



Phil Jones (V2.01, October 2007)

Primers for Predocs: Proteins and Proteomics

19 February, 2008

OLS – The Ontology Lookup Service

www.ebi.ac.uk/ontology-lookup

The Ontology Lookup Service began life as a SOAP web service intended specifically to provide ontology support for PRIDE. It quickly took on a life of its own, acquiring a web application user interface to allow users to browse over 45 ontologies, including all of the ontologies from OBO, the Open Biomedical Ontology web site, the NEWT taxonomy and the PRIDE controlled vocabulary (covering everything that we could not find defined elsewhere!).

This brief set of activities will help to acquaint you with OLS and hopefully give you some idea of how it can assist you in your research and data handling activities.

✍ Navigate to <http://www.ebi.ac.uk/ontology-lookup>

You should now see the home page of OLS. Next you can explore the available ontologies and experience the search capabilities that make use of AJAX technology.

✍ Click on the down-pointing arrow on the select option list labeled ‘Search Ontology’

You will be presented with the full list of available ontologies from which you can search. Note that if you select one of these, your search will be restricted to that ontology. Otherwise, you can select the option ‘Search in All Ontologies’ at the top of the option list.

✍ Select ‘Gene Ontology [GO]’ from the list of available ontologies.

✍ Click on the ‘Term Name’ text box to select it and then start to enter the name of a sub-cellular component of your choice (e.g. mitochondria, golgi apparatus, chloroplast, endoplasmic reticulum)

Note that matching suggestions are made as you type. This search is not case sensitive and looks for a match in any part of the result text. Synonyms, where they are listed in the ontology, are also included.

✍ Click on the suggested match that you think is closest to the term you intended to find.

Selecting a term results in the definition and other details of the term being retrieved from the ontology lookup service and displayed.

You can now explore the ontology browsing mechanisms available from OLS.

✍ Click on ‘Browse’ button at the top of the page.



This work is licensed under the Creative Commons Attribution-Share Alike 3.0 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/3.0/> or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.

You will now be taken to the OLS “GO Ontology Browser” view. On the left you will see a hierarchical ‘tree’ view of the ontology, rooted at the term you selected in the previous step.

Click on the ‘browse full ontology’ link at the top of the page.

The hierarchical view will now be refreshed and should show you the three GO roots ‘biological_process’, ‘molecular_function’ and ‘cellular_component’.

✍ Click on the “cross in a square” icon to the left of the terms on the hierarchical view to expand and see the child terms as follows:

- cellular_component
- organelle
- intracellular organelle
- cytoplasmic vesicle
- cytoplasmic membrane-bound vesicle
- golgi-associated vesicle
- COPI-coated vesicle
- Golgi to ER transport vesicle

✍ Hover your mouse pointer over the various terms on the hierarchical view.

You will notice that the content of the “Relations” box on the right side of the screen will be updated to show the relationship between a term and its parent (e.g. organelle is_a cellular_component).

✍ Click on the term “Golgi to ER transport vesicle”.

You will notice that the content of the “Term information” box will be updated with the GO term id and an external link to the main GO website entry for that term. Not all ontologies provide an external website and such links are only shown where available. The contents of the “Associated information” box will also have been updated to provide any metadata available for this term from the OLS.

✍ Look at the bottom-right of the window where the entire map of terms from the root to “Golgi to ER transport vesicle” will be displayed.

You should now have seen just a few of the various ways in which you can examine the contents of multiple ontologies using OLS. In addition to the web application interface you can also examine the available SOAP interface by examining the contents of the following page:

✍ Navigate to the following URL to examine the OLS SOAP WSDL (SOAP interface description):

✍ <http://www.ebi.ac.uk/ontology-lookup/WSDLDocumentation.do>

✍ (You can also navigate to this from the OLS home page, under the ‘Documentation’ heading.)

This page will give you an idea of the scope of the web service, including all of the methods that you can call and questions that you can ask the service. You will see that this is a more extensive service than that provided through the human interface. We welcome suggestions of additional queries that you would like to pose to the service.

