

---

# **GPMAW 9.1**

---

Users manual  
Rev. 05-11

## **GPMaw version 9.1**

### **GPMaw**

General Protein/Mass Analysis for Windows

#### **Users manual**

Lighthouse data

Engvej 35

DK-5230 Odense M

Denmark

#### **Tech Support**

Fax: (+45) 6619 3396

E-mail: [php@bmb.sdu.dk](mailto:php@bmb.sdu.dk)

Home: <http://www.gpmaw.com>

We have done our best to insure that the material found in this publication is both useful and accurate. However, please be aware that errors may exist in this publication, and neither the authors nor Lighthouse data make any guarantees concerning the accuracy of the information found here or in the use to which it may be put. As the program is continuously being upgraded, please check the on-line help for changes made to the program.

The entire risk of the use of the result of the use of this software and documentation remains with the user. No part of this documentation may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without written permission from Lighthouse data.

# Contents

<b>1</b>	<b><a href="#">Introduction</a></b> .....	<b>1</b>
	What is GPMaw 1.1 ¶ Installation 1.2 ¶ Program layout 1.3 ¶ Starting GPMaw 1.4 ¶ Essential tables 1.5 ¶ Window menu 1.6 ¶ Recent program changes 1.7 ¶ Help 1.8 ¶ Lighthouse data 1.9	
<b>2</b>	<b><a href="#">Reading and saving sequences</a></b> .....	<b>23</b>
	Open sequence 2.1 ¶ How to acquire sequences 2.2 ¶ Save sequence 2.3 ¶ Delete sequence 2.4 ¶ Import ASCII 2.5 ¶ Reading sequences from a database 2.6 ¶ Retrieve sequences from the Internet 2.7 ¶ Export sequence 2.8 ¶ Protein Explorer 2.9	
<b>3</b>	<b><a href="#">The sequence window</a></b> .....	<b>45</b>
	Sequence window display 3.1 ¶ Highlight sequence 3.2 ¶ Highlight residues (motifs) 3.3 ¶ Underline residues 3.4 ¶ Cross-links 3.5 ¶ Amino acid modifications 3.6 ¶ Fragment window 3.7 ¶ Sequence information 3.8 ¶ Annotation 3.9	
<b>4</b>	<b><a href="#">Edit</a></b> .....	<b>75</b>
	Edit sequence 4.1 ¶ Edit mass files 4.2 ¶ Edit modification files 4.3 ¶ Composition formulas 4.4	
<b>5</b>	<b><a href="#">Setup</a></b> .....	<b>91</b>
	Setup system parameters - System 5.1 ¶ Peptide parameters 5.2 ¶ System colors 5.3 ¶ System directories 5.4 ¶ Digest mass search parameters 5.5 ¶ Display 5.6 ¶ BLAST 5.7 ¶ Users 5.8 ¶ Setup proxy 5.9	
<b>6</b>	<b><a href="#">Mass search / FastA files</a></b> .....	<b>107</b>
	Search for masses 6.1 ¶ Mass differences 6.2 ¶ Search for cross-linked peptides 6.3 ¶ Various searches ¶ Find mass in FastA database 6.4 ¶ Search FastA for motif 6.5 ¶ Mass list matching 6.6 ¶ Peptide list mass search 6.7 ¶ .mgf file filtering 6.8 ¶ FastA file handling 6.9	
<b>7</b>	<b><a href="#">Search for composition / BLAST / N-linked glycans</a></b> .....	<b>139</b>
	Search for composition 7.1 ¶ Local BLAST 7.2 ¶ ClustalW multiple alignment 7.3 ¶ N-linked glycans 7.4	
<b>8</b>	<b><a href="#">Database mass search / MS/MS search</a></b> .....	<b>155</b>
	Introduction to digest mass search 8.1 ¶ Setup digest mass databases 8.2 ¶ Digest mass search - data input 8.3 ¶ Digest mass search - status window 8.4 ¶ Digest mass search - results 8.5 ¶ Multiple digest mass search 8.6 ¶ Combine digest mass search 8.7 ¶ Protein mass search 8.8 ¶ MS/MS search 8.9 ¶ MS/MS data input 8.10 ¶ Search Parameters 8.11 ¶ MS/MS search results 8.12 ¶ Analyzing individual hits 8.13 ¶ Running the BSA example 8.14	
<b>9</b>	<b><a href="#">Cleavages / Coverage map</a></b> .....	<b>191</b>
	Automatic digest 9.1 ¶ Other cleavages 9.2 ¶ Cleavage analysis 9.3 ¶ Peptide window 9.4 ¶ Import peptide list 9.5 ¶ Coverage map 9.6	

