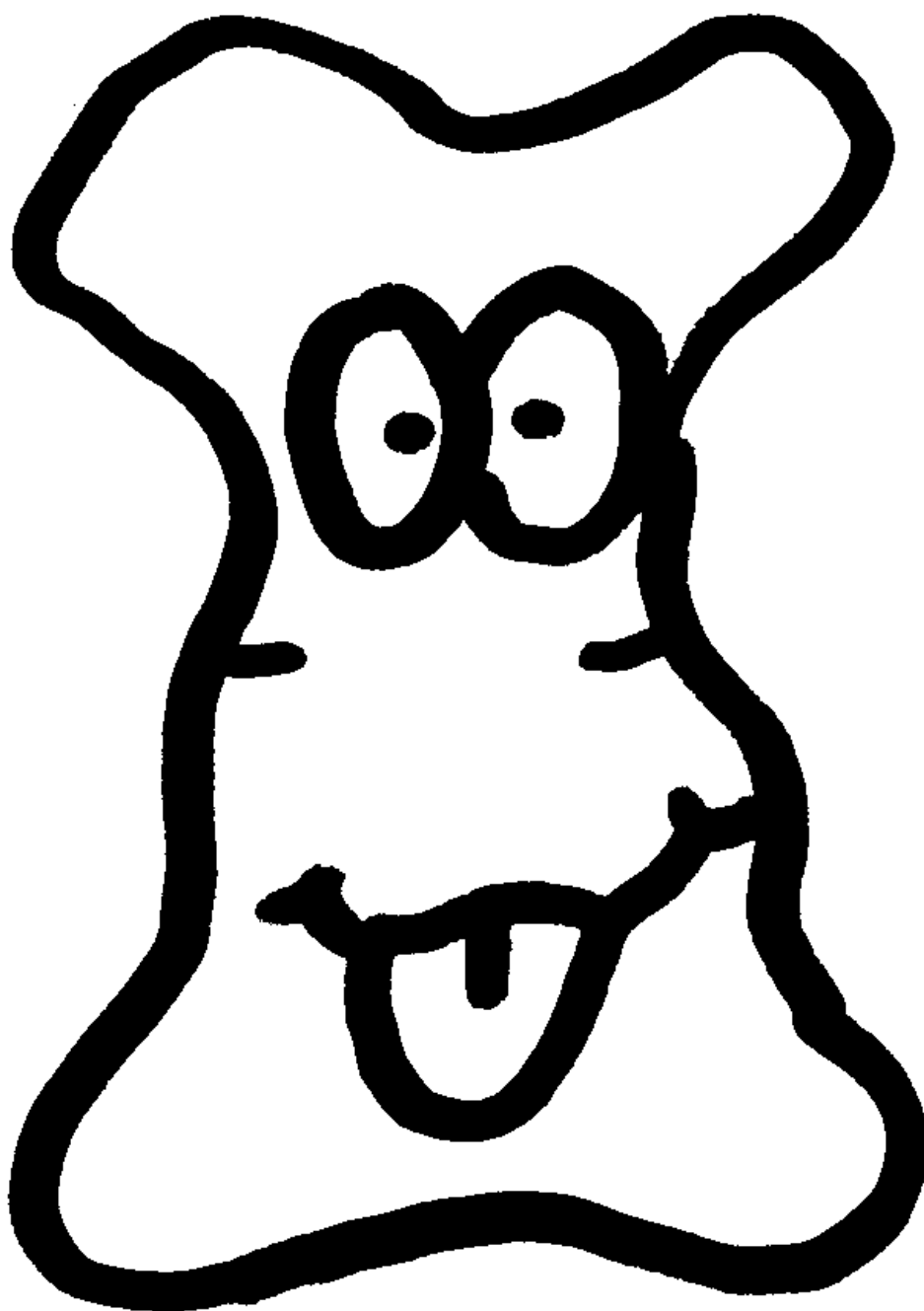


GPMAW for Dummies

Rev. 5

by Peter Højrup



Introduction

The purpose of this booklet is to introduce you to the workings of GPMaw. This booklet is not intended as a replacement for the manual, but rather as an addition. Where the manual gives you the factual workings of the program, “GPMaw for Dummies” shows you how the program works by describing simple examples (probably in more detail than you want to know). In addition, a large number of tips and hints are given in sidebars.

This booklet is based on questions frequently raised by users and encountered through my job as a teacher. The booklet is not a static object, but will (hopefully) develop over time. I will greatly appreciate feedback, both as criticism and as suggestions for additional chapters (e-mail: php@bmb.sdu.dk).

A few conventions are used as follows:

More : The current subject will be covered in more detail in a separate chapter.

Setup : The default value of the current feature can be changed in Setup system (accessed through the menu item **Setup | Setup system** in GPMaw).

Commands accessed through the menu are shown as **File | Export sequence | Export to clipboard** (meaning select the menu item ‘File’ then sub-item ‘Export sequence’ followed by sub-item ‘to clipboard’).

The present guide is written based on GPMaw version 9.10.

Index

Chapter A – Basic sequence handling

1 – The sequence.....	2
2 – The toolbar	3
3 – Selections.....	5
4 – Coloring residues.....	6

Chapter B – Calculating mass values

1 – The mass file	7
2 – Other ways of changing the mass.....	8

Chapter C – Disulfide bridges and multiple chains

1 – Obtaining the sequence – The easy way	10
2 – Obtaining the sequence – Entrez (World Wide Web)	11
3 – Editing the sequence	12
4 – Disulfide bonds	14
5 – Cleaving proteins – also with linked peptides.....	15

Chapter D – Post-translational modifications

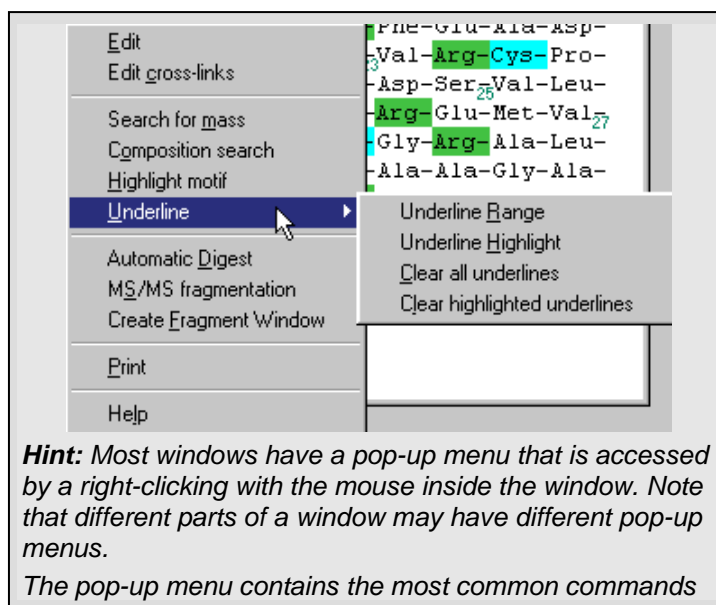
1 – Obtaining the sequence – Swiss-Prot	17
2 – Inserting post-translational modifications	19
3 – N-linked glycosylation	22

Chapter E – Getting data out of GPMW

1 – Protein sequence.....	24
2 – The peptide list	25
3 – Mass search results.....	27
4 – Presenting sequence coverage	28

Chapter F – Mass search of two proteins

1 – Preparations	30
2 – Results	31



Chapter A – Basic sequence handling

1 - The sequence

The basic unit of most work in GPMW is a protein sequence. You may choose to enter a sequence manually, load it from a database or directly from the Web (e.g. see B.1/2 and C.1). However, when you work with a sequence on several occasions, you will normally save it to a file on your hard drive. The main advantage to saving sequences locally, in addition to faster access, is that the sequence saved from GPMW may contain additional information, e.g. modified residues, cross-links, annotations etc. Note that GPMW sequence libraries may contain multiple sequences. This works in the way that when you save to a file, which already contains one or more sequences, the new one is appended to the file, instead of replacing it.

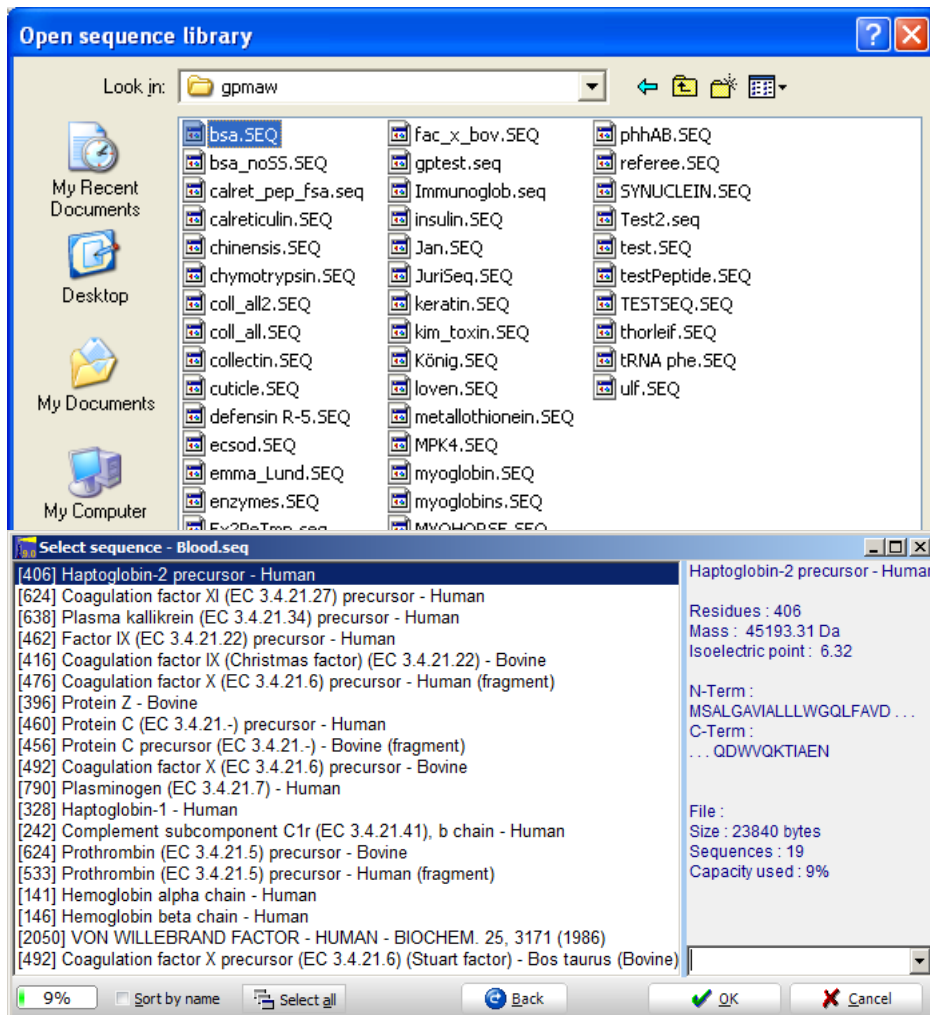
When reading back a sequence already saved in a sequence library, there is a difference between opening a library file containing a single sequence and one containing multiple sequences. If the library file contains a single sequence, it will be read into GPMW immediately you select the file, while you will be asked to select a sequence from a list if the library file contains multiple sequences.

Start by loading a sequence already saved in a file on the hard drive.

Select **File | Open** or click on the 'Open file' icon. This opens the 'Open sequence library' dialog:

Select the 'blood.seq' file and you are greeted by the 'Select sequence' dialog.

This shows a GPMW sequence library files containing multiple sequences. As a sequence file is limited to a size of 264000 bytes (characters), there is a limit to the number of sequences that can be stored in a single library. In the present library, 'blood.seq', you can see (status bar at the bottom) that the file is 9% filled.



Note: The 'Open sequence library' dialog will always show the same directory initially (can be user-defined). **Setup**

Hint: The nine most recently opened files are shown at the bottom of the File menu, thus making it easy to access the same sequences again.

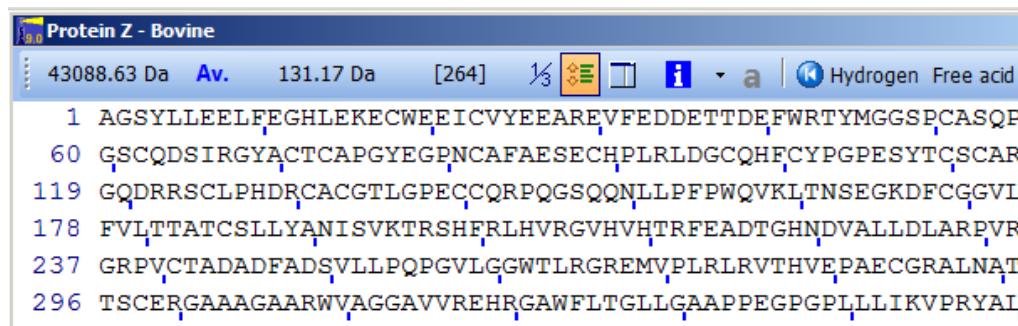
Note: When you close a sequence window, all associated windows will be closed as well. As all associated windows refer to the same sequence, you should not edit the sequence while associated windows are open, as the results are unpredictable.

GPMW for dummies

The status bar is initially green, but turns yellow and red as you get close to full capacity of the file. Furthermore, when a protein is selected, the basic information is displayed in the right-hand information box.

You may now open a sequence either by selecting it with the mouse followed by 'OK' (or the Enter key), or you may double-click on the sequence name to open it directly. Alternatively you can select multiple sequences by holding down the Ctrl key while selecting for a discontinuous selection. Use the Shift key for a continuous selection. You open all the selected sequence by pressing the 'OK' button. The 'Back' button returns you to the file selection dialog. The right-hand drop-down box lists the most recently opened files.

Select Protein Z, and press 'OK' and you will open this sequence window



Note: The functions in the sequence toolbar are local to the sequence window, unlike the main toolbar that contains commands global to all windows.

This is the basic working window of GPMW. Most other windows will be derived from this window and are called daughter windows.

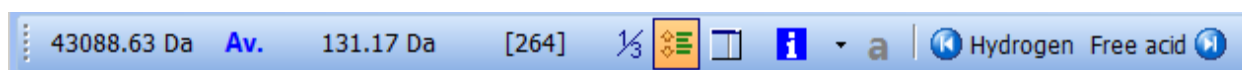
The sequence will always be displayed with the number of the first residue to the left of each line. The number of residues on a line will be the maximum number possible inside the given window, except when the 'Multipla 5' is turned off *Setup*. With the 'Multipla 5' turned on the number of residues on each line will be a multiple of 5 (i.e. 20, 25, 30...).

Every 10th residue is labeled with the residue number as a subscript when showing 3-letter residue code (divided by 10, i.e. residue 120 is labeled with 12) or as a small tick mark when showing 1-letter code. The feature can be turned off *Setup*. Use the 1/3 button in the toolbar to switch between 1- and 3- letter code.

Hint: You can select to have the most recently accessed sequence open automatically next time you open GPMW. *Setup*

The colors of the displayed residues can be changed for easier navigation and to indicate modifications and changes *More*.

2 - The toolbar



The toolbar of the sequence window contains the following:

43088.63 Da **Av.** The leftmost panel shows the total mass of the protein in Daltons. The button next to it shows whether it is the average mass (Av. – blue) or monoisotopic mass (Mo. – red). You can change the mass type by clicking on the Av./Mo. button.

The next two panels from the left can show either of two states:

131.17 Da **[264]** If no peptide is selected the mass of the amino acid under the cursor (i.e. the residue mass + 18 Da) will be displayed in the left-hand panel and the right-hand panel will show the residue number (if the protein is a multi-chain protein, the first chain will be labeled 'a', the second 'b' etc (i.e. 80b is residue number 80 in the second chain counting from the N-terminus)).

Note: The Av./Mo. button will affect all mass calculations made on the sequence. For proteins the average mass makes most sense, but when you highlight peptides, you should switch to monoisotopic mass.

1314.46 Da[1] [13-22]

.FEGHLEKECWEEICVY

If part of the sequence is highlighted, the left panel shows the mass of the selected peptide and the right panel shows the first and last residues of the selection (see next section). Note that the mass value displayed is M, not M+H.



The 1/3 button toggles between 1- and 3-letter code (i.e. KTA vs Lys-Thr-Ala). When you toggle between the two modes, all selections, highlights, coloring, links etc. are conserved. The default setting is done in Setup; separately on the 'Peptide' AND the 'Display' page for the peptide and sequence windows respectively.