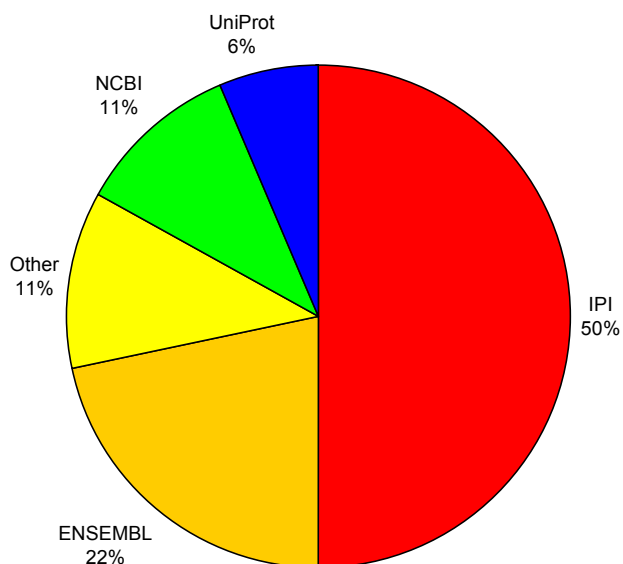
PICR – The Protein Identifier Cross-Reference Service

<http://www.ebi.ac.uk/Tools/picr/>

PICR is a new service offered by the EBI with the intent of solving a problem that has dogged proteomics – the use of multiple disparate protein sequence databases with their own protein accession systems. Two common tasks that are more difficult than they should be are data integration, where data sets coming from multiple sources need to be compared, and identifier translation, where identifiers from one source need to be converted to those of another source to use specific tools or services.

This situation is also very problematic for databases that accept submissions of data that use several different accession systems. The PRIDE database for example contains protein identifications that have been submitted using accessions and IDs from IPI, Ensembl, NCBI (gi and RefSeq identifiers), UniProt and various proprietary or species-specific protein sequence databases.

Data Sources for PRIDE Identifications



This makes it very difficult to query a database like PRIDE – for example a query with a specific IPI accession will return any matches against that accession, but will ignore any identifications of the same protein using a different coordinate system.

This problem led to the development of PICR, the Protein Identifier Cross-Reference Service.

PICR offers both a powerful ‘human’ (web based) user interface as well as comprehensive programmatic ‘web service’ interfaces to allow use of the PICR service to be built into external software systems.

Save the file http://www.ebi.ac.uk/~pjones/camb2007/twenty_random_accessions.txt to your local machine (Use File...Save As on the browser to save this file, e.g. to your desktop.)

- ✍ Take a look at the file you have downloaded – you should see that it contains a random mix of 20 accessions from UniProt, IPI, RefSeq and some Genbank Identifiers (GI numbers).
- ✍ Navigate to PICR at the URL given above.
- ✍ You can either paste the list into the ‘Input data’ box, or just navigate to the file you have saved locally using the ‘Browse...’ button below the ‘Input Data’ box.
- ✍ Initially, leave all other settings as default and click on ‘Search’.

You should see that the majority of the accessions have been mapped successfully to UniProt. (Separated into Swiss-Prot and TrEMBL on this view). You will probably see a couple of input accessions however that have not mapped successfully. If you read the feedback text however, you should see that this is because the search performed was too narrow.

- ✍ Click on the ‘PICR’ logo in the top left hand corner, or the ‘click here to start another search’ link at the bottom of the page.
- ✍ Now examine the search settings on the form and ensure that you know how to change the following default settings:
 - Return only active mappings, or include obsolete mappings.
 - Changing the mapping databases included in the results.
 - ‘All species’ -> Species specific search
 - ‘Simple HTML’ format -> ‘Detailed HTML’, ‘CSV’, ‘XLS’

Note that ‘CSV’ is ‘comma separated values’. This is a simple plain-text format, using commas to separate the values in columns. This is useful for parsing using Perl for example. It could also be used to import into most spreadsheet software too. ‘XLS’ is Microsoft Excel format. This format includes the colour coding present in the HTML formats.

- ✍ Try performing the same search with the settings changed. It is possible to successfully return mappings for all 20 accessions in the file – see if you can achieve this.

By default, PICR only returns mappings to active database entries, though many more might be available. PICR queries the UniProt Archive (UniParc), which is a historical archive of all known protein entries for over 60 protein sequence databases. As entries are deleted or deprecated from the source databases, they are never deleted from UniParc but are marked as ‘inactive’. PICR can include these inactive mappings in the results if the ‘Return only active mappings’ box is unchecked in the search options. These **inactive mappings will be shown in red** in both HTML views. Entries that have been mapped to SwissProt or TREMBL might also have **secondary Uniprot Knowledge Base identifiers, which will be shown in green**.