## GPMaw version 9.1

<b>10</b>	<b><a href="#">Fragmentation</a></b>	<b>223</b>
	MS/MS fragmentation 10.1 ¶ Ladder sequencing 10.2 ¶ Graphical fragment mapper 10.3	
<b>11</b>	<b><a href="#">Graphs</a></b>	<b>233</b>
	Common graph commands 11.1 ¶ Secondary structure prediction 11.2 ¶ Hydrophobicity 11.3 ¶ Primer multiplicity 11.4 ¶ User graph 11.5 ¶ Dot-plot 11.6 ¶ Alpha-helical wheel 11.7 ¶ Charge vs. pI 11.8 ¶ DigestAlyzer 11.9	
<b>12</b>	<b><a href="#">Utilities</a></b>	<b>251</b>
	MS peak analysis - sequence tag and mass difference 12.1 ¶ Composition calculator 12.2 ¶ Fragment analyzer 12.3 ¶ Database indexer (DBindex) 12.4 ¶ Simulated 2D gel 12.5 ¶ Coverage analysis 12.6	
<b>Appendices</b>		
<b>A</b>	<b><a href="#">File formats</a></b>	<b>281</b>
	Files used by GPMaw A.1 ¶ Mass list input files A.2	
<b>B</b>	<b><a href="#">Databases</a></b>	<b>291</b>
	Databases for sequence retrieval B.1 ¶ Databases for index and mass search B.2 ¶ File formats B.3	
<b>C</b>	<b><a href="#">Tables</a></b>	<b>293</b>
	Mass types C.1 ¶ Atomic masses C.2 ¶ Amino acid residues C.3 ¶ Modified residues C.4 ¶ Post-translational modifications C.5 ¶ Carbohydrates C.6 ¶ pI values C.7 ¶ Peptide residue mass values < 304 Da C.8	
<b>D</b>	<b><a href="#">Configuring GPMaw startup</a></b>	<b>305</b>
<b>E</b>	<b><a href="#">License agreement</a></b>	<b>309</b>
	<b><a href="#">Index</a></b>	<b>310</b>

---

### Keyboard shortcuts

<b>F1</b>	Help*	<b>F7</b>	Hydrophobicity graph
<b>F2</b>	Open	<b>Sh F7</b>	Secondary structure
<b>Sh F2</b>	Save	<b>Ctrl F7</b>	Dot-plot
<b>F3</b>	Edit sequence	<b>F8</b>	Automatic digest
<b>Sh F3</b>	Edit new sequence	<b>Sh F8</b>	Manual digest
<b>Ctrl F3</b>	Edit mass file	<b>Ctrl F8</b>	Make fragment window
<b>F4</b>	Highlight residues	<b>F9</b>	Sequence info
<b>Ctrl F4</b>	Close daughter window*	<b>Sh F9</b>	Annotation
<b>Alt F4</b>	Close program*	<b>F10</b>	Menu
<b>F5</b>	Ms/Ms search	<b>Sh F10</b>	Setup system
<b>Ctrl F5</b>	Cascade windows	<b>F11</b>	Digest mass search
<b>Sh F5</b>	Tile windows	<b>Sh F11</b>	Edit modifications
<b>F6</b>	Mass search	<b>Ctrl F11</b>	Edit cross-links
<b>Sh F6</b>	Composition search	<b>F12</b>	MS peak analysis
<b>Ctrl F6</b>	Next daughter window*	<b>Ctrl F12</b>	Edit modification file

---

\*Standard Windows command

---

## Introduction

Getting started with GPMaw (General Protein/Mass Analysis for Windows).

