (Parker, 2002; Eriksson and Fenyö, 2004). Obtaining an accurate molecular mass for a peptide is somewhat similar; however the number of potential residue compositions that can have a particular mass can be quite large. Multiple isobaric combinations of amino acids (e.g., M(GA) = M(Q)) and the fact that leucine and isoleucine cannot be differentiated by simple mass measurement means that any peptide of more than a few residues can have more than one composition corresponding even to an exact mass measurement. The mass spectrometry data make up for this increased ambiguity (when compared to AAA) by providing information about many peptides from the same protein simultaneously, while AAA is a sequential, relatively slow process.

Interpreting the statistical evidence in context

The goal of any real experiment that involves the use of PROWL to discover protein sequences is not simply to name proteins. Instead it is to produce information that can be rationally used in the investigation of biological processes. Therefore, it is very important to interpret the results of the bioinformatics software analysis within the framework of what is known about a biological system. For example, keratin is often identified, and it is regularly dismissed as an experimental artifact caused by contamination from dust and dander. Automated systems may use peak "exclusion" lists that make identifying keratin impossible. However, keratin is expressed in almost all eukaryotic cells as an important element in the cytoskeleton, and its detection may be crucial evidence in a particular experiment. Understanding the types of exclusion lists and filters that have been used to process the data is therefore a necessary part of interpreting the results of bioinformatics results.

The identifications produced by PROWL are all statistical in nature; as mentioned above, they are not the exact equivalent of detailed peptide sequencing. This fact makes the use of the results similar to the use of sequence homology information generated by utilities such as BLAST (Karlin and Altschul, 1990, 1993). One can fairly accurately estimate the likelihood that a particular assignment has happened at random, but one cannot say that an assignment is unambiguously "true." Therefore, it is necessary to clearly indicate the confidence of any assignment when reporting that result either to a colleague or for published research. It is often the case that sequence assignments with fairly low confidence can be of great assistance to a well informed biological researcher who understands the significance of a set of correlated results in context. Important correlations resulting from poor evidence can be tested with more focused informatics analysis of existing data or improved experimental designs.

COMMENTARY

Background Information

PROWL is a collection of tools for the identification of protein sequences, using input data derived from mass spectrometry. It was designed and initially made available in 1996, as a collaboration between the protein mass spectrometry research groups at Rockefeller University and the Skirball Institute at New York University Medical Center (Fenyö et al., 1996, 1998). It was clear at the time that one of the most important applications of mass spectrometry to protein biochemistry was going to be for sensitive determination of which mature gene products are present in a mixture of proteins derived from a biological source. The use of mass spectrometry for the complete sequencing of proteins and peptides had a long history, but practical difficulties in obtaining complete, unambiguous sequences had limited the general applicability of the technique. Instead, Edman degradation chemistry had been the gold standard since its development. The availability of reasonably complete genomic and cDNA sequence databases in the middle 1990s reduced the necessity for the complete sequencing of a protein: the task switched to determining enough sequence to directly match the protein with a known translated nucleotide sequence or to create a primer that could be used to amplify the mRNA coding the protein for subsequent nucleotide sequencing.

The advancements in the use of mass spectrometry to identify proteins were made possible by discoveries made between 1987 and 1989 that produced a new generation of ion sources for peptides: electrospray ionization (ESI, Fenn et al., 1989) and matrix-assisted laser desorption/ionization (MALDI, Hillenkamp et al., 1991). These ion sources were more than 1000× more sensitive than the ion sources that they replaced. With this new sensitivity, it was possible to routinely examine sample amounts of <1 picomole, which is a practical level for the preparation of biologically derived protein samples (Aebersold and Goodlett, 2001). By combining the new ion sources with digestion and separation protocols that had been developed for handling Edman sequencing samples, it was possible to derive enough information from mass spectra to correlate a protein sequence with a known translated nucleotide sequence (Mann and Pandey, 2001).

PROWL is designed to use protein sequence information to identify potential matching sequences; experimental data from various types of mass spectrometers can be input directly into PROWL's component software. For comparison with familiar genomic tools, it would be similar to directly inputting a sequencer's untranslated trace data into BLAST. This is not a perfect analogy, however, as mass spectrometric data rarely contain enough information to unambiguously sequence a peptide. Instead, PROWL utilizes a set of sophisticated statistical techniques to draw reasonable correlations between a protein's sequence and often noisy, incomplete experimental data.

While PROWL was one of the first sites that made tools for this purpose available, it has become part of a group of academic and commercial Web sites that offer a range of tools aimed at the same applications, utilizing different approaches and software. Most notable among these are the ExPasy, Mascot (Perkins et al., 1999), and Protein Prospector suites. Many of the freely available sites also have commercial versions with more flexibility. These include Mascot (Matrix Science), Spectrum Mill (Agilent), Protein Prospector (ABI, licensed from UCSF), and Knexus (the Proteometrics version of PROWL, currently licensed to Harvard Biosciences; Field et al., 2002).

The use of convenient, form-based search engines for retrieving information using the Internet has become such a commonplace way of obtaining detailed information that it scarcely needs comment. PROWL was first made available when the notion of using a remote resource for this type of search was still rather novel. Therefore, the goal of the original PROWL resources, such as PepFrag (Basic Protocol 1), was to obtain sequence information as an end in itself. As the Internet (and various genome projects) has evolved, sequence information has been supplanted by an ever-expanding network of information that is linked together by a set of unique accession numbers. ProFound (Basic Protocol 3) and the GPM (Basic Protocol 2) now mainly fill the role of generating lists of the relevant accession numbers and providing rapid "click-through" access to the additional bioinformatics information available on the proteomic, genomic, and transcriptomic levels.