Entries that can map to an **active SwissProt or TREMBL may also have additional mappings, which will be shown in blue**. These mappings are obtained from the Uniprot Knowledge Base and, while valid, might not have 100% sequence identity to the submitted accession.

Where possible, active mappings will be hyperlinked to the primary source database entries to provide additional information. These links will only be available in the HTML views.

The choices in the 'Limit by species' pull down menu on the main search page contains the most commonly known and abundantly annotated taxons in the Uniprot Archive (UniParc) but is by no means comprehensive. There are over 140,000 distinct taxonomy identifiers in UniParc. If a species is not present in the pull-down menu, you can type its name in the provided text field. This will query the Ontology Lookup Service (OLS) and will provide suitable matches, if available. If taxons are entered in both the pull down menu and in the text field, only the value in the text field will be used.

Querying with taxonomy restrictions was designed to be pessimistic. While taxonomy annotation coverage is improving in UniParc, many databases do not provide taxonomy information. Mappings that are not annotated with taxonomy information or are not an exact match to the query parameter will not be included in the search results.

Further Reading

Côté, R.G., Jones, P., Apweiler, R., Hermjakob, H. (2006) **"The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries."** BMC Bioinformatics, 7, 97.

Côté, R.G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R., Hermjakob, H. (2007) **"The protein identifier cross-reference (picr) service: reconciling protein identifiers across multiple source databases."** BMC Bioinformatics, 8, 401.

The new 'beta' UniProt web-site also includes a mapping service, currently available from: <http://beta.uniprot.org/?tab=mapping>

PRIDE – The Proteomics Identifications Database www.ebi.ac.uk/pride

The PRIDE database is a centralized, standards compliant repository of identifications of proteins, peptides and protein modifications together with supporting evidence from mass spectrometry.

This tutorial will introduce you to how you can use PRIDE in your research and training activities and how you can perform powerful queries using the PRIDE database.

Exploring the PRIDE User Interface

You have a protein sequence upon which you have performed a trypsin digest followed by a mass spectrometry experiment. This experiment has yielded a single peptide, with the sequence SLTNSETEMVQQTQSLR.

PRIDE allows searches on a variety of areas, including peptide fragments, which it holds in its database.

- 📄 Go to <http://www.ebi.ac.uk/pride> and follow the link to Advanced Search in the left hand column. Select the "Peptide Sequence" field and

copy in the trypsin digest result above. Hit the "Submit Query" button to start the search.

This peptide is in the current PRIDE dataset and a summary form appears as a result of the search. The peptide only appears in one experiment, number 1642.

- ✎ Alter the "Select a Format" table to enable both the identifications and the spectra to be viewed in the browser.

The default position is to only show the identifications (i.e. the peptides resulting from the experiment). However some experiments also have the experimental spectra attached to them, and in order not to miss any potential information, this option should be selected.

- ✎ Scroll down to the list of experiments (a very short list in this case!) and hit the red "Download" button to retrieve the experimental information.

The resulting page details the supporting evidence for the experiment with this protein. The initial section of the page indicates the type of experiment, the procedure followed and links to relevant experimental data further down the page.

Protein Identifications

Peptide identifications arising from mass spectrometry can then be used to predict the presence of specific proteins using various algorithms.

The peptide fragment in your search has resulted in a database hit for the SwissProt protein Q9C035 using the MASCOT search engine from Matrix Science¹. MASCOT compares the results of the experimental analysis with an prediction of the protein fragments deriving from a tryptic digest of the protein sequences from a primary database and generates a score based on the similarity.

Additionally, MASCOT also calculates a threshold score to allow false positive identifications to be spotted. The results are displayed in the PRIDE entry as score and threshold values. If the score value is greater than the threshold value, the peptide sequence retrieved by MASCOT is statistically significant at a given confidence level (typically 95%). The larger the difference between score and threshold, the more confidently the sequence has been assigned. The confidence level for this score is given in the 'additional information' section.

In this case we have a score of 44 and a threshold score of 23.

The expectancy (e) of random matching can be calculated using:

$$e = \alpha \times 10^{\frac{T-S}{10}}$$

Where α is the confidence interval (0.05 in this case), T the threshold value and S the score. The expectancy of this being a random match is therefore:

$$e \approx 0.0004$$

This is well below the cut-off of 0.05.