### **What is GPMaw**

**1.1**

As the name implies, GPMaw is a program for the analysis of protein primary structures, particularly using mass spectrometry.

The initial purpose of the program, back in the late 80'es, was to answer the simple question: given a certain peptide mass how many peptides can be constructed that fit with a mass within the given precision (see Chapter 6). The answer to this trivial question is easily solved with a computer while it is rather tedious and error-prone using a calculator.

Since then the program has expanded in a number of directions, but most of these are characterized by the fact that they are based upon previous knowledge of a known sequence. This has necessitated the interfacing to protein databases, with all the problems of exponentially growing data-input (from less than 10,000 proteins in 1988 to more than 700,000 proteins today, EMBL non-redundant database).

Having loaded or entered your sequence (GPMaw interfaces to a number of other file formats and hardware), the sequence can be post-translationally modified and cross-links established (Chapter 3). The protein can be cleaved based on a number of parameters (using enzymes or chemicals) and the resulting peptides can be sorted, viewed, and further characterized based on a number of parameters (Chapter 9).

One of the most recent advances in mass spectrometric protein identification, digest mass search, is shown in Chapter 8, but also features like ms/ms fragmentation are covered (Chapter 10).

When performing mass spectrometry on proteins and peptides you are very likely to perform additional protein chemical procedures. For this purpose a number of features have been included: Searching a protein for amino acid composition, displaying the hydrophobicity index, reversed phase retention data of peptides, predicting secondary structure, comparing sequences with dot-plots, charge vs. pH graphs of peptides and proteins, digest HPLC chromatogram and mass spectra of peptide mixtures, etc.

### Pre-installation notes



**Note:** As the installation system is likely to change over time, particularly the actual content on the distribution media, you should always check the installation package for additional information.

GPMAW is presently (May 2011) distributed on a single CD-ROM containing the program and all auxiliary files including protein databases (see Appendix B for more information on databases). Alternatively, the program may be shipped on a USB stick or downloaded from the web site <http://www.gpmaw.com>. From here you can also download updates for free for up to 18 month after purchase.

The installation takes place by running the setup program from Windows. In the case of CD-ROM based installation:

If you have auto-notification turned on, the installation program will start as soon as you insert the CD-ROM. If you do not have auto-notification turned on, you will have to open the file explorer and navigate to your CD-ROM drive and start the 'Setup' program. Alternatively, you can select the '**Start**' button. From the menu select '**Run...**'. In the 'Run' dialog box you enter '**D:\SETUPGPMAW**' and select '**OK**' (assuming that 'D:' is your CD-ROM drive). Follow the instructions.

The file on the CD-ROM called Setup.exe is the installation program. This file contains all the components for installation except for gpmaw.lcs, which is the separate license file, also present in the root directory. WizardUI.exe is a wizard that runs just after the installation and prompts you for the basic settings of GPMAW.

The installation program also contains some protein databases, the Swiss-Prot (UniProt) database and a couple of IPI databases. The databases cannot be accessed on the CD-ROM, but have to be installed. They can be installed across a network on a shared drive for access by several persons.



**NOTE:** The Swiss-Prot database is copyrighted. If you are part of a commercial company you need a license in order to use the database. For non-commercial users (e.g. universities) there is no license requirement.

### Installation

The default directory for the installation of GPMAW is C:\GPMAW. It is recommended that you choose this directory for installation of GPMAW, as future upgrades and faultfinding will be simplified.