Line distance. When the button is pressed, the distance between the sequence lines will be increased for easier viewing.



The '**Frames**' button toggles the display of an information frame to the left of the main sequence window. This window is dynamic, as the content of the frame will be updated when the content/selections are changed.

Protein	1	EPAYV
Termini:	52	SQDAE
Modified residues	103	IDMHC
Asp104 Methylation [C]	154	LYTLI
Cross-linked residues	205	RAKII
Cys120-Cys146	256	GEWKE
Net charge	307	GTIFI
Molar Ext./Abs. @280	358	KRKEE
Highlights		
Inverted: 7 [1.8%]		
Underlined: 0 [0.0%]		
Sel. mass		
MH1+: 759.834		
MH2+: 380.421		
MH3+: 253.949		

The information available is close to the same information given in the 'sequence information window' (see below), but in the frame it is updated dynamically. The individual terms are initially hidden, but can be expanded by clicking on the small '+', changing it into a '-'.
Information:

Termini: Name and composition of the N- and C-termini.
Modified residues: Name, position and composition of all individually modified residues (not the ones changed globally through the mass table).
Cross-linked residues: Residues that are cross-linked, typically cysteine residues (see B-3).
Net charge: The theoretical charge of the protein at pH 2.0, 7.0 and user-selected pH (Setup).

Molar Ext./Abs.@280: Theoretical extinction coefficient / absorption of the protein at 280 nm.
Highlights: Percentage of the sequence, which is inverted (highlighted) or underlined, updated dynamically.

Sel. mass: Mass of the selected (highlighted) part of the sequence. Shown as singly, doubly and triply charged peptide (i.e. residue mass + 18 + charge). Updated dynamically. Note that unlike the peptide mass shown in the toolbar of the sequence window, this is the charged ion. More on peptide selections in part 3.

As can be seen in the sequence part of the figure, modified residues are colored (red). *More*



The white on blue 'i' opens the sequence information window, which gives you file and statistical information, calculated indices (pI, absorption etc.), amino acid composition and multiply charged masses. The drop-down arrow opens a menu giving direct access to the various pages in the information window.



The next button is a shortcut to the annotation page. If the annotation is empty, the 'a' in

88273.14 Da

1M+ 88274.14 / M1+ 88274.14

2M+ 176547.28 / M2+ 44137.58

3M+ 264820.42 / M3+ 29425.39

4M+ 353093.56 / M4+ 22069.29

5M+ 441366.69 / M5+ 17655.63

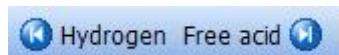
6M+ 529639.83 / M6+ 14713.20

Copy mass value

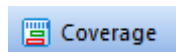
Tip: If you right-click in the mass panel, a window will display multiply charged ions and cluster ions. The same information can be found in the Sequence information window *More*, but is quicker to access here.

Note: Although GPMW can display sequences in both 1- and 3-letter code, all sequence input in the editor and input lines has to be in 1-letter code to prevent ambiguity.

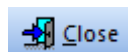
the button will be gray, if the annotation page contains text, it will be green, blue or red depending on the content of the annotation (Swiss-Prot, Entrez or unknown format respectively).



The next two buttons show the status of the N- and C-terminal respectively (double-click to edit).



The Coverage button only appears when the sequence has a corresponding coverage map associated. This can be generated from mass searches, see later.



The Close button closes the sequence window and all daughter windows (i.e. peptide, cleavage, search and graphical windows)

Main toolbar – Search & cleavage section:



Two sections of the main toolbar are of direct interest to the sequence window, the *Control* section (described later in Mass calculations) and the *Search and cleavage* section:



Color residues. You can color residues in the sequence in up to three different colors. Click on the main button to enter residues or sequences (motifs) in specific colors. Alternatively select the drop-down list for quick access to the most common residues (e.g. Lys+Arg for tryptic cleavages sites; Phe, Tyr + Trp for chymotryptic cleavage sites etc.). [More](#) (Coloring residues)



The magnifying glass is a shortcut to peptide mass searching of the protein. [More](#)



Ms/ms search. Search the protein, list of proteins or database using peak lists in either mgf, dta or pkf format.



The scissors are a shortcut to cutting up the protein into peptides. This is usually done using proteolytic enzymes, but may also be carried out chemically. The only requirement is that the process can be specified relative to specific residues. The small down-arrow button opens a menu enabling you to select enzyme cleavage in a single click. Note that the bottom part of the menu enables you to specify one missed cleavage, and/or to digest all opened sequence windows. [More](#)



Fragment button. Create (ms/ms) fragments of your protein. If part of the main sequence is highlighted, this part will be taken as the fragment peptide. If no selection has been made, the whole sequence is taken as input. If the sequence is longer than 50 residues you will be asked for confirmation before the first 400 residues are used as input.

3 - Selections

```
16 Lys-Glu-Cys-Trp-
31 Val-Phe-Glu-Asp-
46 Met-Gly-Gly-Ser-
61 Ser-Cys-Gln-Asp-
```

A useful feature is to determine the mass of a peptide that are part of the displayed protein. This is easily carried out in GPMW by pointing the mouse at the first or last residue in the peptide, press the left mouse button and drag the mouse cursor across the sequence.

The mass of the selected peptide and the region covered will be shown in the toolbar **1041.13 Da[1] [103-112]**.

The [1] displayed in the first pane indicates that only a single peptide has been highlighted. You can highlight multiple sequences by holding down the shift button while selecting additional regions. Up to three regions can be selected at a given time.

Tip: If you rest the mouse cursor above either panel, the fly-by help will open as a small pop-up window showing the chemical formula of the terminal.

Note: When multiple sequences are highlighted, the mass displayed will then be the combined mass of all the selected regions. Each region will be calculated as a peptide, i.e. residue masses + 18 Da.

The region covered will only be shown for the last selection, while the mass will be for the total..
You deselect all regions by clicking once in the sequence without holding down the Shift key.

The **arrow keys** can alter the most recently selected region:


The left/right arrow will change the position of the C-terminal residue of the selection one residue back or forward. Holding down the Ctrl key will similarly change the N-terminal residue of the selection. Holding down the Shift key will move the whole selection. This sound complicated, but is straightforward once you try it.

If you copy the sequence to the clipboard (**Edit|Copy to clipboard** or Ctrl+C), you will only copy any peptide(s) selected. If no part of the sequence is selected, the whole sequence will be copied to the clipboard in the format in which it is displayed (i.e. 1- or 3-letter code).

Note, when copying this way, only the sequence and not the name will be copied. Pressing Ctrl+F you will copy the sequence in FastA format (i.e. 1-letter code including the name of the sequence). If you want to copy the complete protein information or if you want to format the sequence (e.g. for a report), you should use the **File|Export sequence** menu option.

4 - Coloring residues.

One of the most efficient tools to examine and manipulate a sequence is to color the background of specific residues. This is done either through the 'Highlight residues' dialog

(**Search | Highlight residues (motifs) ...**), F4, the highlight button  (in the main toolbar) or by right-click and select 'Highlight residues' from the pop-up menu.

In the example (right) the basic residues have been colored one color, cysteines another color and N-glycosylation sites (N-X-S/T/C) a third color.

```

-Leu-Gly-Pro-Glu-Cys-Cys-
-Phe-Pro-Trp-Gln-Val-Lys-
-Val-Leu-Ile-Gln-Asp-Asn-
-Ala-Asn-Ile-Ser-Val-Lys-
-His-Val-His-Thr-Arg-Phe-
  
```

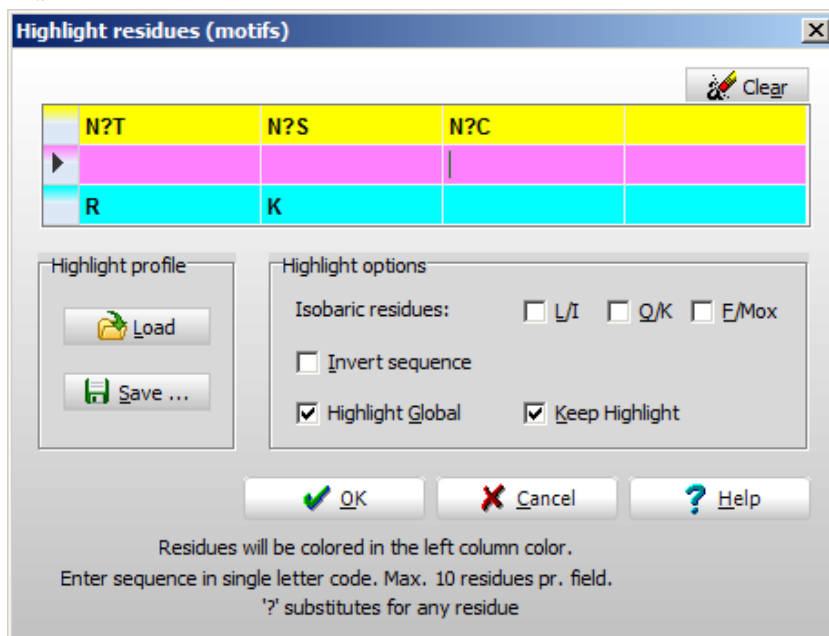
The coloring of residues is done through a simple edit box (right). Three colors are available (presented in the left-hand

column) **Setup**, and for each color you can have four different entries (max. 10 residues in each entry). Notice that the question mark can substitute for 'any residue' (do not use 'X' as this is recognized as a specific residue – unknown). If the **'Highlight global'** check-box is checked, all currently open sequence windows will be colored. If the **'Keep highlight'** is not checked, the entries will be cleared whenever the 'Highlight residues' dialog is accessed.

'Invert sequence' will result in highlighting of sequences

found in both N- and C-terminal directions (e.g. if entering KLGFT both the sequence KLGFT and the sequence TFGLK will be highlighted). Useful for searching for a ms/ms sequence tags (you may not know whether it is a y-ion or a b-ion series). Isobaric residues will, if checked, highlight both kinds of residues, i.e. if the **'Q/K'** checkbox is checked and the sequence LKT is entered, the sequence LQT will also be highlighted. Again, this is for the benefit of ms/ms sequence tags.

The **'Quickcolor'** button (the down-arrow next to the **'Highlight'** button) opens a menu that



enables you to color specific residues with a single click. The choice of selections cannot be changed by the user, but has been chosen to highlight the most common situations (e.g. highlighting R and K for a tryptic digest; C for identifying cross-links etc.).

It is also possible to highlight individual residues. This is done through the '**Modify residue dialog**' (double-click on a residue and click on one of the colored frames at the bottom of the dialog box).

When you highlight a residue you are changing the background color. When you modify or underline residues the colors of the letters are changed (by default modified residues are colored red), so be careful when selecting colors for background as residues will 'disappear' if front and back colors are identical.

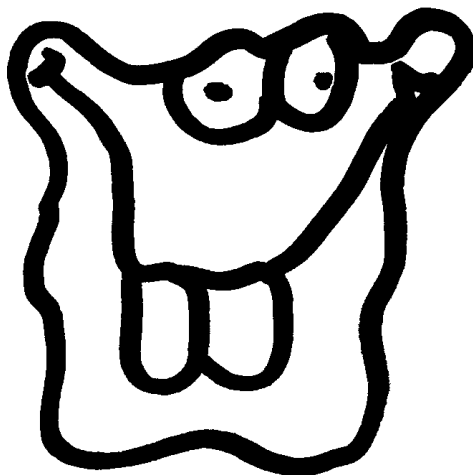
The colors used to highlight residues can be changed by the user *Setup* (Colors page).

Underlines.

Underlines are different from highlights and can be a different way of drawing attention to specific residues/sequences:

- 1) They can be persistent, i.e. you can save the underline information along with the sequence (**File|Save w. highlights**).
- 2) You only have a single color to work with (red by default *Setup*).
- 3) Are often used to transmit information from daughter windows (i.e. mass search) to the main sequence window.
- 4) As they are residue-related, they are always specific to a given sequence.

The underlines are controlled from the pop-up menu in the sequence window and the relevant daughter windows. For more information see the manual and the on-line help.



Chapter B – Calculating mass values.

1 – The mass file.

Calculating mass values is central to handling sequences in GPMW. As the way mass values are used are quite varied, they are subsequently calculated in a flexible (although a little complicated) way.

Peptide/protein mass values are calculated based on 'mass files'. These are constructed of residues. Each residue contains: 1- and 3-letter code, name of residue, atom composition, pKa and charge of side chain. The residue is defined by the 1-letter code, which is what the sequence is stored as in memory/on disk.

1-lett	3-lett	Name	Composition	Monoiso.	pKa	Ch.
X	Xxx	Unknown	C6H8N1O1	110.06	0.00	0
A	Ala	Alanine	C3H5N1O1	71.04	0.00	0
C	Cys	Cysteine	C5H7N1O3S1	161.01	10.30	-1
D	Asp	Asp. acid	C4H5N1O3	115.03	3.50	-1
E	Glu	Glu. acid	C5H7N1O3	129.04	4.50	-1

The table is edited through the menu **Edit | Edit mass file** where you have tables for the amino acid residues, peptide termini and atom mass values.

Atom mass table: The basis of the mass calculations is the composition, which again is based the table of atom mass values. By default this list contains the monoisotopic and average mass value of the 16 most common atoms, but it can be extended to 24 values.

Note that the mass of a proton is defined separately.

The mass file: The mass files are loaded upon startup of the

Mass of a proton

1.00727647

program and form the basis for calculating peptide and protein sequence mass values.

Atom	Name	Ave. mass	Mono mass
C	Carbon	12.011000	12.000000
H	Hydrogen	1.007940	1.007825
N	Nitrogen	14.006700	14.003074
O	Oxygen	15.999400	15.994915

By default GPMW is delivered with a number of files, which only differ in the compositions (mass value) of cysteine. The currently loaded mass file is displayed in the main toolbar next to

the 'SS' button **SS AA_mass.MSS**. The default file is called AA_mass.MSS. If you change the active mass file by selecting a new in the drop-down box (i.e. click on the arrow and select), you will change the mass value of all proteins and peptides opened, to the values defined in the new file. **Note:** this is the easiest way of changing the mass of all residues of a given kind.

Please note that **Cysteine** is calculated differently from all other amino acids. If you define Cys as being in the reduced state, this will be changed by the program to the oxidized state (i.e. a mass value of 103 will be changed to 102). The oxidation step of Cys is controlled through the 'SS' button. When the button states 'SS', Cys will be calculated as 102 Da, when the button is pressed and state 'SH', Cys will be calculated as 103 Da. If Cys is defined as something else but 102/103 Da in the current mass file, the SS button will be inactive.

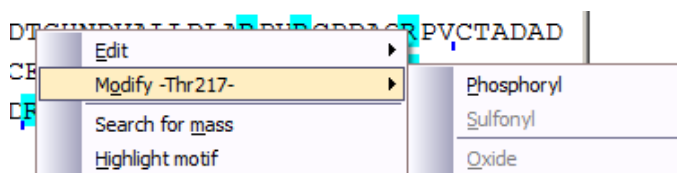
The most typical *example* for the use of a mass file is when you alkylate a protein. E.g. if you reduce and alkylate your cysteines with iodoacetamide, you will just change the mass file from AA_mass.MSS to acetamide.MSS, and all your mass values will be based on Cys = 161 Da.

Termini: This is a separate table from mass files, which specifies the status of the N- and C-

termini. By clicking on the termini buttons in the sequence window **Hydrogen Free acid**, you will open a dialog box enabling you to select the chemical status of the termini. The termini are edited in the same dialog box as above.

2 – Other ways of changing the mass.

In addition to changing the mass of a residue in a global way as described above, you can modify each residue individually. If you right-click on a protein residue, the pop-up menu displays a



'Modify -Xxx-' option, which have a sub-menu detailing the 'simple' modifications available for the particular residue. Selecting one of these will put the modification into the sequence table and the mass value will be recalculated. Whenever a residue has been modified, it will be displayed in **red** in the sequence.

Note: the modification will be carried on into the peptide window and several other.

When you have modified a sequence, you need to save it to 'keep' the modification.

If you 'un-check' the 'Strict modification check' at the bottom of the menu, you can select any of the 'simple' modifications for any residue (although it may not have any biological or chemical relevance).

You may also double-click on a particular residue to bring up the 'Insert modification' box. In the top left-hand box you will see the residue and number, which can be changed. Two drop-down buttons to the right will enable you to either replace a residue or insert a 'simple' modification (the same menu as above).

In the central box is listed the available modifications in the currently loaded **modification file** (see below). You can either double-click or use the 'Select' button transfer values to the edit box above. These can also be filled directly.