¹ <http://www.matrixscience.com> Electrophoresis, 1999 Vol 20(18) pp3551-67

Identified Peptide Lists

In this case there is only one peptide shown as we searched for a particular peptide sequence.

PRIDE is therefore only showing us the relevant information from the experiment, omitting all other identified peptides and their corresponding proteins (of which there are several thousand in this experiment). It is a 16 residue peptide situated between amino acids 213-229 in the Q9C035 protein. Whilst the high score above would suggest a good match, the prediction has been derived from a single peptide fragment.

In addition, peptide length is only 16 residues and whilst there is currently no information on how long the SwissProt entry is, it is likely to be greater than this. These results would suggest that we should treat the analysis and subsequent inferred protein with caution and try to corroborate the findings with information from other sources.

A read-through of the paper abstract relating to this analysis informs us that this particular experiment relies on a peptide-isolation strategy that only picks up Met-containing peptides, thereby substantially lowering the theoretically 'coverable' size of the protein. When this is taken into account, the one peptide becomes less surprising than it otherwise might be. (PUBMED link at top of PRIDE)

MzData – Mass Spectrometry Data


This details the sample that has resulted in the final identification of the peptide fragment. It also provides information describing the machine and processing software used as well as the principal investigator involved.

The "Additional Sample" information lists the NEWT² taxonomy reference (in this case human). Other controlled vocabularies, including the BRENDA tissue ontology, the cell type ontology and the disease ontology are used to annotate the sample.

This peptide was not the only result of this experiment, and searching PRIDE using the experiment identification number displays the sequences of over 2100 other peptide fragments and their corresponding protein assignments. Some of these proteins have been inferred by the detection of several peptide fragments in the experiment, whilst others have used a single fragment, as is the case in our example.

Spectrum List

The list of spectra associated with the chosen experiment. In this case there is only one spectrum, so the file is not too large. Other queries may return several spectra.

 In the "Spectrum List" section of the results page, follow the "View Spectrum" link to see the ionisation peaks of the spectrum.

This is the actual fragmentation spectrum resulting from the mass spectrometry experiment. It shows the ionization intensity (blue values on the peaks) versus the mass/charge ratio (red values).

² <http://www.ebi.ac.uk/newt/display> UniProt taxonomy browser

- 🖱️ “Roll” your mouse over any of these peaks to see the mass over charge ratio and intensity values.

The difference between the masses of two peaks is an indication of which amino acids were present in the fragmented peptide. Since the flanking masses are also known, rough positional information can also be obtained in this way. When a series of a few consecutive amino acids have been determined, the result is called a ‘sequence tag’. These take the form of: ‘396.5 GTHE 412.7’, with the flanking masses indicated on either side of the one-letter notation sequence. The very first search algorithms were often based on sequence tags, and even today such algorithms persist for specific tasks.

- 🖱️ Click on one of the peaks, and then on another to see the difference in mass and a potential amino acid suggestion appear on the spectrum display. Couple that to another amino acid in series.

The ionization peaks are marked in red - denoting the Y series, or in blue denoting the B series of ions – this can be toggled between using the option in the top right hand corner of the spectrum view. The b- and y-ions (Biemann nomenclature³) refer to the fragments that contain an intact peptide amino-terminus or carboxy-terminus, respectively.

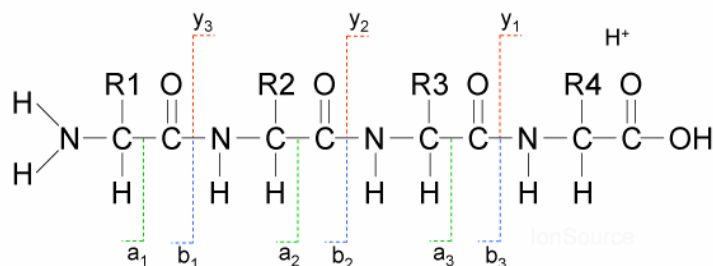


Figure 1: Diagram of ion cleavage sites. a and b and y indicate which bond is cleaved within the amino acid. The peptide bonds that break to generate complementary b and y fragments are generally seen as being the most susceptible to fragmentation, thus it is oft en only these that are represented.

An example for the first 6 fragment ions of each type is given below.