The program starts with checking for previously installed components, specifically it checks for the location of the gpmaw3.exe program file. If the program is found, its location will be taken as the default for an upgrade, otherwise c:\gpmaw will be taken as the default.

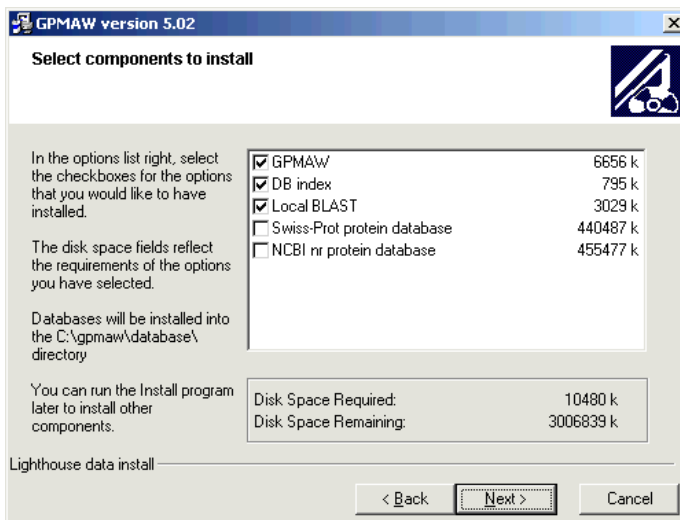
The installation starts with a splash screen (a lighthouse) followed by a welcome dialog. **Note:** the actual installation of files will not take place until all dialog boxes are handled (the last box ends with a 'Finish' button).

## 1 - Introduction

You will then be asked to acknowledge the End User License Agreement before proceeding with the installation.



**Hint:** If you want to perform an additional installation, you can rename the \gpmaw\bin\ directory to something else (e.g. \gpmaw\bin1\ ) before running the installation program. In this case you have to specify a different destination directory.



The next dialog box gives you the following options:

- GPMaw:** Installs GPMaw. This option is selected by default – you may deselect it if you only want to install a database.
- DBindex:** Installs the FastA database indexing utility. For more information, please see Chapter 12.4).
- Local BLAST** Installs a local copy of NCBI BLAST program. The program works on a local FastA formatted database. Details are presented in Chapter 7.2.
- Swiss-Prot:** Copies the Swiss-Prot/UniProt database to a directory on your harddisk (default C:\gpmaw\database\). **Note:** Commercial use of this database demands a separate license obtainable from <http://www.ebi.ac.uk>.
- IPI-databases:** Copies the human, rat and mouse IPI databases to your harddisk. The IPI databases (from [www.ebi.ac.uk/IPI/](http://www.ebi.ac.uk/IPI/)) aims to provide a minimally redundant yet maximally complete sets of proteins for the features species. The database format is similar to Swiss-Prot. Are also available for zebrafish, arabidopsis and chicken.