Selecting 'OK' will transfer the information to the selected residue in the sequence.

The four bottom panels will just transfer the corresponding color as a background to the residue.

The modification file

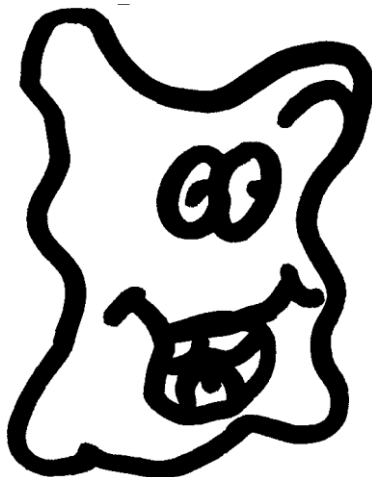
This is a file which contains up to 30 modifications specific to one or more residues. You may have as many files as you like, but only one can be loaded at any given time. In addition to the 'Insert modification' box, the modification file is used in a number of other functions (e.g. mass searches).

The modification file is edited through Edit | Edit modification file. For each modification you have to edit the Name, the Formula, Valid residues (if none are specified, all are valid), the OK box enables you to enable just some of the entries. The Charge, pKa and Terminal are optional. The mass of the currently selected

Name	Formula	Valid residues	OK	Charge	pKa	Term.
Oxygen	O1	M	<input checked="" type="checkbox"/>	0	0.00	-
Methylatio	C1H2	DE	<input checked="" type="checkbox"/>	0	0.00	-
Phospho	H103P1	STY	<input checked="" type="checkbox"/>	-1	3.14	-
thr_ala	-H2C1O1	T	<input checked="" type="checkbox"/>	0	0.00	-
Me-ester	H2C1	DEST	<input checked="" type="checkbox"/>	0	0.00	-
D-Succ	-H2O1	D	<input checked="" type="checkbox"/>	0	0.00	-
Sodiated	-H1+Na1	DE	<input checked="" type="checkbox"/>	0	0.00	-
Deamidation	-H1	DE	<input checked="" type="checkbox"/>	0	0.00	-
Acetyl	H2C2O1	K	<input checked="" type="checkbox"/>	0	0.00	-
di-Methylation	H4C2	K	<input checked="" type="checkbox"/>	0	0.00	-
Methyl	H2C1	K	<input checked="" type="checkbox"/>	0	0.00	-
			<input type="checkbox"/>			
			<input type="checkbox"/>			
			<input type="checkbox"/>			
			<input type="checkbox"/>			

entry is shown to the right. The Unimod file can be opened at the bottom, and you can select entries to the modification file by double-clicking on the Unimod entries. If you do not know the chemical composition of a modification, you can enter it by selecting an empty line and click the 'Mass only' button.

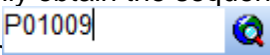
A modification file has to be saved to disk before you can use it in GPMW.



Chapter C – Disulfide bridges and multiple chains.

1 - Obtaining the sequence – The easy way

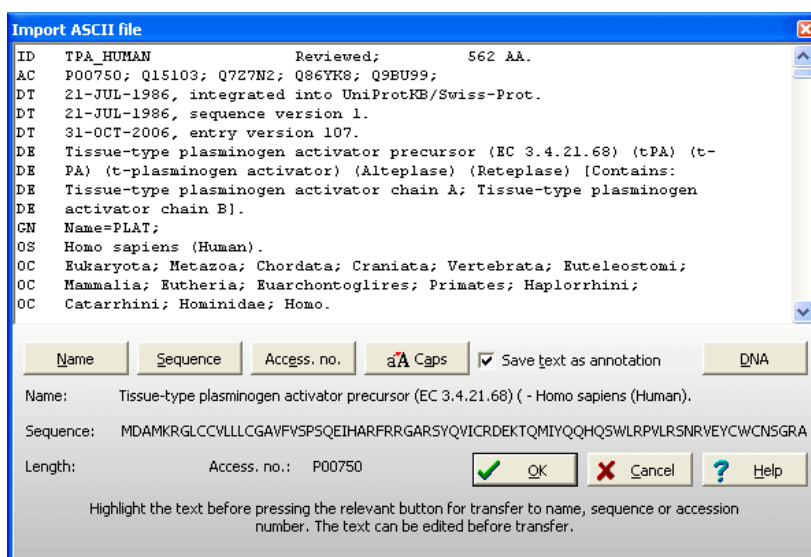
If you know the accession number of a given sequence, does not matter from which protein database, you can most easily obtain the sequence by entering the number in the web access

input box in the main toolbar . If this section is not available, right-click in an empty section of the toolbar and select 'Web' from the pop-up menu.

You retrieve a sequence by entering the accession number and either press 'Enter' or click on the 'Web' icon. The edit box has a second function, as you can enter a residue or small sequence and press the 'Mark' button to color the relevant residues in the sequence. Settings for the coloring are taken from the 'Color residues' dialog box, see above.

GPMW will search the **UniProt** database (ExPASy web site) for all accession numbers entered that start with O, P or Q. All other accession numbers will be searched in the **NCBI nr** database (the ExPASy web site). I strongly recommend that you extract sequences from the UniProt database, as the sequences here are curated, and GPMW is able to extract sequence modifications directly from this format, see below.

No matter which database was searched, the results will be



Import ASCII file

ID TPA_HUMAN Reviewed; 562 AA.
AC P00750; Q15103; Q727N2; Q86YK8; Q9BU99;
DT 21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT 21-JUL-1986, sequence version 1.
DT 31-OCT-2006, entry version 107.
DE Tissue-type plasminogen activator precursor (EC 3.4.21.68) (tPA) (t-
DE PA) (t-plasminogen activator) (Alteplase) (Retelase) [Contains:
DE Tissue-type plasminogen activator chain A; Tissue-type plasminogen
DE activator chain B].
GN Name=PLAT;
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC Catarrhini; Hominidae; Homo.

Name Sequence Access. no. aA Caps ☒ Save text as annotation DNA

Name: Tissue-type plasminogen activator precursor (EC 3.4.21.68) (- Homo sapiens (Human)).
Sequence: MDAMKRGGLCCVLLCGAVFVSPSQEIHARFRRGARSYQVICRDEKTMQMIYQQHQSGLRPVLRSLNRVEYCWCNNGRA
Length: Access. no.: P00750

OK Cancel Help

Highlight the text before pressing the relevant button for transfer to name, sequence or accession number. The text can be edited before transfer.

presented in the 'Import ASCII file' dialog box:

As GPMW recognizes the format of both Swiss-Prot and Entrez, the record will be parsed into the relevant sections (i.e. name of sequence, the sequence itself, and the accession number). These can be reviewed below the record.

If the sequence is not displayed, you have to select it manually: Highlight the part of the record representing the name, and click on the 'Name' button. Highlight the accession number, and press the 'Access. No.' button. Scroll to the bottom of the record, highlight the sequence and press the 'Sequence' button. **Note** as GPMW only imports 1-letter codes that are defined in the current mass file, space characters, numbers, backslash etc. are ignored and not imported.

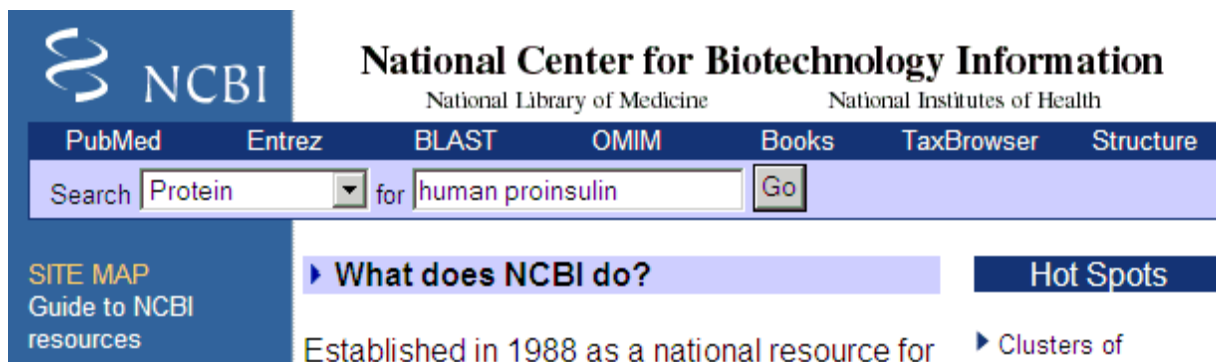
Press the 'OK' button to import the sequence into a GPMW sequence window. If the 'Save text as annotation' is checked (default), the entire annotation will be saved in the annotation window and will be saved along with the sequence, allowing you to access it at a later date.

Note: The top part of the dialog box is an edit box. This means that you can edit the text prior to importing it into a sequence window.

2 - Obtaining the sequence – Entrez (World Wide Web).

If you do not know the accession number of a given sequence, if the web retrieval doesn't work or you are just browsing the web and happen to meet an interesting sequence, it is nice to know that GPMW has a very flexible sequence input system:

For this example we will obtain our sequence from one of the most popular molecular biology sites on the web, the NCBI site (<http://www.ncbi.nlm.nih.gov/>). The web site is powered by the Entrez search engine, and we will search in the protein database.



Select "Protein" in the left-hand drop-down box, and enter 'human proinsulin' in the search box. Press Enter or click on the go button.

In the results page select number '9' by clicking on the underlined accession number 'P01308'. Most of the results from the search are human insulin, but from different databases. P01308 is from the Swiss-Prot database. We select this entry because it is the best annotated database. You can usually recognize Swiss-Prot entries (or the associated TrEMBL entries) by starting with 'O', 'P' or 'Q' followed by 5 characters or ciphers. There is more information on Swiss-Prot in chapter C.1.

- ☐ 8: [NP_000198](#)
proinsulin precursor [Homo sapiens]
gi|4557671|ref|NP_000198.1|[4557671]
- ☐ 9: [P01308](#)
Insulin precursor
gi|124617|sp|P01308|INS_HUMAN[124617]

The result of the search is by default shown in GenPept format. This is OK, as we will get most additional information this way.

GPMW for dummies

1: [P01308](#) . INSULIN PRECURSOR...[gi:124617]


```
LOCUS      INS_HUMAN      110 aa      PR
DEFINITION INSULIN PRECURSOR.
ACCESSION  P01308
PID        gi124617
VERSION    P01308 GI:124617
DBSOURCE   swissprot: locus INS_HUMAN, accession P013
           class: standard.
           created: Jul 21, 1986.
           sequence updated: Jul 21, 1986.
           annotation updated: May 30, 2000.
           xrefs: gi: 33930, gi: 758088, gi: 186437,
```

Highlight the entry starting with 'LOCUS' and move all the way down beyond 'ORIGIN', remember to include the ending '//' (including the '//' is also important when loading Swiss-Prot records).

```
ORIGIN
      1 malwmrllpl la
     61 lqvvgqvelgg gp
//
```

Now press Ctrl-C to copy the entry to the clipboard.

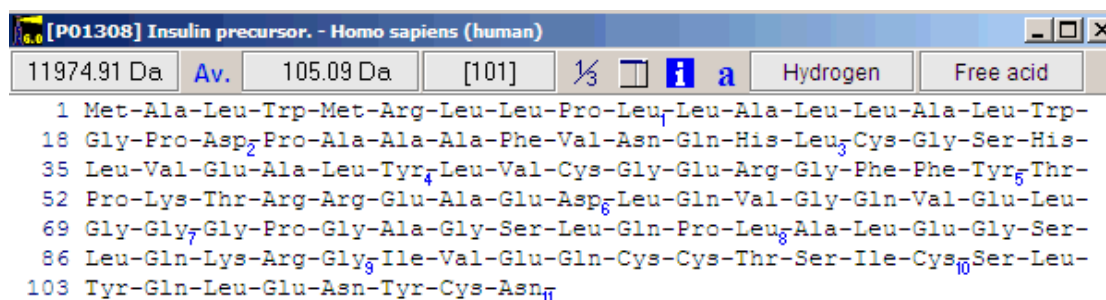
Move to GPMW and select **File|Import text (ASCII) |from clipboard**.

Hint: Alternatively you can press the "Import from clipboard" button  in the main toolbar.

The 'Import ASCII file' dialog will open with the entry from Entrez in the top edit box, please refer to the picture on the previous page.

As GPMW recognizes the GenPept format, the database entry is already parsed into "Name", "Sequence" and "Accession number". Make sure the '**Save text as annotation**' is checked in order to save the complete entry in the annotation page of the GPMW sequence.

Select '**OK**' and the entire sequence is imported into GPMW and opens a separate window.



Notice that the 'a' button in the local toolbar is blue, indicating that there is information in the annotation page (click on the button to view the complete annotation). The color of the button indicates the content type: **Gray** : no content; **Blue** : Entrez format; **Green** : Swiss-Prot format; **Red** : content, but not in a recognized format.

Now select **File|Save as** in order to save the sequence. In the 'Save sequence' dialog you enter 'human insulin' followed by 'OK'. Alternatively you can save to an already existing sequence file, thus creating a sequence library. By saving several sequences to the same library, you greatly reduce the clutter on your hard drive.

3 - Editing the sequence.

We now have the insulin precursor, but we want to work with the active form of insulin.

First we need to know where in the sequence the active part is. Click on the red 'a' button (or

GPMW for dummies

select **Info|Annotation**). This opens the annotation page that contains the complete database record from Entrez (Swiss-Prot). The interesting part is close to the bottom of the page where it reads:

FT	SIGNAL	1	24	
FT	CHAIN	25	54	INSULIN B CHAIN.
FT	PROPEP	57	87	C PEPTIDE.
FT	CHAIN	90	110	INSULIN A CHAIN.
FT	DISULFID	31	96	INTERCHAIN.
FT	DISULFID	43	109	INTERCHAIN.
FT	DISULFID	95	100	

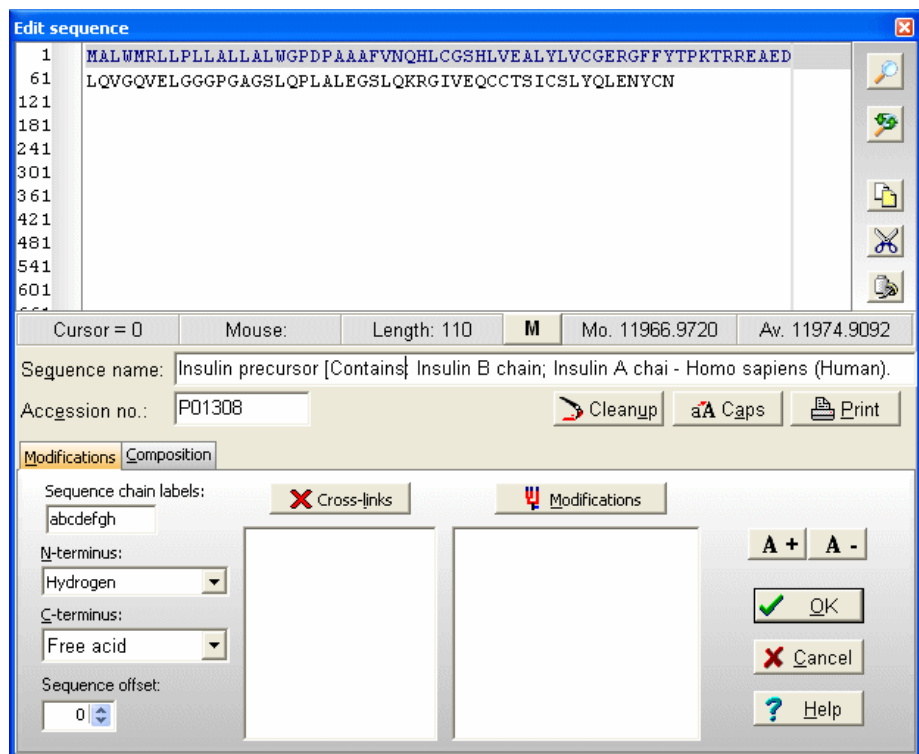
The information we need here is that the A-chain is from 39-59 and the B-chain is from 1 to 38.

There are of course several ways of making these chains, but the easiest is to start by opening the sequence editor: Select the appropriate sequence window and select **Edit|Edit sequence...**; alternatively you can right-click in the sequence window to open the pop-up menu and select **Edit|Edit sequence**.

This opens the sequence editor with the insulin precursor in the edit field.

Now we start from the C-terminus, in order not to we don't change the original numbering. The status line just below the edit box indicates the residue to the right of the text cursor. The first step is to separate the chains using the dash ('-') character, which is used as chain delimiters.