Sequence: **YSFVATAER**

	1	2	3	4	5	6
b-ion	Y	YS	YSF	YSFV	YSFVA	YSFVAT
y-ion	R	ER	AER	TAERT	ATAER	VATAER

You will notice that b3 and y6 taken together actually constitute the whole peptide. This is to be expected as the fragmentation process actually breaks the peptide backbone and as a result always generates complementary fragments. Each resulting fragment can potentially fragment again, yielding a child fragment of the same type as well as an ‘internal’ fragment (one that does not include either terminus). The prevalence of the latter depends on the instrument and the fragmentation process used.

It is also important to note that whenever a peptide breaks into two complementary fragments, you do not necessarily see both fragments. The mass spectrometer can only detect charged

³ Biemann K. Contributions of mass spectrometry to peptide and protein structure. Biomed Environ Mass Spectrom. 1988 Oct;16(1-12):99-111

fragments (the 'fragment ions'). If, for instance, the precursor peptide is singly charged, only one of the daughter fragments can inherit the charge, leaving the instrument blind to the uncharged sibling fragment.

When a series of connected peaks containing amino acid information becomes sufficiently large, it is called de novo sequencing. It is relatively complex to do this reliably, but once the masses of individual amino acids are known, the task becomes much easier and gaps between the peaks more meaningful. Appendix I contains a list of each amino acid and its corresponding mass. A similar list is actually built into the spectrum viewer, which is how it manages to report matching amino acid masses when you select two peaks.

As the amino acids are ionised, it is not necessarily the case that they are ionised with only a single charge, and extra protons may attach themselves to an amino acid. To compensate for the correspondingly reduced mass-over-charge ratio of these multiply charged amino acids, there is an option to include a list of amino acid mass-over-charge ratios with a proton charge greater than +1.

- 🔑 Hit the refresh button of your browser to reset the spectrum view.
- 🔑 Check the box "Include ions where $z > 1$ " in the top panel of the spectrum view. In order to make our de novo sequencing as accurate as possible, alter the "Mass Error" to read 0.199 daltons (this is the error rate quoted in the original paper for this experiment) by dragging the blue arrow towards the left of the log triangle.

Whilst many of the peaks are individually visible, some overlap. The zoom arrows on the top panel can be used to zoom into a region on the X (m/z) axis or increase the height of the peaks along the y (Intensity) axis. To enhance this effect, gridlines for the graph may be toggled on or off, and the "Peak List" can be accessed to see the actual peak values in a list.

- 🔑 You can actually sequence a fair part (about 4 residues) of the C-terminus of this peptide directly from its spectrum. For this, leave the Y-ion series selected (remember, y-ions are the ones that include the C-terminus of the original peptide) and click on the really big peak at 175 m/z units. This peak actually represents the singly charged arginine residue ($C_6H_{14}O_2N_4$) $1+$ at the actual C-terminus.

? 1 Why is it no surprise that it is arginine? Could it have been something else as well?

If you now look at your table of amino acid masses in the Appendix, you'll find that R (arginine) has a monoisotopic mass of approximately 156.10 Da – a difference of 19 Da from the 175 m/z peak. This apparent difference is the result of the way the masses for the single amino acids are tabulated. Look at the composition in this table; they omit one oxygen atom, two hydrogen atoms and one proton (for the +1 charge) when compared to a y-ion (if you add these up you should get the required 19 Da difference).

An arginine residue has a really basic sidechain (which can be verified by looking up the pKa for it) and as such, it is no surprise that a single R residue ionizes well – hence the highly intense peak.

- ✎ Try to extend the C-terminal coverage by attempting to pick out the y2, y3, y4 and y5 ions. The PRIDE spectrum viewer will assist you with this as it displays potential amino acids in the blue box once you mouse over a peak.

A simple way to query PRIDE is to use the Browse PRIDE page.

- ✎ In the main PRIDE menu on the left, click on “Browse Experiments”.
- ✎ From the “Browse by Tissue” section near the bottom of the page, click on “brain” (or the corresponding BTO accession). Take a look at the summary page listing all of the experiments that match this search.
- ✎ Find the experiments on the summary with accession number 1636. What tissues has this experiment been annotated with? Note that brain is not one of them – this is an example of the functionality offered by the Ontology Lookup Service.
- ✎ Tick any two or three experiments in the “Compare Protein Identification Sets” column and then click on the “Compare Experiments” button at the top of the table. This will display the overlap between the protein identifications in the selected experiments.