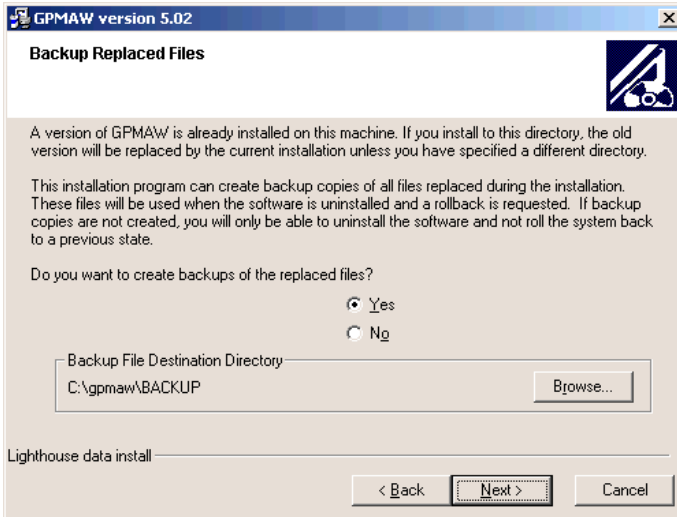


**Notice:** The actual databases included in the compilation may change over time.

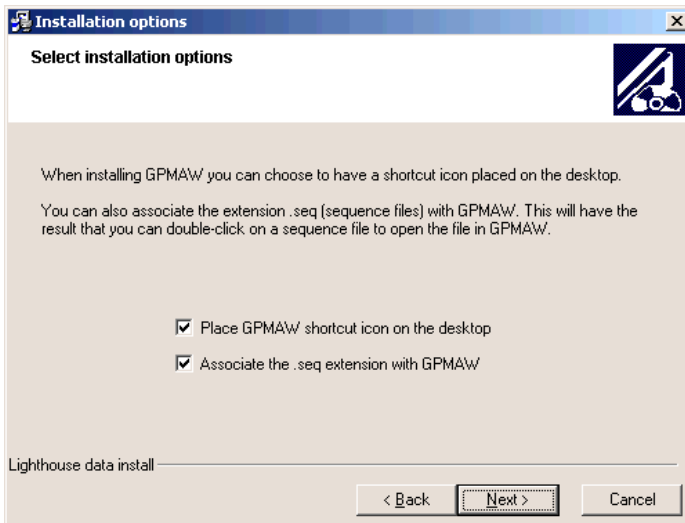
## 1 - Introduction

Note that when you an item in the list box is selected the combined size necessary for installation is displayed below the check-box.

If a previous version of GPMaw was detected in the initial part of the installation, the following screen will give you the option of saving all files replaced during the installation in a directory called \gpmaw\backup\). The main files are gpmaw3.exe and gpmaw.hlp. Restoring these to the \gpmaw\bin\ directory will in most cases restore previous installation.



If the installation program did not detect a previous installation of GPMaw it will jump directly to the following options:





## 1 - Introduction

**Place a GPMW shortcut icon on the desktop.** A shortcut will be created on the desktop. Double-clicking on this will open GPMW.

**Associate the .seq extension with GPMW.** When you associate the .SEQ extension it means that when you double-click on a file in Explorer that has the extension .SEQ GPMW will automatically open and read the file.

You are then asked for a startup folder for the program. As default a name of 'GPMW' is suggested, but you can also choose another name, or select an existing startup folder from the list below the edit line.

The next (last) dialog box informs you that installation will proceed when you press the 'Finish' button.

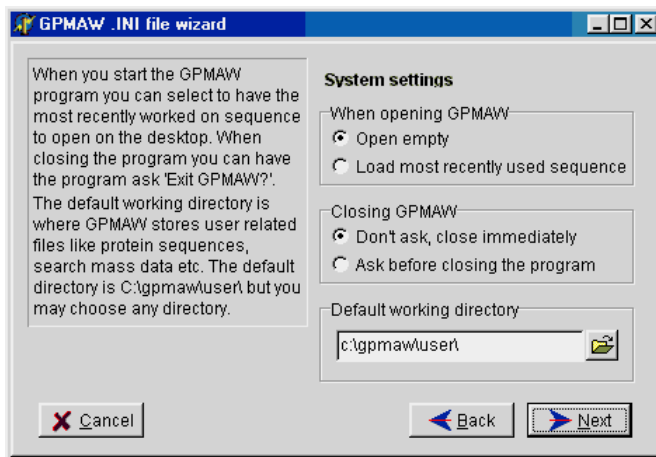
### The GPMW setup wizard

When the program has been installed, the setup wizard takes over. This program quickly walks you through the basic settings of GPMW. You can at any point select '**Cancel**' whereupon the default settings of GPMW will take effect. Only when you run through the complete wizard, the settings chosen on the various pages will take effect.