To do this, enter a dash ('-') at the end of the sequence. Move the cursor so it is between 89 and 90 and enter another dash. Do the same between 54 and 55.



Position the cursor after residue 24 and highlight to the beginning of the sequence. The editor

looks like this:

```

1 MALWMRLRLPLALLALLWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKT-RREAEDLQVGQV
68 ELGGGPGAGSLQPLALEGSLQKR-GIVEQCCTSIICSLYQLENYCN-
  
```

Delete the highlighted portion as this is not part of the mature protein chain. Highlight the middle portion from RRE to QKR- and delete it. Highlight the last peptide from GIVE to YCN- (the A-chain). Cut to clipboard (Ctrl-X or use the buttons in the right-hand control panel), move the cursor to the beginning of the line and paste the sequence. Finally move to the end of the sequence and remove the dash.

You now have the final insulin molecule:

1 GIVEQCCTSICSLYQLENYCN-FVNQHLCGSHLVEALYLVCGERGFFYTPKT]

Delete the word 'Precursor' from the name line and you are done. Select 'OK' takes you back to the sequence display. The disulfide bonds could have been entered in the sequence editor, but it is just as easy from the sequence window.

5807.65 Da	Av.	132.12 Da	[3b]			Hydrogen	Free acid	
1	Gly-Ile-Val-Glu-Gln-Cys-Cys-Thr-Ser-Ile-Cys-Ser-Leu-Tyr-Gln-Leu-Glu-Asn-Tyr-							
20	Cys-Asn-----Phe-Val-Asn-Gln-His-Leu-Cys-Gly-Ser-His-Leu-Val-Glu-Ala-Leu-Tyr-							
39	Leu-Val-Cys-Gly-Glu-Arg-Gly-Phe-Phe-Tyr-Thr-Pro-Lys-Thr-							

A few things worth noting when working with multiple chains:

The 'dash' chain delimiting character becomes three dashes in 3-letter code with a dash before and after, in total five dashes. The chains are named a, b, c etc. (counting from the N-terminus) as you can see from the cursor pointing to residue 3 in the B-chain (the third position panel in the toolbar shows [3b]). For every chain delimiter, 18 Da is added to the molecular mass relative to the single chain molecule. **Note:** in the sequence editor you can change the naming of the chains in the 'Sequence chain labels:' field, i.e. if you want l and h (for light and heavy chain) you just enter l and h as the two first characters in the field.

4 – Disulfide bonds

We still need to define the disulfide bonds. From the annotation page we looked at above, we can see that the following disulfide bonds are present: A-chain first to third, second to first on the B-chain, and the last on the A-chain to the last on the B-chain. When looking at the annotation information remember that the A-chain comes after the B-chain in the linear sequence of the precursor.

Right-click on the sequence and select **Edit | Edit cross-links** from the pop-up menu or **Edit cross-links** from the **Edit** main menu (you can also use the keyboard shortcut Ctrl + F11). As soon as the cross-link dialog opens, the Cys residues will be colored, as this is the default residue to cross-link.

The sequence now looks like this

1 Gly-Ile-Val-Glu-Gln-Cys-Cys-Thr-Ser-Ile-Cys-Ser-Leu-Tyr-Gln-Leu-Glu-Asn-Tyr-
20 Cys-Asn-----Phe-Val-Asn-Gln-His-Leu-Cys-Gly-Ser-His-Leu-Val-Glu-Ala-Leu-Tyr-
39 Leu-Val-Cys-Gly-Glu-Arg-Gly-Phe-Phe-Tyr-Thr-Pro-Lys-Thr-

To define the links, you now just have to click on the Cys residues to link in the correct order and they will be entered into the table as you click on them. If you click on the wrong residue, just click on the 'X' in the top left of the table, and the corresponding line in the table will be cleared.

If you want to link other residues, just select this in the drop-down box to the right, and click the 'Update' button to refresh the sequence window.

If you have multiple sequences where you want the same disulfide pattern (e.g. if you have multiple IgG sequences), you can save the pattern to disk and re-load it for the next sequence. The pattern is based on link-residue 1 to link-residue 4 etc, e.g. not on specific sequence positions. This enables the pattern to be transferred even if there are insertions and deletions in the sequences.

In the bottom of the window, you can select the color with which to paint the lines connecting the Cys.

X	Cys-1	Cys-2
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		

Enter residue numbers to link into table.
Alternatively click on residues to link in the sequence window.
To clear a link, click in the left-hand column.
To highlight other residues, select in residue box below

Residue to highlight:

Cysteine

Update

SS profile:

Load

Save

OK

Cancel

Help

Line colors:

Select **'OK'** and you move back to the sequence window with the defined cross-links shown in red:

```

1 Gly-Ile-Val-Glu-Gln-Cys-Cys-Thr-Ser-Ile-Cys-Ser-Leu-Tyr-Gln-Leu-Glu-Asn-Tyr-
20 Cys-Asn-----Phe-Val-Asn-Gln-His-Leu-Cys-Gly-Ser-His-Leu-Val-Glu-Ala-Leu-Tyr-
39 Leu-Val-Cys-Gly-Glu-Arg-Gly-Phe-Phe-Tyr-Thr-Pro-Lys-Thr-
  
```

Cysteines can be in the oxidized state (S-S, cross-linked) or in the reduced state (SH). This is controlled in GPMW by the SS button in the main toolbar. When the button shows **'SS'** cysteines are oxidized and the cross-links are shown in red colors (Cys is calculated with a mass of 102 Da). When the button shows **'SH'** cross-links are broken and shown as gray lines (Cys is then calculated as 103 Da). The activity of the SS button is also connected to the currently selected mass file (shown in the drop-down box next to the SS button) as Cys has to be defined as mass 102/103 Da. **Note:** the action of the **'SS'** button is global to all sequences opened in GPMW.

SS AA_MASS.MSS

Save the file using command. This will save to the same file (and position if it is a library file). If you use the **File|Save as** command to save to the same file, the sequence will be appended to the sequence file, thus generating a sequence library (if the file is not already a library). If you append it to the existing file, it will be advantageous to rename the name of the sequence to differentiate from the previously saved sequence (e.g. add 'Cross-linked' to the beginning of the name).

Note: The default state of the SS button can be defined in setup. [Setup](#)

Note: If you imported the sequence from the Swiss-Prot database, the database annotation will end up on the annotation page of the sequence. GPMW is able to interpret the 'Feature' section of this annotation, so you can import the disulfide bridges directly into to sequence with a few mouse clicks (that is if they are part of the annotation which is usually the case). The features should be imported prior to changing the sequence length. For more information please see section C.1-2, the manual and the online help.

5 – Cleaving proteins – also with linked peptides

Cleaving a protein into peptides is usually done using specific proteases or using chemistry. GPMW uses a very flexible notation that enables you to specify up to approximately 16 positions with 'required residues', 'non-cleaving residues', 'multiple independent specificities' etc. In addition you can enable 'missed cleavages', focus on a mass range, modify peptide terminals, perform deuterium exchange etc. For details check the on-line help and chapter 9 of the manual. Cleaving linked peptides generally works just like cleaving non cross-linked sequences with only a few minor differences.

The first thing to do when selecting an enzyme to use for cleavage, is to highlight the residues participating in the particular cleavage (e.g. Arg and Lys for trypsin, E for endoproteinase Glu-C, Trp, Phe and Tyr for chymotrypsin etc.). As the program use different colors for the different residues selections, you get a view of the resulting peptides. Particular regions where cleavages are difficult can usually be seen clearly (i.e. regions where cleavages are far apart or particularly close – generating very long or very short peptides that both can be difficult to separate and analyze). As in most cases, it is much easier to have an overview of the sequence when viewed in 1-letter code.

Another way is to select the **Cleavage|Cleavage analysis** dialog.

The window consist of three tabbed pages, where the first page enables you to view the peptides generated by the first 8 enzymes listed in your automatic cleavage list.

In the right-hand panel you select the cleavage reagent. The main window shows the peptides generated in individual colors.

GPMAW for dummies

The bottom panel enables you to show only some peptides, i.e. excluding small or large peptides or excluding very hydrophilic or hydrophobic peptides.

If you check the 'No limits' checkbox, all peptides will be shown.

By clicking on the different cleavage

agents you can quickly select an enzyme that gives the best-sized peptides.

The **graph** page shows the number of peptides divided into mass ranges and the **Single cleavage** page shows a peptide summary along with a sequence where highlighted residues show cleavage points.

Note: The coverage page can be saved to disk in 'Coverage analysis' format, enabling you to compare with actual sequence coverage obtained. Please see end of this handbook.

For the plasminogen analyzed in the setting above, chymotrypsin seems to be a very appropriate enzyme to use for general analysis (except that it doesn't always cleave as cleanly as several other enzymes).

You now switch back to the sequence window and click on the down-arrow next to the scissors button. This opens the "Quick-cleavage" menu where you can select among the 10 top entries in the 'Automatic digest' list. This menu does not give you many options, only '1 missed cleavage' and 'Digest all sequences' at the bottom of the menu, but this is usually sufficient. If you need more options, click on the 'scissor' button for the full options. From the drop-down menu you now select '**Chymotrypsin /W,/Y,/F-/P**', and peptide window opens:

[1] Chymotryp -> Plasminogen (EC 3.4.21.7) - Human

Mo. S 1/2 Alt i Low MS Seq. sort Sync. windows

Chymotryp [/W,/Y,/F-/P] - p1

Num	From-To	MH+	HPLC	M3H+	Alt.MS	Av/Mo	Sequence
42	497-510	1589.7918	13.05	530.6021	1588.785	1590.7324	TPETNPRAGLEKNY
78 ¹	132-145	1662.6424	21.53	554.8856	1661.635	1663.7905	CRNPDNDPQGPWCY
67 ¹	1-15	1668.7752	20.07	556.9299	1667.768	1669.7844	EPLDDYVNTQGASLF
86 ¹	214-226	1689.7737	24.76	563.9294	1688.766	1690.9492	CRNPDRELRPWCF
19	200-213	1690.9639	16.37	564.3262	1689.957	1692.0134	IPSKFPNKNLKNY
32	366-381	1727.8255	17.43	576.6134	1726.818	1728.9028	RGTSSTTTTGKKCQSW
25	264-279	1750.8316	18.70	584.2821	1749.824	1751.9455	RGNVAVTVSGHTCQHW
17	174-189	1752.8170	23.15	584.9438	1751.810	1754.0171	DGKISKTMGLECQAW
40	470-485	1775.8732	18.57	592.6292	1774.866	1776.9932	RGKRATTVTGTPCQDW
9	92-107	1837.9286	19.54	613.3144	1836.921	1839.1722	RGTMSKTKNGITCQKW
52	597-613	1857.9402	21.66	619.9849	1856.933	1859.1356	VLTAHCLKESPRPSSY
73 ¹	75-91	1960.9559	18.26	654.3235	1959.949	1962.2102	EKKVYLSECKTGDGKNY
23	235-253	2031.9150	25.12	677.9765	2030.908	2033.2896	ELCDIPRCTTPPPSSGPTY
129 ¹	761-779	2075.1205	22.22	692.3784	2074.113	2076.4555	GLGCARPKNKPGVYVRVSFF
16	156-173	2102.6905	27.57	701.5683	2101.683	2104.3044	CDILECEEECHMCSGENY
128 ¹	753-773	2215.1567	26.71	739.0571	2214.149	2216.5913	ILQGVTSWGLGCARPKNKPGVY
58	692-712	2314.2462	23.58	772.0869	2313.239	2315.7214	GAGLLKEAOLPVIENKVCNRY

The first column always shows the peptide number in the cleavage of the protein. In this case 66 peptides are generated as 'clean' cleavages, so numbers above signifies that the peptides contain a missed cleavage site. The blue superscript after the number indicates the number of missed cleavages. **Note** that peptide 19 contains a FP site that is not cleaved as the enzyme specifications are that cleavage does not take place in front of a proline residue. This is thus not counted as a missed cleavage.

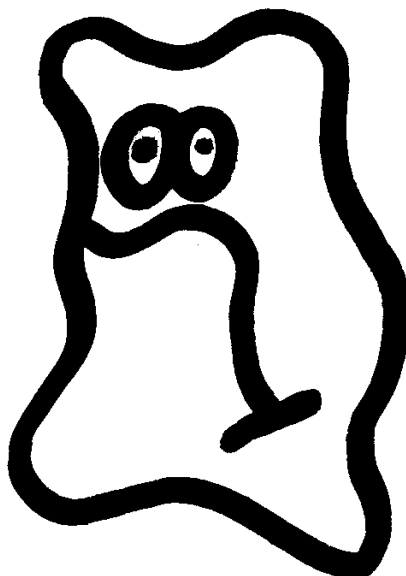
Each column contains specified information for the given peptide. Currently 19 different formats can be chosen, from mass, m/z at various charges, through pI, HPLC retention index to alternate mass tables etc. For a complete list, please refer to the on-line help, the manual and the sidebar.

Note: The actual information displayed can be configured in 'Setup'. Click on the white-on-blue button in the peptide window toolbar and the 'Setup' dialog opens on the 'Peptide' page. At the bottom are the two column layouts. Click on the 'Setup' button to configure columns. Choices are: Mass, negative charge mass (-1 or -2), positive charge mass (1 to 4), location, HPLC index, Bull & Breeze index, charge, pI, mass from alternate mass file, addition of a set mass. **Setup**

Notice in the window that the coloring of residues in the sequence window has been carried on into the peptide window. In addition to the 'normal' information, linked peptides are shown at the bottom after all non-linked peptides. The first column shows the number of the peptides linked, and the last column the mass of the peptide(s).

The **non-linked** peptides may be sorted by any column by clicking on the respective header. Click a second time to reverse the sort order. The toolbar of the peptide window gives access to a number of functions. From left to right: Change mass type; Setup peptide list properties; 1-/3-letter residue display; Alternate column display; Peptide information (select appropriate peptide first); Remove low mass peptides from list (cutoff is set in **Setup**); Show partial modifications in list; ms/ms cleavage (select peptide first); simulated HPLC chromatogram; simulated HPLC chromatogram; charge vs. pI graph (select peptide first), and isoelectric focusing gel. Checking the "Sync. Windows" checkbox will result in underlining of the selected peptide in the parent sequence window.

If you right-click in the window, you will get additional choices in the pop-up menu.



D – Post-translational modifications.

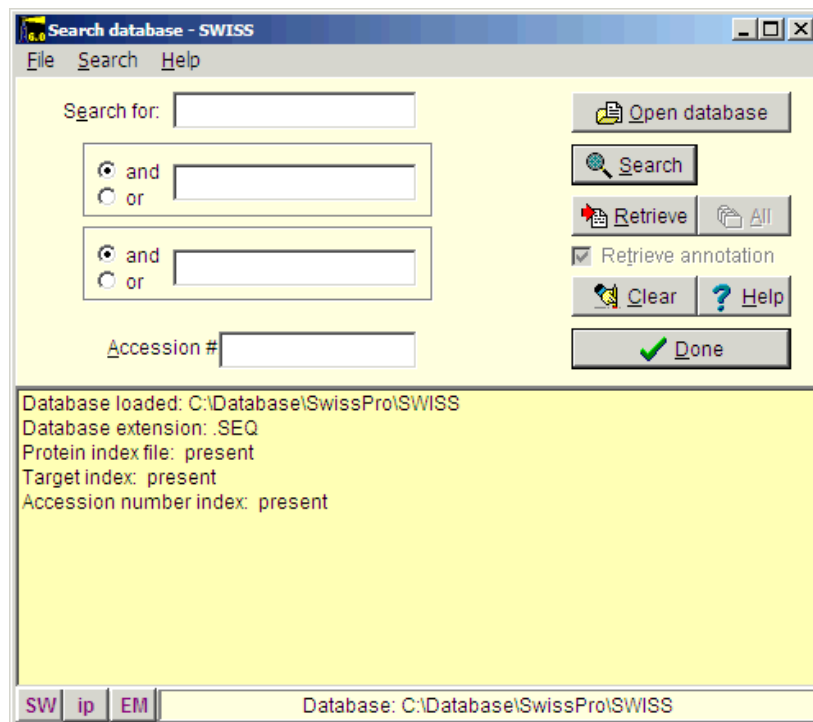
1 - Obtaining the sequence – Swiss-Prot.

GPMW for dummies


This time we will retrieve a sequence from the Swiss-Prot database. The Swiss-Prot/UniProt database is supplied on the original GPMW installation CD-ROM. The database cannot be accessed directly on the CD-ROM so you have to copy it to the hard drive using the installation program (can be done after the main installation of GPMW). If you do not have the CD-ROM, or if you need a more recent version of the database, you can download it from the Internet (e.g.