The following tutorial activities are intended to illustrate the potential of the PRIDE BioMart for querying PRIDE.

- ✎ Navigate to <http://www.ebi.ac.uk/pride> and click on the “PRIDE BioMart (Beta)” link in the menu on the left.
 - ✎ Use the BioMart to find out which literature reference(s) are available for PRIDE experiment accession 2.
- ? 2 How many different references are there?
- ✎ Create a link from BioMart to access the abstracts for these references on the CiteExplore system.
 - ✎ Use the PRIDE Biomart to create and open an Excel spreadsheet containing data for peptide sequences identified in PRIDE in Experiment Accession 2 where the corresponding protein identification has a score of > 85. Include the experiment accession number, protein accession number, species, tissue and peptide sequence.
 - ✎ Find out how many times the (exact) peptide sequence HEVININLK has been identified in PRIDE.
 - ✎ Refine your query to find out how many distinct / different mass spectra are included in PRIDE that provided evidence for this peptide.

- ✍ Look at these spectra using the PRIDE spectrum viewer (directly linked from BioMart) to compare (visually) how similar they are.
- ✍ Is this a potential proteotypic peptide? If it is definitely not, how many proteins have been identified by this peptide in PRIDE?
- ✍ In which species and tissues has this peptide been identified?
- ✍ Take a look at the NEWT (taxonomy) entries for these species by following a link directly from BioMart.

Preparing Your Data for Submission to PRIDE using Excel

If you are interested in submitting data to PRIDE using a spreadsheet, you may wish to complete this tutorial activity in your own time. Note that you must use Microsoft Excel on a Windows computer to use the Proteome Harvest spreadsheet.

One of the major constraints in submitting data to PRIDE is the complexity of the data model. To try to alleviate this problem, the PRIDE team have developed a Microsoft Excel based submission tool that uses embedded code written in Visual Basic to generate a valid PRIDE XML file. This spreadsheet is also able to connect directly to the Ontology Lookup Service to allow the user to search for suitable terms to annotate their data.

- ✍ Navigate to <http://www.ebi.ac.uk/pride/proteomeharvest/index.html>

This page gives an overview of the Proteome Harvest project and provides links to download an example, populated spreadsheet, and an empty Proteome Harvest spreadsheet that you can use to prepare a submission. There is also a series of tutorial movies that will introduce the spreadsheet and its use.

- ✍ Scroll to the bottom of the page and click on the “Introduction” tutorial. You may wish to watch this tutorial (~ 5 minutes) to get a feeling for how the spreadsheet operates.
- ✍ Click on the back button to return to the main Proteome Harvest page and then click on the link “PRIDE Proteome Harvest Spreadsheet Demonstration, Version 0.1 beta” (the second link) which will download a version of the spreadsheet that is already populated with data.

The code in the spreadsheet has been digitally signed to assure you that it has been written at the EBI and has not been modified before getting to you. The first time you open the spreadsheet, you are likely to be asked if you wish to accept this signature. You should accept this signature. You can also indicate that you ‘trust the EBI’ so you will not need to respond to this question again.

- ✍ Click on the Sample tab.
- ✍ Throughout PRIDE, controlled vocabulary terms can be added to annotate the data. You will add a term to annotate the experiment on this page. First of all, double-click the “Add Param” button in cell F22.

- ✎ Now double-click the “Search for CV Term” button to open the “Parameter Entry Form”. Try to find a term “Disease Free” and add it to the list of parameters on the Sample sheet.

Next you will add an additional protein identification that has nine associated peptides and a number of protein modifications to show you some of the support for this kind of data entry in the spreadsheet.

- ✎ Click on the “Proteins” tab to open the protein identification spreadsheet. Add a “Gel Free” protein identification for the protein “Q14974” which has been identified from the “UniProt” database. Enter “9” in the “Peptide Count” column and then double click the “Add Peptides” button in column A on this row.

At this point you should see nine rows added to the “Peptides” sheet that are cross-referenced to this entry on the “Proteins” sheet.

- ✎ Navigate in an Internet browser to <http://www.ebi.ac.uk/~pjones/peptides.txt> and save this file to your machine. Open this file in a text editor (such as notepad) where you will see that it is a list of 9 peptides. You should also notice that some of the amino acids are annotated with either * or # to indicate the presence of a modification. Keep this file open for the time being.
- ✎ On the spreadsheet, click on the “PTM Codes” tab. Here you will see that * and # have been defined as indicating the presence of specific modifications. Next, copy the nine annotated peptides from the text file that you opened in the previous step. Click on the “Peptides” tab and paste these nine peptides into the first empty space available on column “E”.
- ✎ Finally, take a look at the PTMs sheet to see the additional modification rows that have been added, depending upon the position and presence of the annotations on the peptide sequences that you pasted in.