**Notice:** All the settings in the wizard can be set at any point in the program by selecting **Setup | Setup system** (see Chapter 5 for details).

The individual pages in the wizard work with '**Back**' and '**Next**' buttons that enable you to go both forward and backwards in order to customize GPMW.



The left-hand side of each page gives a short description of the functions, further details can be found in the respective chapters (sequence – Ch. 3; peptide – Ch. 9; general setup – Ch. 5).

### Errors during installation

If something goes wrong during installation, you may still have the right programs copied to the hard disk, but in the wrong places. Please see Appendix A for directory structure and file information.

## 1 - Introduction

If you are unable to reconstruct the program, try to delete all parts of the program and perform a re-installation. If you still do not have a functional installation, contact Lighthouse data (Chapter 1.9) either by fax or e-mail.

### Un-installation – removing the program

The best way to remove GPMaw from your system is to open the Control Panel and select the 'Add/Remove Programs' applet. This shows you a list of installed programs. From the list you select GPMaw and click on the 'Remove' button. All files and registry entries created during installation will be removed, only files created after installation (e.g. sequences etc.) will remain to be removed manually.

Alternatively you can uninstall manually by activating the uninstall program called `unwise.exe` located in the `\gpmaw\` folder.



As all files are installed under the same directory (`C:\GPMaw`) you can remove everything by deleting this directory and everything beneath it (remember to move valuable sequence files first). The shortcuts on the desktop and/or the Start menu may have to be removed manually.

### Upgrading

Starting with version 3.0 the program has been made 'upgradeable' compared to previous versions. This means that if you can connect to the Internet you will be able to download upgrades free of charge within 12-18 month from purchase. In this way, we hope to make bug fixes and improvements immediately available to users without the usual hassle and time delays. Please check the GPMaw web site <http://www.gpmaw.com> for more details. If this site is unavailable contact the author on `php@bmb.sdu.dk`. Free upgrades are available for at least one year after purchase, but may be available for a longer time.

When you buy an upgrade, you will get a full installation while downloaded upgrades contain only the `gpmaw3.exe` and `gpmaw3.hlp` files. In the case of the full installation your existing mass and modification files may be overwritten if the names coincide with files from the installation. In this case the installation program should warn you, but the safest course is to make a backup of your `\gpmaw\system` directory before upgrading. In the case of a 'raw' upgrade, only the `.exe` and the `.hlp` files are replaced.

When you upgrade, you will have an installation program like that for a normal installation except that the 'Desktop icon' option will be checked, and there will be an extra option 'Upgrade only'. If you un-check this option you can specify a different installation directory for the upgrade (`gpmaw3.exe` and `gpmaw3.hlp`). You should only do this if you know you have an installation that is not detected by the installation program, or if you want the new files in a separate location.

When you buy an upgrade, it will always be accompanied by a manual, while upgrades downloaded from the Internet will have to do with the on-line help.

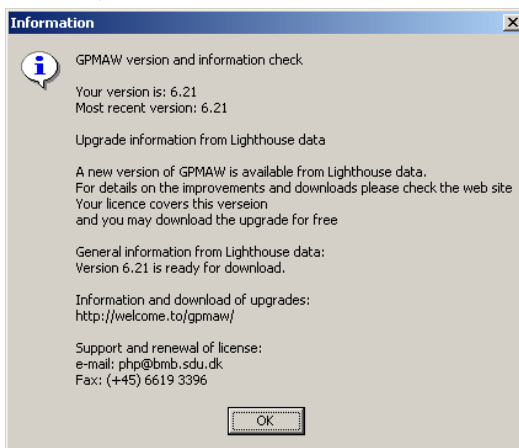
The frequent upgrades have the downside that the printed documentation will often be slightly out of date. In this case you are referred to the on-line documentation, which will be kept up to date and available along with the

## 1 - Introduction

upgrades. Pressing the <F1> key or selecting 'Help' from the main menu can access the online help. <F1> will usually give context-sensitive help that is help information pertaining to the dialog box currently open.