<ftp.ncbi.nlm.nih.gov>,
<ftp.ebi.ac.uk>, <ftp.expasy.ch>), but you have to convert it to FastA format and index it using the Dbindx utility before searching the database.

The main reason for using the Swiss-Prot database is that this database is the best-annotated and curated protein database. This also means that the database is less redundant and the sequences found here are more likely to be 'correct' than auto-translated sequences from a nucleotide database.



Select **File | Open**

Database | FastA (or click the FastA button  in the main toolbar) and the search database dialog box (right) will open. The first time you open the dialog it will be empty and you will have to navigate to the relevant database file (with a .trg extension) in order to get access to it. Once accessed, the program 'remembers' the database (or rather the 5 most recently accessed), and you can open it directly by either pressing the relevant button in the bottom left corner of the dialog (the buttons show the first two characters of the database name, the fly-by help shows the full name), or you can access it at the bottom of the **File** menu item.

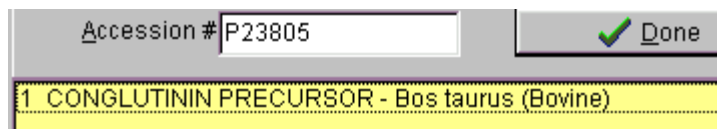
After you have installed the Swiss-Prot database from the CD-ROM, press the '**Open database**' button, and in the '**Open**' dialog you navigate to the Swiss-Prot database. When you have found the correct database directory, you select the file 'swiss.trg'.

When the 'Search database' dialog opens, you are greeted with information on the database (see above). You can see the selected database in the title (Swiss) and in the status line at the bottom of the dialog where the complete path to the database is displayed.

On future access to a FastA search, you can select **File | Open FastA database** in the main menu or the buttons at the bottom of the dialog box.

The fastest way to retrieve a sequence is by using the accession number (if known). Just enter the accession number (e.g. P23805) in the 'Accession #' field, press the search button and the name of the protein will show in the result box:

Highlight the name and press the '**Retrieve**' button and the sequence will be imported into GPMW as a sequence window. If the 'Retrieve



Tip: You may also access the database across a network, thus allowing several people to share the same databases. When copying the database make sure to copy all required files (5 files with the same name but the extensions .seq .ndx .acc .trg .fac + .idx and the annotated database in case of a swiss-prot format database (swiss, EMBL, IPI)).

annotation' tick-box is selected (it is 'on' by default), the entire database entry will be copied into the annotation page of the sequence window.

If you do not know the accession number, you will have to search on the basis of the protein name and, perhaps, species. Please remember that you are searching the FastA formatted version of the database, not the entire database (even though you will retrieve the full database entry). You are thus limited to words that are present in the name line of the database entry. This line will usually also hold the species name. If you need to search in other parts of a database entry, you will need use one of the web search engines (e.g. the EBI, www.ebi.ac.uk, or Expasy, www.expasy.ch for the Swiss-Prot database).

If you want to retrieve Bovine Coagulation factor X from cow you should enter:

You could also have entered 'factor', but the search returns so few entries that it does not matter. Normally you use 'and' for the search parameters (in this case 'coagulation and bovine'), but you can also use 'or' (works only reliably for two parameters). If you get an 'I/O error 87' your search terms are too loose and you have to narrow them. You should always put the most selective term first (e.g. if you use 'human' make it the last term).

The results of the above search returned 7 hits. Double-click on 'factor X' or highlight and press **'Retrieve'**. If you wanted to retrieve several sequences, you can hold down the Ctrl key while selecting (clicking on) multiple hits and finally select the **'All'** button to read all sequences into GPMW.

```
COAGULATION
No. of hits: 36
BOVINE
No. of hits: 1284
1 COAGULATION FACTOR XIII A CHAIN (EC 2
2 COAGULATION FACTOR X PRECURSOR (E
3 COAGULATION FACTOR XII PRECURSOR (
4 COAGULATION FACTOR V PRECURSOR (A
5 COAGULATION FACTOR VII (EC 3.4.21.21) -
6 COAGULATION FACTOR IX (EC 3.4.21.22) (
7 TISSUE FACTOR PRECURSOR (TF) (COAG
```

Select **'Done'** when you are finished retrieving sequences.

Save the sequence(s) to disk (remember you can save multiple sequences to the same file).

2 – Inserting post-translational modifications

If you have retrieved your sequence from a Swiss-Prot database as illustrated above you should have a green **'a'** in the sequence toolbar

If the **'a'** is not green and there is no information in the annotation page, you should go back and make sure that the **'Retrieve annotation'** tick-box is ticked in the 'Search database' dialog (alternatively there was no annotation to retrieve from the database, either because the full database was missing or the cross-index file was not functional). If you still do not get the annotation, your sequence database is not set up correctly and you should reinstall it (e.g. by using the install program on the CD-ROM or re-index a downloaded database). Note that you will not get an annotation when you use other databases like the EMBL-nr or NCBI-nr as these databases are originally in FastA format and does not contain other information than name, accession number and sequence.

Click on the green **'a'** in the sequence window toolbar. Looking at the secondary modifications in the annotation, they should look like this for human Factor X (2 columns)

FT SIGNAL	1	?		FT MOD_RES	46	46	GAMMA-CARBOXYGLUTAMIC ACID.
FT PROPEP	?	40		FT MOD_RES	47	47	GAMMA-CARBOXYGLUTAMIC ACID.
FT CHAIN	41	180	FACTOR X LIGHT CHAIN.	FT MOD_RES	54	54	GAMMA-CARBOXYGLUTAMIC ACID.
FT CHAIN	183	492	FACTOR X HEAVY CHAIN.	FT MOD_RES	56	56	GAMMA-CARBOXYGLUTAMIC ACID.
FT PROPEP	183	233	ACTIVATION PEPTIDE.	FT MOD_RES	59	59	GAMMA-CARBOXYGLUTAMIC ACID.
FT CHAIN	234	492	ACTIVATED FACTOR XA, HEAVY	FT MOD_RES	60	60	GAMMA-CARBOXYGLUTAMIC ACID.
FT			CHAIN.	FT MOD_RES	65	65	GAMMA-CARBOXYGLUTAMIC ACID.
FT PROPEP	476	492	MAY BE REMOVED BUT IS NOT	FT MOD_RES	66	66	GAMMA-CARBOXYGLUTAMIC ACID.
FT			NECESSARY FOR ACTIVATION.	FT MOD_RES	69	69	GAMMA-CARBOXYGLUTAMIC ACID.
FT DOMAIN	86	122	EGF-LIKE 1, CALCIUM-BINDING	FT MOD_RES	72	72	GAMMA-CARBOXYGLUTAMIC ACID.
FT			(POTENTIAL).	FT MOD_RES	75	75	GAMMA-CARBOXYGLUTAMIC ACID.
FT DOMAIN	125	165	EGF-LIKE 2.	FT MOD_RES	79	79	GAMMA-CARBOXYGLUTAMIC ACID.
FT DOMAIN	234	492	CATALYTIC.	FT MOD_RES	103	103	HYDROXYLATION.

GPMW for dummies

FT BINDING 200 200 SULFATE (IN SOME MOLECULES). FT CARBOHYD 208 208 FT CARBOHYD 218 218 N-LINKED (GLCNAC...) FT CARBOHYD 485 485 FT ACT_SITE 275 275 CHARGE RELAY SYSTEM. FT ACT_SITE 321 321 CHARGE RELAY SYSTEM. FT ACT_SITE 418 418 CHARGE RELAY SYSTEM. FT DISULFID 90 101 FT DISULFID 95 110	FT DISULFID 112 121 FT DISULFID 129 140 BY SIMILARITY. FT DISULFID 136 149 BY SIMILARITY. FT DISULFID 151 164 BY SIMILARITY. FT DISULFID 172 341 INTERCHAIN. FT DISULFID 240 245 FT DISULFID 260 276 FT DISULFID 389 403 FT DISULFID 414 442 BY SIMILARITY.
---	---

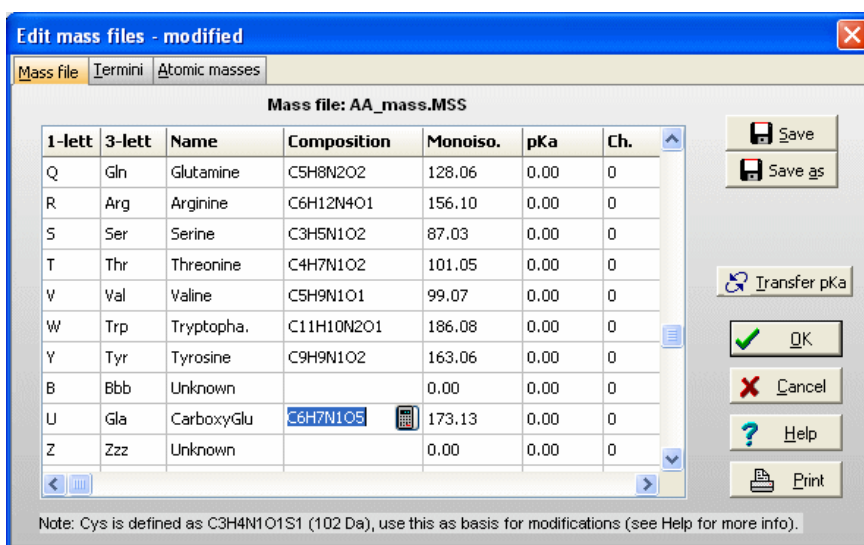
As this is a Swiss-Prot annotation, GPMW has two ways of inserting the modifications into the sequence, a manual and a semi-automatic one. As the manual one is generally applicable, it will be presented first.

Manual method for posttranslational modifications:

When you want to add post-translational modifications to your sequence, you can either add these as a 'new' residue or as an add-on modification (see below). If you have a large number of residues and/or you analyze this residue regularly, you should use 'New residue'. If you on the other hand only have a few modifications of a given kind, you should use 'Add-on modification'.

'New' residue: As there is a large number of modified residues of the same type (gamma-carboxyglutamic acid) you can use the 'extra' amino acid residue feature of GPMW and use individually modified residues for the rest.

The first task is then to create a 'new' amino acid residue in GPMW. From the main menu select **Edit|Edit mass file**. Scroll the list of residues until you come to the 'Unknown' section of the mass list.



As all 1-letter codes have to be unique (and not a punctuation mark, '-' or '\$') select the 1-letter code 'U'. The default for 'unknown' residues is that the 3-letter code is a triplet of the 1-letter code, but as it only has to be unique, so changed the 3-letter code to 'Gla', and the name to CarboxyGlu. The amino acid composition you copy from Glu (C5H7N1O3) and paste it into the composition field. Then double-click on the composition to invoke the 'Elemental composition' editor and increased the number of 'C' atoms by one, and the number of 'O' by two. Alternatively you can just edit the field directly (click twice or select and press F2).

Save the file using the **'Save'** button. If you only want the modification for special occasions, you can use the **'Save as'** button and give it a unique name. You can then select the file through the mass file selection box in the main toolbar.

Return to the sequence window and start editing (**Edit|Edit sequence**). Start by changing the 12 Glu residues (res. 46, 47 ... 79) in the N-terminal to 'U' (our new carboxyglu). Remember that the 'Cursor' field in the sequence editor reports the residue number before the cursor.


'Add-on' modification: The modifications of individual residues can either be carried out from the sequence editor or from the sequence window:

- I) In the sequence editor click on the **'Modifications'** button and in the resulting **'Insert**

Note: If you need the Gla modification both with and without a given Cys modification (e.g. a pyridylethylated cysteine – defined in *pe_cys.mss*), you have to enter the modification in both mass files (e.g. both in the default *aa_mass.mss* file and in the *cys* modification file *pe_cys.mss* (pyridylethylated *cys*)).

modification' dialog you enter the residue number. Alternatively you can double click on the residue.

- II) In the sequence window you double-click on the residue to be modified and you get the "Insert modification" dialog box.
- III) You right-click on a residue and from the pop-up menu you select 'Modify Xxx-' and from the sub-menu, you can select among pre-defined modifications.

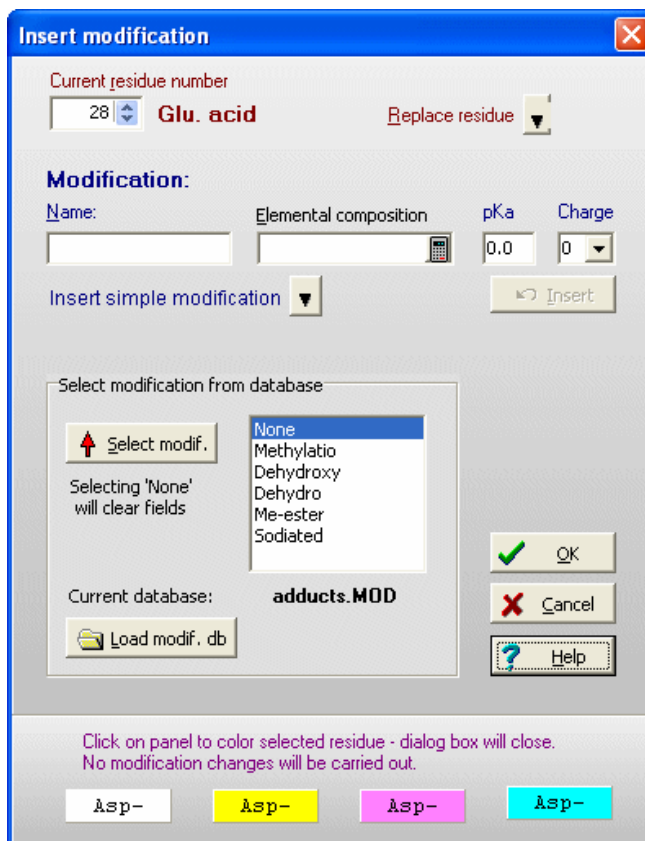
For both I) and II) you get the same 'Insert modification' dialog box (right). If the modification is in a modification database (in the example the adducts.mod file has been loaded), you can select it from the list box. If not, you enter a name (hydroxylation) and elemental composition (O1 – one extra oxygen) and click 'OK' (the composition can easily be entered with the composition calculator ). Pre-defined simple modifications can be selected using the 'Insert simple modification' drop-down list (identical to the selection in III). If you need to exchange a residue (i.e. perform a mutation) you can use the 'Replace residue' drop-down list instead of opening the sequence editor.

The panels at the bottom of the dialog box will color the residue in the sequence window using the displayed background color (and the dialog will close immediately).

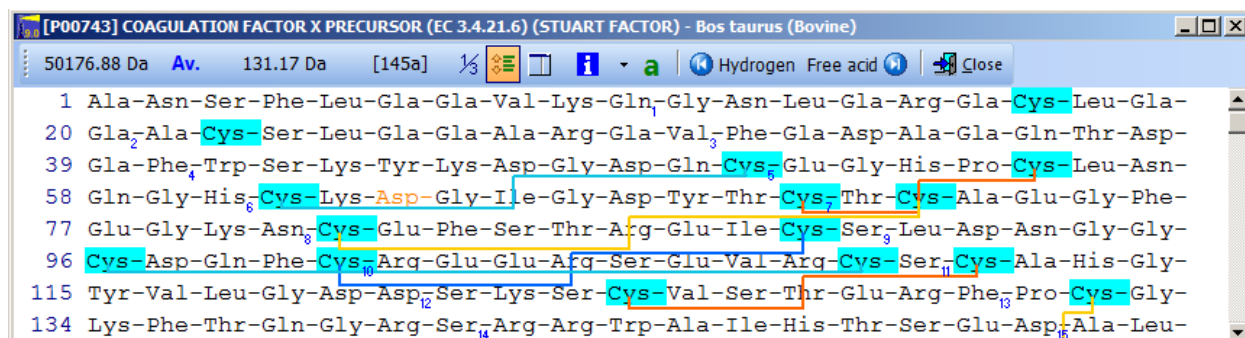
The residues in position 208 and 485 are O-glycosylated and 218 is N-glycosylated. The actual carbohydrate groups are not mentioned in the annotation, and you will have to refer to the main literature to determine the actual modification to enter in the sequence. N-linked glycosylations can be quite heterogeneous, and identifying the exact glycosylation pattern can be difficult – see the glycosylation tool in section C.3 below.

Now you can enter the disulfide links by clicking the '**Cross-links**' button as described for insulin in the previous example.