Hopefully this series of activities will have demonstrated to you that the Proteome Harvest spreadsheet provides the tools and automation necessary to make submitting to PRIDE an achievable task without the necessity for programming.

Further Reading

Martens L, Hermjakob H., Jones P, Taylor C, Gevaert K, Vandekerckhove J, Apweiler R. “**PRIDE: The PRoteomics IDentifications database**” *Proteomics* (2005) Vol 5 Issue 13 Pages 3537-3545.

Jones P, Cote R, Martens L, Quinn A, Taylor C, Derache W, Hermjakob H, Apweiler R. “**PRIDE: a public repository of protein and peptide identifications for the proteomics community**” *Nucleic Acids Research* (2006) Vol 1 Issue 34 (Database issue) D659-D663

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W
“**BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.**” *Bioinformatics*. (2005) Aug 15;21(16):3439-40.

Biemann K. “**Contributions of mass spectrometry to peptide and protein structure.**” *Biomed Environ Mass Spectrom*. (1988) Oct;16(1-12):99-111

Appendix 1: Amino Acid Masses

1-letter code	3-letter code	Chemical formula	Monoisotopic ⁴	Average ⁵
A	Ala	C ₃ H ₅ ON	71.03711	71.0788
R	Arg	C ₆ H ₁₂ ON ₄	156.10111	156.1875
N	Asn	C ₄ H ₆ O ₂ N ₂	114.04293	114.1038
D	Asp	C ₄ H ₅ O ₃ N	115.02694	115.0886
C	Cys	C ₃ H ₅ ONS	103.00919	103.1388
E	Glu	C ₅ H ₇ O ₃ N	129.04259	129.1155
Q	Gln	C ₅ H ₈ O ₂ N ₂	128.05858	128.1307
G	Gly	C ₂ H ₃ ON	57.02146	57.0519
H	His	C ₆ H ₇ ON ₃	137.05891	137.1411
I	Ile	C ₆ H ₁₁ ON	113.08406	113.1594
L	Leu	C ₆ H ₁₁ ON	113.08406	113.1594
K	Lys	C ₆ H ₁₂ ON ₂	128.09496	128.1741
M	Met	C ₅ H ₉ ONS	131.04049	131.1926
F	Phe	C ₉ H ₉ ON	147.06841	147.1766
P	Pro	C ₅ H ₇ ON	97.05276	97.1167
S	Ser	C ₃ H ₅ O ₂ N	87.03203	87.0782
T	Thr	C ₄ H ₇ O ₂ N	101.04768	101.1051
W	Trp	C ₁₁ H ₁₀ ON ₂	186.07931	186.2132
Y	Tyr	C ₉ H ₉ O ₂ N	163.06333	163.1760
V	Val	C ₅ H ₉ ON	99.06841	99.1326

⁴ Monoisotopic masses can be used if the mass spectrometer has enough resolution to visualize each isotope independently (note that the higher the charge state, the smaller the distance in m/z units between isotopes). Most present-day instruments can do this - resulting in the typical 'mountain range' of peaks for each peptide, separated by 1; 0.5 or 0.33 m/z units (for +1, +2 and +3 charges, respectively) and declining almost exponentially towards higher m/z (less common, heavy isotopes).

Iontraps (especially the popular 'LCQ' machines by Thermo Finnigan) are notoriously bad at this and therefore usually produce a single 'big lump' peak, the top of which is usually skewed towards the lightest isotope (because this is the most abundant for masses < 1800 Da) but not positioned where the lightest isotope would actually be.

⁵ Average masses are calculated by taking into account several possible isotopes of CHONPS atoms and their relative abundances to come up with a mass that is most likely to correspond to the top of the single peak (and therefore to the assigned m/z as reported by peak-picking algorithms).

Answers to selected Questions

1. Trypsin predominantly cleaves proteins at the carboxyl side (or "C-terminal side") of the amino acids lysine and arginine, except when either is followed by proline. As a consequence, it is most likely that the C terminus is either lysine or arginine.