### Auto upgrade notification

From version 5.02 GPMAW will have an automatic upgrade notification. Every 20<sup>th</sup> time the program is started, it will contact the GPMAW web server (if your computer is connected to the Internet) and read a small text file containing update and error corrections. Each new message on the server will have a unique number, and if the number has been read previously, GPMAW will just ignore the message. If the message is new (i.e. not read before), it will be displayed to the user:



The dialog will show information on the latest version of GPMAW, whether your copy of GPMAW is upgradeable (i.e. the license has not yet expired) and whether there are any error messages.

The dialog is similar to the one you obtain by pressing the 'Info' button in the **Help | About** dialog (Chapter 1.8).

### Problems during upgrading

When GPMAW is upgraded, the first time it is started, it will sometime complain about a missing file. As the program will in most cases auto-generate a default copy of this file, you should close the program and restart it. If the problem persists, you may contact Lighthouse data (Chapter 1.9).

If, after performing an upgrade, you are informed that your license is no longer valid you will have to 'downgrade' to your previous version of GPMAW. You may then contact Lighthouse data to obtain a renewed license (an upgrade). The upgrade price of GPMAW is 50% of the current price for a full version. The upgrade includes a copy of the most recent manual.

### Conventions

Angle brackets '<' and '>' are used to denote function and special keys. <F1> to <F12> are the function keys, <Esc> the escape key, <Tab> the tabulator


## 1 - Introduction

key, <up>, <down>, <left>, <right> the arrow keys and <Ins> (insert), <Del> (delete), <Home>, <End>, <PgUp> (page up), <PgDn> (page down) the page control keys. <Enter> is the 'enter' or 'return' key.

When the main or sequence/daughter window has the focus you can always access the menu by pressing <F10> or you can access individual menu items by pressing <Alt> + the underlined letter in the menu. When a dialog box has the focus, you can move between controls using <Alt> + the underlined letter of the text describing the control or by using the <Tab> key to cycle through all controls. Once a control has the focus, indicated by highlighting or a stippled line around the associated text, you can activate it using the space bar if the control is a toggle function (button, check box etc).

Menu selections like selecting the menu option 'Edit sequence' from the main menu 'Edit' are shown as: **Edit|Edit sequence**.

A large number of notes, hints and tips are partly framed and marked with the

 symbol in the margin.

**Mouse functions:** Controls are activated in the usual manner using the left mouse button after positioning the mouse cursor on the control. Most windows also have an associated local menu, which is activated by positioning the mouse cursor inside the dialog box or control and pressing the right mouse button. Commands can then be activated by using the left mouse button as usual, or by pressing the underlined character.

**References** are either shown in brackets [ ] or listed at the end of each chapter.

### Program layout

1.3

#### Main content

The GPMW program is built around a few central themes:

**Protein sequence management:** Reading and importing protein sequences from a variety of sources. Modifying the sequences (cross-links, chemical modifications etc.) and saving the sequences to local files for easy retrieval. The sequence can be viewed and colored in various ways for easy navigation.

**Protein cleavage:** Cleavage of the protein into specific peptides by enzymatic or chemical means. From the resulting peptide list a large number of parameters can be calculated and the list can be viewed, sorted and displayed in various ways.

**Mass search:** Given a peptide mass, where in a given protein can it be found. This question can be expanded to search for modifications, ion types etc.