Critical Parameters

Difficulties with "non-redundant databases"

At the moment, most protein identification work is performed using the collection of protein sequences compiled by the U.S. National Center for BioInformatics (NCBI), called nr, short for "non-redundant." This list of sequences is compiled by comparing a new sequence with all of the sequences currently contained in the collection. If a sequence is exactly the same as one currently available, it is given an accession number, but the protein sequence is not added to the list: the accession number is added as an additional description item for the sequence. If a sequence is exactly the same as an existing sequence, except for as little as one residue, then it is assigned an accession number and the sequence is added to the collection. Therefore, when using this collection for protein identification, large lists of protein sequences are often retrieved, all of which contain the same peptide. This problem can be compounded by the existence of similar gene families in an organism and the fact that the sequences of some proteins are very highly homologous across broad taxonomic families. In fact, some of these highly homologous protein sequences are over-represented in nr (e.g., myoglobin and cytochrome *c*) because these sequences have been used in studies meant to understand the basis of this type of sequence homology. When presented with a long list of potentially homologous sequences generated from nr, it is important to realize that the protein present in one's original sample may not be the description at the top of the list and that multiple genes, resulting in very similar sequences, may have been active. Splice variants are an additional complication, making the detailed review of the evidence supporting a particular identification necessary to ensure that results are not reported with undue confidence.

Much of the confusion associated with using nr can be alleviated if the organism that is the source of the proteins has a known genome. By limiting the search to the proteins predicted for a particular genome, it is often possible to greatly reduce the complexity of the resulting list of sequences that correspond to the measured data. It is also much clearer when a measurement is sufficient to distinguish between the members of multigene families of similar sequences. Designing a search strategy to limit the complexity of the results is part of the overall design of an experimental protocol, and should be considered as early on in the planning process as possible.

Instrumental mass accuracy can be deceptive

Part of the specification for a particular mass spectrometer's performance is its mass accuracy, often given in parts per million (ppm). If a mass spectrometer can measure a 2000 Da parent ion with an accuracy of ±10 ppm, then the mass error for the measurement would be $2000 \times 0.000010 = \pm 0.020$ Da. Using this type of calculation to set the parent ion mass measurement tolerance for an MS/MS identification can lead to errors, however. Identifications are made based on the assumption that the all-^{12}C peak (A_0) in the isotope cluster corresponding to a peptide has been correctly determined. Unfortunately, most commonly available software will often assign the most abundant peak in the isotope cluster, rather than the A_0 peak. This effect can result in a systematic mass-assignment error of 1 or 2 Da, even though the accuracy of the instrument is considerably less than 1 Da. The result of being overconfident in the true accuracy of parent ion mass assignments is often to completely miss a good identification or to misinterpret the spectrum in such a way as to imply that one or two chemical deamidations of asparagines or glutamine residues have occurred, which will compensate for the systematic error in mass assignment. Reasonable care should be taken to ensure that the correct peak has been assigned (when possible), and any software settings should take into account the possibility of this type of error. It is important to note that very accurate determination of a parent ion mass does not necessarily help MS/MS peptide identifications, as it is the pattern of fragment ion masses that is used to ensure a good identification. Relaxing the parent ion mass tolerance to be as much as ±2 Da, even when high-accuracy instruments are used, is a good way to ensure that correct identifications are not missed.

Literature Cited

Aebersold, R. and Goodlett, D.R. 2001. Mass spectrometry in proteomics. *Chem. Rev.* 101:269-295.

Craig., R. and Beavis, R.C. 2003. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* 17:2310-2316.

Eriksson, J. and Fenyö, D. 2004. Probity: A protein identification algorithm with accurate assignment of the statistical significance of the results. *J. Proteome Res.* 3:32-36.

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., and Whitehouse, C.M. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246:64-71.

Fenyö, D., Zhang, W., Chait, B.T., and Beavis, R.C. 1996. Internet-based analytical chemistry resources: A model project. *Anal. Chem.* 68:721A-726A

Fenyö, D., Qin, J., and Chait, B.T. 1998. Protein identification using mass spectrometric information. *Electrophoresis* 19:998-1005.

Field, H.I., Fenyö, D., and Bevies, R.C. 2002. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimizes protein identification and archives data in a relational database. *Proteomics* 2:36-47.

Hillenkamp, F., Karas, M., Beavis, R.C., and Chait, B.T. 1991. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem.* 63:1193A-1203A.

Karlin, S. and Altschul, S. 1990. Methods for assessing the statistical significance of molecular sequence features using general scoring schemes. *Proc. Natl. Acad. Sci. U.S.A.* 87:2264-2268.

Karlin, S. and Altschul, S. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. U.S.A.* 90:5873-5877.

Mann, M. and Pandey, A. 2001. Use of mass spectrometry–derived data to annotate nucleotide and protein sequence databases. *Trends Biochem. Sci.* 26:54-61.

Mann, M. and Wilm, M. 1994. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66:4390-4399.

Nielsen, M.L., Bennett, K.L., Larsen, B., Moniatte, M., and Mann, M. 2002. Peptide end sequencing by orthogonal MALDI tandem mass spectrometry. *J. Proteome Res.* 1:63-71.

Parker, K.C. 2002. Scoring methods in MALDI peptide mass fingerprinting: ChemScore and the ChemApplex program. *J. Am. Soc. Mass Spectrom.* 13:22-39.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20:3551-3567.

Yates, J.R. III, Eng, J.K., McCormack, A.L., and Schietz, D. 1995. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* 67:1426-1436.

Zhang, W. and Chait, B.T. 2000. ProFound: An expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* 72:2482-2489.

Contributed by Ronald Beavis
Beavis Informatics Ltd.
Winnipeg, Manitoba, Canada

David Fenyö
GE Healthcare
Piscataway, New Jersey