In the sequence editor you can now insert a cleavage (the dash character '-') after residue 233 (the activation peptide) and remove the initial 40 residues (the signal- and the pro-peptide). Click on 'OK' and you have your final edited sequence:



The 'Insert modification' dialog box is shown. It has a title bar with a close button. Inside, there's a 'Current residue number' field with '28' and a 'Glu. acid' label. A 'Replace residue' dropdown is next to it. Below is a 'Modification:' section with fields for 'Name:', 'Elemental composition', 'pKa' (0.0), and 'Charge' (0). There's an 'Insert simple modification' dropdown and an 'Insert' button. A 'Select modification from database' section shows a list with 'None' selected. Below this is a 'Current database:' field with 'adducts.MOD' and a 'Load modif. db' button. At the bottom are 'OK', 'Cancel', and 'Help' buttons. A footer note says: 'Click on panel to color selected residue - dialog box will close. No modification changes will be carried out.' Below the note are four colored buttons: 'Asp-' (white), 'Asp-' (yellow), 'Asp-' (pink), and 'Asp-' (cyan).



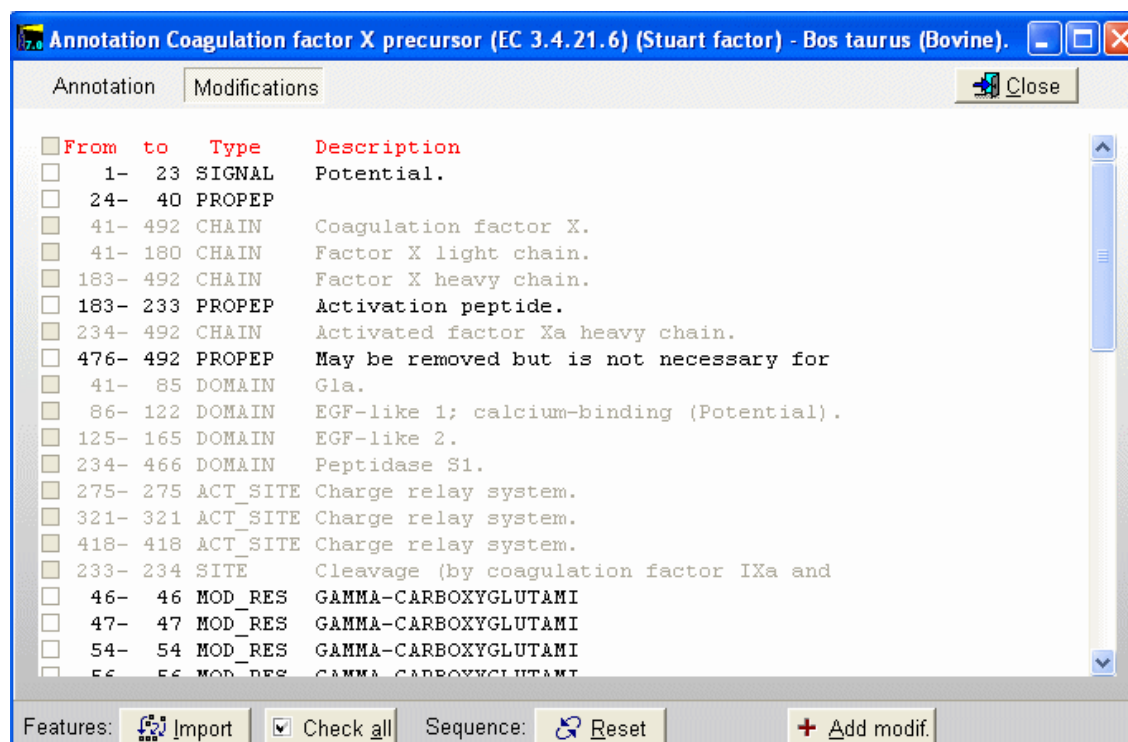
The sequence editor window shows the protein sequence: [P00743] COAGULATION FACTOR X PRECURSOR (EC 3.4.21.6) (STUART FACTOR) - Bos taurus (Bovine). The sequence is displayed with residue numbers 1 to 134. Cysteine (Cys) residues are highlighted in cyan. The sequence is: 1 Ala-Asn-Ser-Phe-Leu-Gla-Gla-Val-Lys-Gln-Gly-Asn-Leu-Gla-Arg-Gla-Cys-Leu-Gla-20 Gla-Ala-Cys-Ser-Leu-Gla-Gla-Ala-Arg-Gla-Val-Phe-Gla-Asp-Ala-Gla-Gln-Thr-Asp-39 Gla-Phe-Trp-Ser-Lys-Tyr-Lys-Asp-Gly-Asp-Gln-Cys-Glu-Gly-His-Pro-Cys-Leu-Asn-58 Gln-Gly-His-Cys-Lys-Asp-Gly-Ile-Gly-Asp-Tyr-Thr-Cys-Thr-Cys-Ala-Glu-Gly-Phe-77 Glu-Gly-Lys-Asn-Cys-Glu-Phe-Ser-Thr-Arg-Glu-Ile-Cys-Ser-Leu-Asp-Asn-Gly-Gly-96 Cys-Asp-Gln-Phe-Cys-Arg-Glu-Glu-Arg-Ser-Glu-Val-Arg-Cys-Ser-Cys-Ala-His-Gly-115 Tyr-Val-Leu-Gly-Asp-Asp-Ser-Lys-Ser-Cys-Val-Ser-Thr-Glu-Arg-Phe-Pro-Cys-Gly-134 Lys-Phe-Thr-Gln-Gly-Arg-Ser-Arg-Arg-Trp-Ala-Ile-His-Thr-Ser-Glu-Asp-Ala-Leu-

Notice that the Cys residues have been highlighted in order to better locate them. The Glu residues in the N-terminal region are correctly labeled as Gla. The modified Asp residue is red (if you move the mouse cursor over it you can see the actual modification in the top right) and

the cross-links are displayed as red lines. If you click on the 'SS' button in the main toolbar, you will 'reduce' the cysteines (the mass of each Cys will increase by one Da) and the cross-links will be grayed.

Semi-automatic method for posttranslational modifications:

If you have imported a Swiss-Prot record with full annotation, the 'a' button on your sequence window will be green. If you press this, the 'annotation window' will open showing you the full annotation. This is a two-page window that upon activation of the second page, 'Feature table', will present you with the following view:



This is the FT (feature) section of the Swiss-Prot record. The parts that are recognized by GPMW are shown in black, while unrecognized ones are grayed out. You can now check those features that you want to import into the sequence and pressing the "Import" button will close the window and transfer the modifications to the sequence window. If the program is unable to transfer some items a dialog box will inform you. Modified residues will be transferred as individual modifications of residues. Signal and propeptides will be removed from the sequence. Remember to import the residue modifications first and then modifications that change the sequence length. You may include both types in the same 'import' session, as GPMW will do the size modifications last. Once you have changed the size of the protein, GPMW will not be able to transfer residue specific modifications – in this case you will have to 'Reset' the sequence (i.e. remove all modifications and return to the sequence as listed in the annotation).

You can add your own modifications through the '+ Add modif.' button. Note that only simple modifications defined in Swiss-Prot (and GPMW) can be added. Remember to save after making changes.

3 – N-linked glycosylations

N-linked glycosylations are in MALDI mass spectra often detected by observing a mass difference of approximately 291 Da between peaks, arising from sialic acid differences between different glycosylation forms (or partial loss of sialic acid in the mass spectrometer).

Using factor X protein and doing a tryptic digest you can proceed as follows:

From the quick color menu you select '**Basic residue**' and then '**N-glycosylation**'.

GPMW for dummies

You can now easily locate both tryptic cleavage sites and the N-glycosylation sites.

```

120 Asp-Ser-Lys-Ser-Cys-Val-Ser-Thr-Glu-Arg-Phe-Pro-Cys-Gly-Lys-Phe-Thr-
137 Gln-Gly-Arg-Ser-Arg-Arg-Trp-Ala-Ile-His-Thr-Ser-Glu-Asp-Ala-Leu-Asp-
154 Ala-Ser-Glu-Leu-Glu-His-Tyr-Asp-Pro-Ala-Asp-Leu-Ser-Pro-Thr-Glu-Ser-
171 Ser-Leu-Asp-Leu-Leu-Gly-Leu-Asp-Arg-Thr-Glu-Pro-Ser-Ala-Gly-Glu-Asp-
188 Gly-Ser-Gln-Val-Val-Arg----Ile-Val-Gly-Gly-Arg-Asp-Cys-Ala-Glu-Gly-
205 Glu-Cys-Pro-Trp-Gln-Ala-Leu-Val-Asn-Glu-Glu-Asn-Glu-Gly-Phe-Cys-

```

Looking carefully at this sequence you will notice that the identification of this glycosylation may be difficult:

Asn178 is located in a fairly large peptide (4082 Da unmodified). The terminating Arg is just after the glycosylated residue and may thus interfere with cleavage. The cleavage before the peptide is a double Arg, which again can lead to heterogeneity (missed cleavages).

Taking this into account, you can simulate a cleavage (**Cleavage | Automatic digest...**). In the digest parameters select a partials level of 2 (= up to two missed cleavages) in order to include the heterogeneity in the resulting peptides.

In the resulting peptide box you can easily locate the potential N-glycated peptide as number 18.

16	140-141	261.14	1,95	2,0	9,46	SR
17	142-142	174.11	1,41	2,0	10,76	R
18	143-179	4079.92	31,38	3,9	3,61	WAIHTSEDALDASELEHYDPADLSPTESLDLLGLNR
19	180-193	1430.66	11,99	1,9	3,93	TEPSAGEDGSQVVR
20	195-199	500.31	10,46	2,0	10,35	IVGGR
21	200-242	4718.10	37,30	4,0	4,26	DCAEGECPWQALLVNEENEGFCGGTILNEFYVLTAACHLHQA

Right-click on the peptide and select '**N-glycosylation**' from the pop-up menu.

Predict Search

WAIHTSEDALDASELEHYDPADLSPTES Av. ☐ Bisecting ☐ Fucose in arm ☒ Glyco type

Peptide mass: 4082.36 Da (Average)

Complex type glycosylation

Core 892.82Da

Chains	Bare	+1Sia	+2Sia	+3Sia	+4Sia	+5Sia
None	4975.18					
Mono	5340.51	5631.77				
Di	5705.85	5997.11	6288.37			
Tri	6071.19	6362.45	6653.70	6944.96		
Tetra	6436.53	6727.78	7019.04	7310.30	7601.56	
Penta	6801.86	7093.12	7384.38	7675.64	7966.90	8258.15

Core 892.82Da + fucose 146.14Da

Chains	Bare	+1Sia	+2Sia	+3Sia	+4Sia	+5Sia
None	5121.32					
Mono	5486.66	5777.92				
Di	5851.99	6143.25	6434.51			
Tri	6217.33	6508.59	6799.85	7091.11		
Tetra	6582.67	6873.93	7165.19	7456.44	7747.70	

In the resulting dialog you now get a mass list of the most common 'complex type' glycosylations linked to the peptide in question. Check the '**Bisecting**' box to add a bisecting GlcNAc or check the '**Extra fucose**' to add an extra fucose residue to the carbohydrate chain. The '**Glyco type**' button switches the display to show masses of high mannose structures.

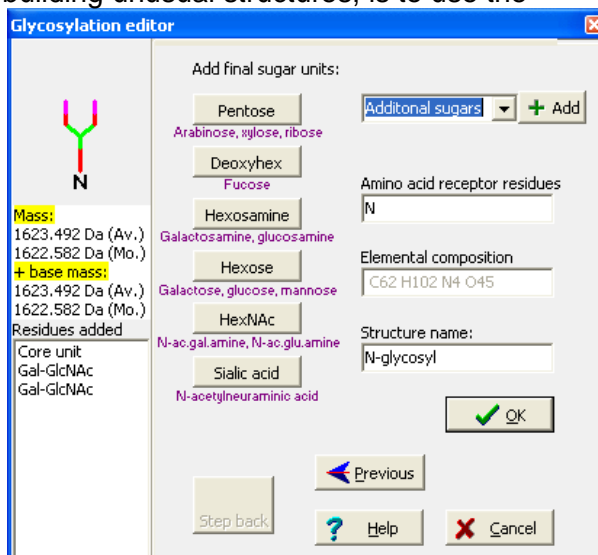
If you switch to the '**Search**' page you can use a mass list (e.g. typically from an ms/ms experiment) to search for valid glycostructures. As the number of structures is astronomical, the search is only carried out for the 'standard' types as shown above (+/- fucose) plus and minus any sugar defined in the 'sugars.mod' modification file.

GPMW for dummies

An alternative approach, which is most useful for building unusual structures, is to use the 'Glycosylation wizard' (**Seach | Glycosylation | Glycosylation wizard**) – also available in the pop-up menu for 'Simple modifications'.

In the wizard, you start by entering your base mass. This is anything attached to the carbohydrate, and can be peptide or derivatization agent. Then you select either N-linked, O-linked or other glycosylation. For the N-linked you then chose the kind of structure to extend a core structure with before ending at the editor where you can add any kind of carbohydrate defined in the 'Sugars.mod'.

Standard sugars can be entered by pressing the appropriate buttons, note that sugars are listed as base type only (e.g. Hexose covers galactose, glucose and mannose) as you cannot distinguish between these isomers using mass spectrometry. As you add residues, the 'Step back' button lights up, enabling you to 'undo' selections and rebuild your structure.



E – Getting data out of GPMW

In a field as varied as protein chemistry, it is not possible just to use a single rather specialized program like GPMW for all your protein analyses. Particularly when using the Internet with the ready availability of (free) programs, it is of interest to be able to quickly and efficiently transport protein sequences around. Another aspect is the handling of larger projects, where you typically use a word processor or spreadsheet to keep track of your data. For these programs, GPMW also have some functions that enable you to do as little handling as possible in the target program.

1 – Protein sequence.

The most obvious way of getting a sequence from GPMW to a report is to copy to the clipboard.

Select the sequence window you want to copy, press Ctrl-C or select **Edit | Copy to clipboard**. GPMW shows a timed dialog box (the box is on screen for a few seconds and terminates automatically) telling you that the sequence has been copied. This operation copies

GPMW for dummies

the sequence to the clipboard in the format on screen (1- or 3-letter code).

Note: If you have one or more peptides (regions) highlighted, only those portions of the sequence will be copied to the clipboard! Different highlighted regions will be copied as separate lines.

When you paste a sequence into a report, you should in most cases select a monospaced font for display, as the sequences will line up correctly eg.

```
Courier      AGSYLLEELFEGHLEKECWEEICVYEEAREVFEDDETDE  40
              FWRTYMGGSPCASQPCLNNGSCQDSIRGYACTCAPGYEGP  80
              NCAFAESECHPLRLDGCQHFCYPGPESYTCSCARGHKLQ  120

Arial (Swiss) AGSYLLEELFEGHLEKECWEEICVYEEAREVFEDDETDE  40
              FWRTYMGGSPCASQPCLNNGSCQDSIRGYACTCAPGYEGP  80
              NCAFAESECHPLRLDGCQHFCYPGPESYTCSCARGHKLQ 120
```

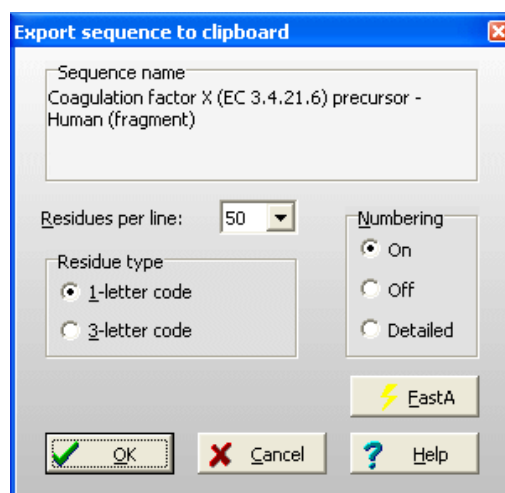
Although the Arial font is nicer to look at, it is much more difficult to find your way around a sequence. If you export in the detailed mode, things get really screwed up unless you choose a monospaced font.

The method above has one disadvantage: you only copy the sequence and not the name. In order to include the name you have to 'export' the sequence.

Select **File|Export sequence|To clipboard**.

This opens the dialog shown on the right. Here you have a number of options to format the output to fit with your report. The **Residues per line** is a dropdown box that lets you select from 10 to 100 residues per line. The **Residue type** is set as your screen display, but can be changed. **Numbering** can be selected **On** (see example above), **Off** (no numbers) or **Detailed**:

```
Protein Z - Bovine (396 res.)
      10      20      30      40
AGSYLLEELFEGHLEKECWEEICVYEEAREVFEDDETDE
      50      60      70      80
FWRTYMGGSPCASQPCLNNGSCQDSIRGYACTCAPGYEGP
      90     100     110     120
NCAFAESECHPLRLDGCQHFCYPGPESYTCSCARGHKLQ
```



The **FastA** button puts a '>' in front of the name (e.g. '> Protein Z – Bovine'), selects 60 residues per line, 1-letter code and numbering off. This makes it easy to copy a FastA formatted sequence to another program (e.g. on the Internet).

When you click '**OK**' the sequence is copied to the clipboard.

The **annotation page** has a small trick when copying to the clipboard. If you select the **Edit|Copy to clipboard** or Ctrl-V command you will always get the whole annotation copied. However, you can copy part of the annotation by highlighting the relevant portion, right-click in the window and select the '**Copy**' command from the pop-up menu.

2 – The peptide list.

The peptide list is the result of the cleavage of a protein. It is one of the main features of GPMW. This is mainly due to the fact that although the primary structure of a large number of proteins is known (mostly based on nucleotide data), the analysis of intact proteins is still very difficult. Thus you have to cleave the protein into specific smaller fragments, peptides, using specific enzymes or chemicals. Although the calculation of chemical/physical parameters like mass, pI and HPLC retention times is possible, either by hand or programs freely available on the web (e.g. www.expasy.ch), these programs are often cumbersome to work with and are often not designed for the mass spectrometrists.

The generation of the peptide list is fairly straightforward, and will not be covered here (see the manual and online help for more details).

GPMW for dummies

[1] Staph.au. -> Protein Z - Bovine						
Staph.au. [E-V] - p0						
Num	From-To	Mass	HPLC	Ch	pI	Sequence
13	87- 88	234.09	2.00	1.0	3.85	Ser-Glu-
4	16- 17	275.15	1.51	2.0	6.11	Lys-Glu-
7	28- 30	374.19	2.81	2.0	6.36	Ala-Arg-Glu-
9	34- 36	377.11	2.03	0.9	3.07	Asp-Asp-Glu-
8	31- 33	393.19	10.47	1.0	3.85	Val-Phe-Glu-
2	9- 11	407.21	14.60	1.0	3.85	Leu-Phe-Glu-
3	12- 15	454.22	9.39	2.0	5.36	Gly-His-Leu-Glu-

Once generated you may sort the list based on any of the displayed parameters by clicking on the header. The first click will sort in descending order, while clicking a second time will sort in ascending order.

Pressing Ctrl-C or selecting **Edit|Copy to clipboard** (or from the pop-up menu) will copy the entire content to the clipboard.

The copy will be just like the list displayed, so you should make sure the sorting, monoisotopic / average mass, 1-/3- letter code etc. is correct before copying.

You may copy only part of the list by selecting only some lines. You can select a continuous range of entries by clicking on the first one and holding down 'Shift' while clicking on the last one. If you hold down 'Ctrl' you can add and remove single entries from the selection. If you start by selecting a continuous stretch, you can add and remove single entries afterwards.

When you want copy only part of the list, you have to select at least two peptides, as GPMW will otherwise just go ahead and copy the whole list.

30	345-347	370.23	4.81	2.0	10.35	Val-Pro-Arg
35	369-372	416.21	4.53	2.0	10.35	Gly-Gln-Gly
33	362-365	429.23	4.48	2.0	10.35	Ala-Ser-Pro
28	318-320	440.21	3.70	3.0	7.21	Glu-His-Arg
16	202-205	523.32	9.82	3.0	10.55	Leu-His-Val
15	198-201	545.27	10.43	3.0	9.92	Ser-His-Phe
8	118-122	587.30	9.49	1.9	5.94	Leu-Gly-Gln
25	296-300	593.24	9.94	2.0	6.36	Thr-Ser-Cys
26	301-308	643.34	5.70	2.0	10.35	Gly-Ala-Ala
21	268-273	743.40	13.60	2.0	6.36	Glu-Met-Val
12	150-156	747.38	7.14	2.0	6.36	Leu-Thr-Asn
17	206-212	804.44	8.23	4.0	10.36	Gly-Val-His
31	348-353	820.46	21.76	2.0	9.00	Tyr-Ala-Leu
10	124-130	825.37	15.37	3.0	6.75	Ser-Cys-Leu
27	309-317	913.51	15.51	2.0	10.35	Trp-Val-Ala
36	373-381	920.43	9.68	1.9	9.07	Asn-Glu-Glu

☐ Copy table to clipboard

☐ Copy table as text

☒ Copy tab delimited

☒ Limited sequence

☐ Full sequence

Two entries in the **Peptide|Setup** are important when copying: copy as text vs. tab delimited and copy full sequence vs. limited sequence.

When you copy as text, each column is separated from the next by a space character. This makes it easy to align columns if you use a monospaced font (e.g. Courier). If you select 'tab delimited', each column is separated from the next by a 'tab' character. This means that you have to set the tabs properly in the report. E.g.

Table as text delimited (**Courier** font):

Num	From-To	Mass	HPLC	Ch	pI	Sequence
33	362-365	429.23	4.48	2.0	10.35	Ala-Ser-Pro-Arg-
28	318-320	440.21	3.70	3.0	7.21	Glu-His-Arg-
16	202-205	523.32	9.82	3.0	10.55	Leu-His-Val-Arg-
15	198-201	545.27	10.43	3.0	9.92	Ser-His-Phe-Arg-

Arial font:

Num	From-To	Mass	HPLC	Ch	pI	Sequence
33	362-365	429.23	4.48	2.0	10.35	Ala-Ser-Pro-Arg-
28	318-320	440.21	3.70	3.0	7.21	Glu-His-Arg-
16	202-205	523.32	9.82	3.0	10.55	Leu-His-Val-Arg-
15	198-201	545.27	10.43	3.0	9.92	Ser-His-Phe-Arg-

Table as tab delimited (**Arial** font):

Num	From-To	Mass	HPLC	Ch	pI	Sequence
33	362-365	429.23	4.48	2.0	10.35	Ala-Ser-Pro-Arg-
28	318-320	440.21	3.70	3.0	7.21	Glu-His-Arg-
16	202-205	523.32	9.82	3.0	10.55	Leu-His-Val-Arg-
15	198-201	545.27	10.43	3.0	9.92	Ser-His-Phe-Arg-

Courier font:

Num	From-To	Mass	HPLC	Ch	pI	Sequence
33	362-365	429.23	4.48	2.0	10.35	Ala-Ser-Pro-Arg-
28	318-320	440.21	3.70	3.0	7.21	Glu-His-Arg-
16	202-205	523.32	9.82	3.0	10.55	Leu-His-Val-Arg-
15	198-201	545.27	10.43	3.0	9.92	Ser-His-Phe-Arg-

When you copy columns to a **spreadsheet** (e.g. Excel) you should always use the '**Copy tab delimited**' as this will transfer columns to individual columns. You can set your spreadsheet up to accept space-delimited columns, but this is fraught with errors.

Instead of copying the complete table, you may be interested in only copying a few of the columns. You could go into **Peptide|Setup** and change the layout of the peptide table, but it is much easier to right-click in the table and select **Copy|Export|Copy columns to clipboard** from the pop-up menu. In the copy to clipboard dialog box, you can then select the columns to copy. The title for each tick-box is taken from the actual header of the peptide table.

The 'Sequence' column is always selected.

Like when copying the complete table you can select a range of peptides before you start the copy operation.

3 – Mass search results

The results of a mass search are reported in a two-page window. The **first** page (**Analyze**) displays the 'hits' in a tabular format that enables you to fine-tune the results by changing the displayed properties, change precision, perform recalibration etc. The **second** page (**Report**) displays a report based on the selections made on the first page.

Copying the results presented on the **first** page works very much like the peptide list described above. The main difference lies in the selection of lines to report. Where the peptide list is a standard multiple selection list, the results of the mass search is a check box selection list.

This works in the way that if no lines have been selected (checked) the whole list is copied. If one or more lines have been checked you are asked whether you want only the selected lines copied (Yes – selected lines only; No – whole list; Cancel – cancel copy operation).

The '**Check**' button atop the check boxes, works to check/uncheck all lines in a single operation.

The individual check boxes can be checked/unchecked by clicking on them with the left mouse button. Alternatively you can use the arrow keys to move up and down the list and use the space bar to check/uncheck lines (usually faster than using the mouse).

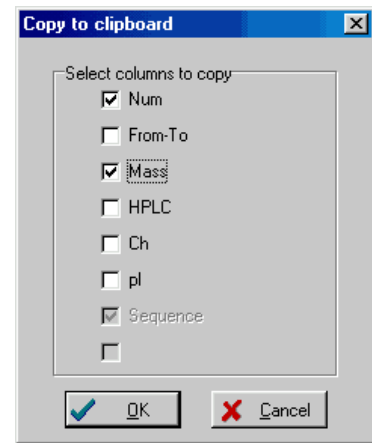
A shortcut exists in the pop-up menu to check all peptides that fits with the cleavage pattern of the enzyme used in the search (**Selected peptides|Check perfect fits**). Another option inverses all selections (e.g. unchecks all checked items and visa versa - **Selected peptides|Toggle selections**).

Unlike the peptide list, you cannot select individual columns for transfer.

Depending on how you make out your report and whether you copy to a spreadsheet, you have to set the **Peptide|Setup** correctly (see 'Peptide list' above).

The **second** page (**Report**) contains two scroll boxes, where the top one displays the sequence, and the bottom one statistics and the identified peptides. The top box shows the sequence and the identified peptides in color, which are not preserved when copied to the clipboard. Here the format changes to show the sequence in lower case with the cleavage residues (blue on screen) in upper case. The identified peptides are shown as double underlines.

Although some information and the easy navigation of the screen is lost most of the essential information is transferred. **Note:** you can copy the coverage map presented below for a much clearer display.



```

GLSDGEWQQVLNVWGKVEADIAGHGQEVLRIRI
1814.90

LKKHGTUULTALGGILKKKGHHEAEKPLAQSL
1350.63 1377.83

GDFGADAQGAMTKALELFRNDIAAKYKELGFI
1501.66 747.43

Residue coverage: 80% [123 of 153]
Peptide hits: 10 Modified: 0 Not identified

- Peptides identified without modifications:
input found dev. mc from-to sequen
748.384 / 747.428 69 0 134-139 ALELFR
1271.634 / 1270.656 23 0 32- 42 LFTGHP
1351.568 / 1350.634 54 1 51- 62 TEAEMK
1378.798 / 1377.834 32 0 64- 77 HGTVVL
1502.619 / 1501.662 33 0 119-133 HPGDFG
1606.863 / 1605.847 -5 0 17- 31 VEADIA
1815.849 / 1814.895 29 0 1- 16 GLSDGE

glsdgewqqvlnvvgkveadiaghgqevlrlftghpetleKf
=====

lKKhgtvultalggilKKKGhheaelkplaqshatKhKipiKy
== ===== =
=====

gdfgadaqgamtKalelfrndiaaKyKelgfqq 153
=====

== unmodified, -- modified

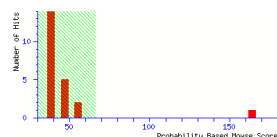
Residue coverage: 80% [123 of 153]
Peptide hits: 10 Modified: 0 Not identified

- Peptides identified without modifications
input found dev. mc from-to se
0 748.384 / 747.428 69 0 134-139 AL
0 1271.634 / 1270.656 23 0 32- 42 LF

```

4 – Presenting sequence coverage.

One of the common and tedious jobs for a protein chemist (except for presenting 1500 annotated ms/ms spectra as an appendix) is to present the sequence coverage of an enzyme digest in order to show that you have done a good job in characterizing a given protein.



Concise Protein Summary Report

Format As: Concise Protein Summary Help

Significance threshold p<: 0.05 Max. number of hits: 20

Re-Search All Search Unmatched

1. [MYG EQGBU](#) Mass: 16941 Score: 164 Expect: 9.6e-12 Queries matched: 11
 Myoglobin - Equus burchelli (Plains zebra) (Equus quagga)

In GPMW you get a sequence coverage when performing a mass search (see previous section), but this coverage has a couple of shortcomings:

- 1) Beauty is not one of its main attributes, as it is meant for ease and compactness
- 2) Often you have additional information you may want to incorporate (e.g. you may find additional peptides manually or you may have multiple digests that you want to combine in a single figure).

In addition to the primitive ‘automatic’ coverage above, GPMW have the option to make nice coverage maps, which can easily be edited and exported to reports, both in Word, PowerPoint and other programs.

A coverage map in GPMW consists of a sequence and up to eight levels. Each level consists of a number of peptides, each of which is defined by first and last residue number. The numbering (and thus the level) is not tightly coupled to a sequence, so you have to be careful not accidentally to paste a level into the wrong sequence. Each peptide may further have a label (16 characters) and a comment (40 characters). The peptide mass is not saved along with the peptide in the coverage map, but is calculated based on the sequence, current mass file, and the peptide limits. Furthermore, each level has an associated color that can be edited from the ‘Edit level’ dialog box.

To work with a coverage map you select **Utilities| Coverage analysis** from the main menu, which opens a window with a large empty field and a right-hand toolbar.

Coverage map:

Load new Paste level Save map

Levels Edit level

#	Name	Col
1	Mass search results -	Red
2	Mass search results -	Blue
3		
4		
5		
6		
7		
8		

Labels

☐ None ☒ Label ☐ Mass ☐ Comment ☐ First-last

Copy Done

Myoglobin - Equus caballus (Horse), and Coverage: 105/ 68.6% [Single 88/ 57.5%] Mass search results - abrf_horsemyo.PEP

GPMW for dummies

Starting a coverage map:

Manually: Click on the down-arrow in the 'Load new' button. From the drop-down menu select 'Sequence from desktop', and you can now choose any sequence that is currently open on the desktop.

Select a level in the 'Levels' table and click on the 'Edit level' button. This opens a dialog box with a table where you can enter start, end, label, and comment for each peptide.

As this is rather tedious, and as you are likely to have the data on electronic form anyway, a number of shortcuts are available.

If you have copied an intact level to the clipboard from the mass search window, you can paste it using the 'Paste level' button.

It is usually more interesting to import a table from most tabular listings:

Making a peptide mass fingerprint search using the Mascot search engine (www.matrixscience.com) you may get a result like this with a clear hit.

Matched peptides shown in Bold Red

1 GLSDGRNQVQV LIPWGRVAD IAGHGGVLI RLTFHPETL ERFDFKHLK
51 TEARMKASRD LKSHQTVVLT ALQILKQKQ RHEARLKPLA QSHATKHKIP
101 IKYLEPISDA IIVHLSKHP GDFGADAQA MTKALELFRN DIAARYKELG
151 FQG

Show predicted peptides also

SortPeptides By ☒ Residue Number ☐ Increasing Mass ☐ Decreasing Mass

Start	End	Observed	Mr (expt)	Mr (calc)	Delta	Miss	Sequence
1	16	1815.8490	1814.8417	1814.8951	-0.0534	0	- GLSDGRNQVQV LIPWGRV
17	31	1606.8630	1605.8557	1605.8474	0.0083	0	K V EAD IAGHGGV LIR L
32	42	1271.6340	1270.6267	1270.6557	-0.0290	0	R LPTGHPET LK F
51	62	1351.5680	1350.5607	1350.6337	-0.0729	1	K TEARMKASRD LK
64	77	1378.7980	1377.7907	1377.8343	-0.0436	0	K HGTAVLTALGGILK K
79	96	1982.0440	1981.0367	1981.0493	-0.0126	1	K KHHHEARLKPLAQSHATK H
80	96	1853.9450	1852.9377	1852.9543	-0.0166	0	K GHHEARLKPLAQSHATK H
103	118	1884.9890	1883.9817	1884.0145	-0.0328	0	K YLEPISDAIIVHLSK H
119	133	1502.6190	1501.6117	1501.6619	-0.0502	0	K HPGDGFADAQAGAMTK A
119	133	1518.6130	1517.6057	1517.6568	-0.0511	0	K HPGDGFADAQAGAMTK A Oxidation
134	139	748.3840	747.3767	747.4279	-0.0512	0	K ALELFR N

No match to: 732.4720, 993.4030, 1262.5730, 1486.0880, 2010.0630, 2394.1270

Load the protein in question into GPMW (use the accession number to retrieve it from the Internet, or copy and paste from the detailed information window).

Click on the sequence accession number link, and from the 'detailed information' window, you now highlight and copy the table to the clipboard:

Go back to GPMW and press the 'Paste table' in the 'Edit coverage level' dialog. GPMW will now parse the table into columns in a new dialog box. In order to import the table, you have to define the column that contains the 'from' and 'to' values. This is done through the spin edit controls in the right-hand panel. GPMW will make a guess (first integer column as 'from' and the second as 'to') and the selected columns will be highlighted. The label and comments can also be selected as columns.


Select 'OK' to transfer to the Edit level dialog. If the peptides overlap, you have the option of dividing the peptides into separate levels. The different levels are indicated by different colors in the Edit level dialog. Note that only three levels can be created this way. However, if the third level contains overlapping levels, you can edit this level to create multiple levels (if levels are available).

When you have a coverage map, as in the example above, the calculated coverage in residues and in percent is displayed in the footer. The first value is in residues, and the second is in percent. In sharp parenthesis the single peptide coverage values are displayed. The sequence

is coloured according to coverage: Yellow: no coverage, red: single coverage, white: multiple coverages.

If you copy the coverage to the clipboard, you can paste it directly into Word, Powerpoint and other programs that accept vector formats. As it is a vector display, you can scale it without losing any resolution (i.e. magnifying the picture will not end in large pixels). Furthermore, in Powerpoint you can ungroup the picture after converting it into a Microsoft Office drawing (just right-click and select 'Grouping|Ungroup'). You then have complete control over text and drawings and can add any kind of embellishment you like, i.e. changing the color of individual residues, adding circles, arrows etc. Remember to re-group the picture when done as the picture otherwise easily becomes 'un-stuck'.

Other ways of obtaining a sequence coverage: From the **peptide window**, you can save a coverage map of the entire peptide digest – right-click in the window and select 'Copy special' from the pop-up menu and in the resulting dialog box you select 'Copy as coverage file'.

From the **Mass search** window you can save the coverage through the 'Save' button  on the **report** page.

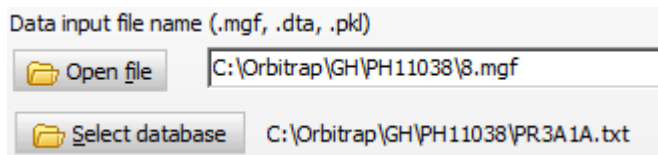
Chapter F – Ms/ms search of two proteins.

1 – Preparations.

The background for the analysis is an analysis of the interaction of alpha-1-antitrypsin and proteinase 3. A gel-band is observed at the position of PR3, but it is suspected that it may be a-1-a. Instead of just analysing by MALDI, we also want to have as good a sequence coverage as possible, so we analyse it by ms/ms using an Orbitrap XL after tryptic digestion.

We want to use the X!Tandem search engine of GPMW, and as there are only two proteins involved, there is no need to search the large database.

Search data: The data are collected by a short 30 min separation using nanoLC, and the data are converted into a peak list. An mgf file format is chosen as this contains a bit more information than dta or pkl. The data are saved in a separate directory.




Data input file name (.mgf, .dta, .pkl)

Open file C:\Orbitrap\GH\PH11038\8.mgf

Select database C:\Orbitrap\GH\PH11038\PR3A1A.txt

Getting the proteins: The accession number of each of the two proteins (AAH96186 and P01009) are entered in main toolbar retrieval box and the sequences retrieved from the web.

Saving the proteins: As X!Tandem reads FastA formatted files, the two sequences are saved to file by selecting File | Export sequence | all sequences as FastA file. The file is saved with the name PR3A1A.TXT in the same directory as the ms/ms data to keep the project organized.

Setting up the search: The ms/ms search dialog is called through Search | MS/MS Search (an alternative would be pressing the F5 key or the shortcut  in the toolbar).

The files created are now selected either through the 'Open' buttons, or simply by drag-and-drop from File Explorer.

The enzyme chosen is trypsin, the output file name is left at default, the 'Search crap list' (contaminations) is checked.

GPMW for dummies

The parameters are chosen to suit the Orbitrap XL, in particular precision is set at 5 ppm (an initial search of a previous file had shown that the instrument was properly calibrated), fragment error at 0.4 Da, e-value at 0.01, maximum charge at 4 and missed cleavage at 2.

As the search is quite fast (on my machine 3 sec) you can play around with the parameters to get the best results. If you do not change the 'output file name' the result files will overwrite each other and you will not clutter up your disk with temporary data. If you want to save your results, change the output file name when changing parameters. When you load a new search file, the name will be changed automatically.

Oxidation of Met and Carbamidomethyl modification of Cys are chosen as variable and fixed modifications.

Variable modifications [1]	Fixed modifications [1]
Oxygen [M] 15.99	Carbamidomethyl [C] 57.02
Methylation [DE] 14.02	Carboxymethyl [C] 58.01
Phospho [STY] 79.97	PyridylCys [C] 105.06
thr_ala [T] -30.01	

Parameters:	
Parent, mono. error:	- 5.0/+ 5.0 ppm
Fragment, error:	0.4 Daltons
Max. valid expect value:	0.010
Minimum ion count:	4
Scoring ions:	Y A B
Spectrum	
Minimum number peaks:	13
Maximum number peaks:	99
Maximum parent charge:	4
Minimum parent m+H:	300
Minimum fragment m/z:	146
Protein	
Max. missed cleavage:	2
Cleavage N-term mass change:	1.011
Cleavage C-term mass change:	17.003
N-term residue mod. mass:	0.000
C-term residue mod. mass:	0.000

'Enable refinement' is chosen, with 'Semi cleavage' checked, as we may find unusual cleavages (i.e. the gel band is at a lower Mr than expected for a-1-a).

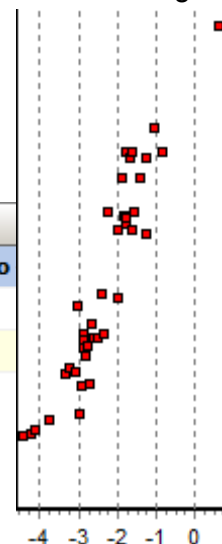
2 – Results.

The results of the search show that alpha-1-antitrypsin is likely to be the major component in the gel band analyzed.

#	log(e)	hits	uniq	name
1	-250.3	40	40	RecName: Full=Alpha-1-antitrypsin; AltName: Full=Alpha-1 pro
2	-176.3	28	28	Proteinase 3 [Homo sapiens]. - Homo sapiens (human)...
3	-32.8	6	0	SERUM ALBUMIN PRECURSOR - Bos taurus (Bovine)

The precision of the parent ions are quite good, although the calibration is not perfect (between -4 and 0 ppm for Alpha-1-antitrypsin, right).

Alpha-1-antitrypsin is the first hit with the major score, so it is clear that the gel-band contains a major proportion of this protein. A little surprisingly bovine serum albumin is also on the hit list, so now the search goes in for finding the contamination.



Double clicking on the first line with a-1-a, fills out the result page and displays the sequence

```

MPSSUSWGIL LLAGLCCLUP USLAEDPQGD AAQKTDTS HH DQDHPTFNKI TPNLAFAFS 60
LYRQLAHQSN STNIFFSPUS IATAFAMLSL GTRADTHDEI LEGLNFNLTE IPEAQIHEGF 120
QELLRTLNPQ DSQQLTTGN GLFLSEGLKL UDKFLEDVKK LYHSEAFUN FGDTEEAKKQ 180
INDYVEKGTQ GKIU DLKEL DRDTUFALUN YIFFKGKWER PFEUKDTEEE DFHUDQTTU 240
KUPMMKRLGM FNIQHCKKLS SWULLMKYLG NATAIFFLPD EGKLQHLENE LTHDIITKFL 300
ENEDRRSASL HLPKLSITGT YDLKSULGQL GITKUFNGA DLSGUTEAP LKLSKAUHK 360
ULTIDEKGT E AAGAMFLEAI PMSIPPEVKF NKPFUFLMIE QNTKSPLFMG KUUNPTQK
  
```

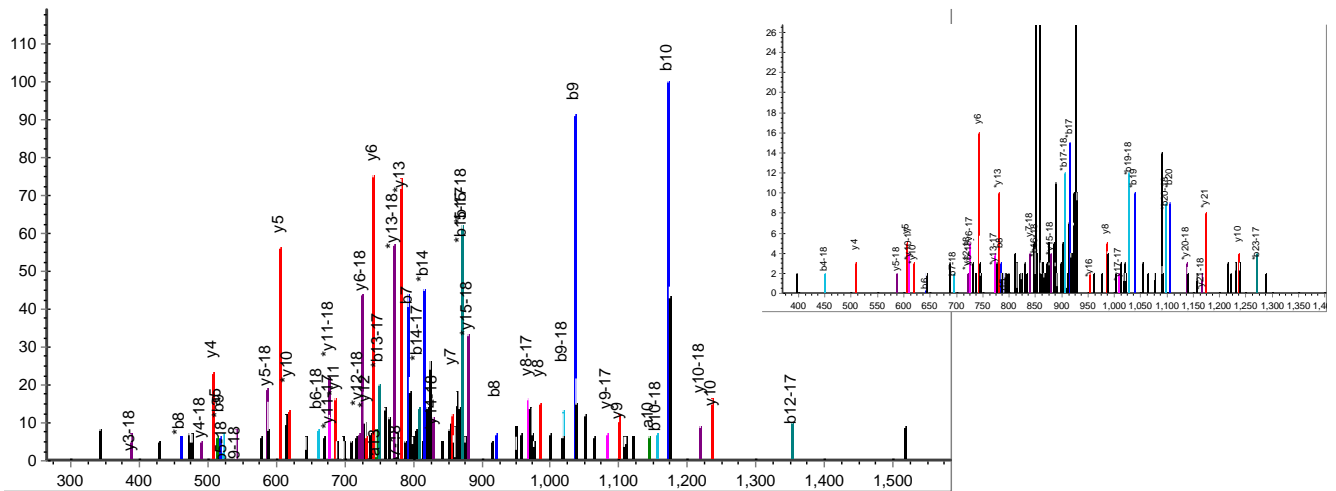
with identified peptides shown in red.

From the annotation page of the sequence the reported signal sequence is 1-24, which fits perfectly with the first peptide starting at residue 25. However, comparing with runs of tryptic digests of the intact protein, a large number of peptides are missing up to residue 150. If the protein started at residue 150 or slightly earlier, the resulting mass of the protein would be just over 30 kDa, which would fit with the relative molecular weight observed in the gel. The reason

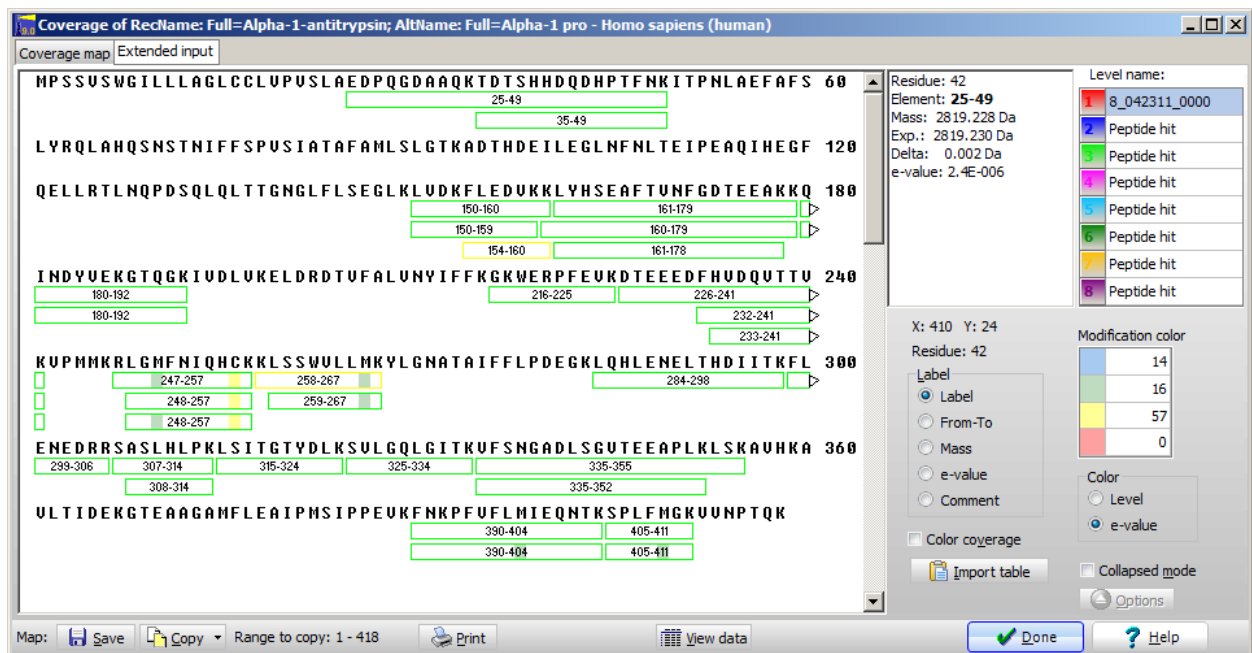
GPMAW for dummies

for the first peptide to be observed may be caused by the presence of three histidines, giving the peptide high proton affinity and thus making it easy to observe.

The ms/ms spectrum of the 25-49 peptide is not convincing, having low intensity and the major peaks not identified (below right), however, the 35-49 peptide is difficult to refuse (below left).



Clicking on the 'Sequence' button opens the coverage window, which shows a detailed picture of the peptides identified.



The conclusion must be that we most likely have a truncated form of alpha-1-antitrypsin cleaved at residue 150 or slightly before, but it is 'contaminated' by a small amount of full-length protein which demands further analysis.

Note: When you perform a number of ms/ms searches, the results are saved in xml files in the search directory. These can be re-opened through the 'Open result file' button, or more easily by drag-and-drop from the File Explorer. If you want to compare with another search directly, you can open a second search window and load the result file into this.

Atomic mass values

Name	Abbr.	Monoiso.	Average
Hydrogen	H	1.0078250	1.00794
Carbon	C	12.0000	12.011
Nitrogen	N	14.0030740	14.00674
Oxygen	O	15.9949146	15.9994
Flour	F	18.99840322	18.99840322
Phosphor	P	30.9737634	30.97376
Sulfur	S	31.972018	32.066
Chlorine	Cl	34.968852721 36.96590262	35.452737
Iodine	I	126.904473	126.904473

Mass values of the commonly occurring amino acid residues

Name	3-lett.	1-lett.	Compos.	Monoiso.	Average	
Alanine	Ala	A	C ₃ H ₅ NO	71.03711	71.0788	
Arginine	Arg	R	C ₆ H ₁₂ N ₄ O	156.10111	156.1875	
Asparagine	Asn	N	C ₄ H ₆ N ₂ O ₂	114.04293	114.1038	
Aspartic Acid	Asp	D	C ₄ H ₅ NO ₃	115.02694	115.0886	
Cysteine	Cys	C	C ₃ H ₅ NOS	103.00919	103.1448	NB! Cys-H
Half-cystine	Cys	C	C ₃ H ₄ NOS	102.00137	102.1369	NB! Cys-S
Glutamic Acid	Glu	E	C ₅ H ₇ NO ₃	129.04259	129.1155	
Glutamine	Gln	Q	C ₅ H ₈ N ₂ O ₂	128.05858	128.1307	
Glycine	Gly	G	C ₂ H ₃ NO	57.02146	57.0519	
Histidine	His	H	C ₆ H ₇ N ₃ O	137.05891	137.1411	
Isoleucine	Ile	I	C ₆ H ₁₁ NO	113.08406	113.1594	
Leucine	Leu	L	C ₆ H ₁₁ NO	113.08406	113.1594	
Lysine	Lys	K	C ₆ H ₁₂ N ₂ O	128.09496	128.1741	
Methionine	Met	M	C ₅ H ₉ NOS	131.04049	131.1986	
Phenylalanine	Phe	F	C ₉ H ₉ NO	147.06841	147.1766	
Proline	Pro	P	C ₅ H ₇ NO	97.05276	97.1167	
Serine	Ser	S	C ₃ H ₅ NO ₂	87.03203	87.0782	
Threonine	Thr	T	C ₄ H ₇ NO ₂	101.04768	101.1051	
Tryptophan	Trp	W	C ₁₁ H ₁₀ N ₂ O	186.07931	186.2132	
Tyrosine	Tyr	Y	C ₉ H ₉ NO ₂	163.06333	163.1760	
Valine	Val	V	C ₅ H ₉ NO	99.06841	99.1326	