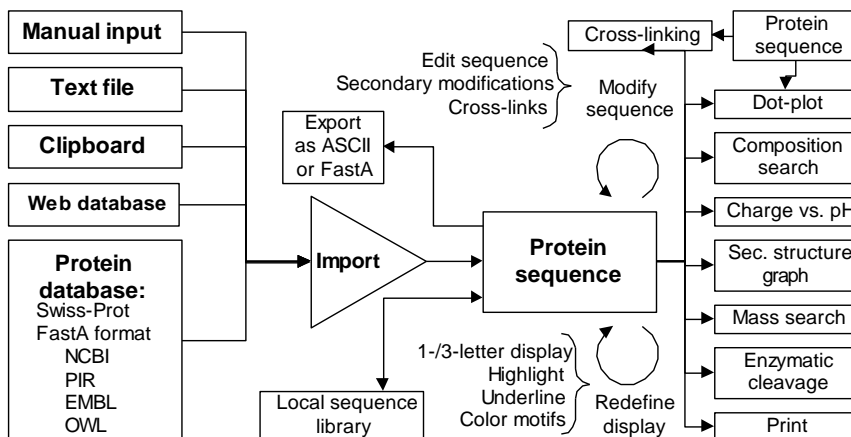
**Database mass search:** Given a list of peptide masses that originate from a single or a few proteins, what protein(s) does it originate from? A large number of options can be specified, both when searching and when analyzing the results. You may also search a database for a protein of a given mass.

## 1 - Introduction

**Ms/ms search:** Using an external search engine (Xtandem!) you can search a sequence or database using a list of ms/ms fragments in .mgf or .pkl format.

**Various:** A number of functions which are helpful in protein analysis, but not directly related to the above groups. This comprises functions like composition search, charge vs. pH graph, dot-plot analysis, BLAST searches, ClustalW multiple sequence alignment etc.

### Protein sequence management



The simplest way of getting a protein sequence into GPMW is to enter it directly in the sequence editor (Ch. 4.1). Although perfectly feasible, it is a tiresome and error prone undertaking. The easiest alternative is to import directly from the web (Ch. 2.7), but you may also download a database in FastA or Swiss-Prot format for even speedier access (Ch. 2.6). If you are already in possession of a protein sequence, you can transfer it directly through the clipboard or as a text file (Ch. 2.5). The main criterion for importing a sequence into GPMW is that it has to be in standard 1-letter code.

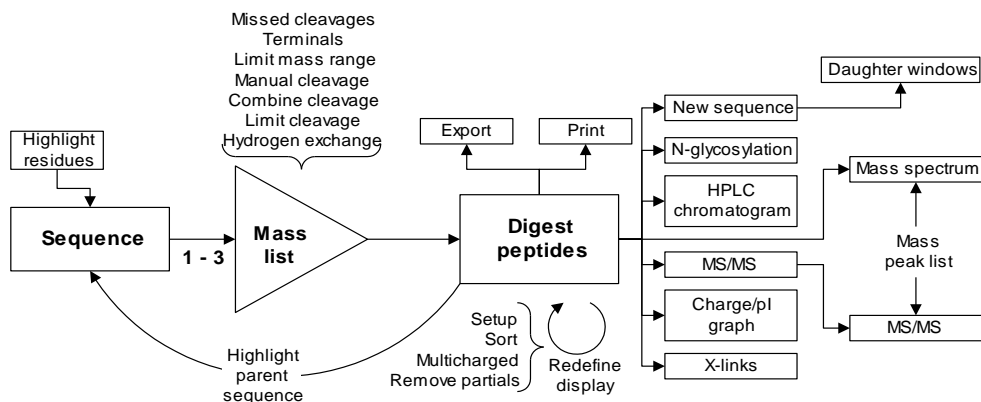
Once a sequence is entered into GPMW you should save it to a file on disk. You have the choice of either saving a single sequence to a file or saving several sequences to the same file, resulting in a sequence library. The sequence is saved in a format proprietary to GPMW (the format is explained in Appendix A) and contains in addition to the name and the sequence, information on cross-linked residues (Ch. 3.5), modified residues (Ch. 3.6) and a free text annotation page (Ch. 3.9).

In order to navigate the sequence without effort, you can display in either 1- or 3-letter code, you may color specific residues or sequences (Ch. 3.3) and you may underline residues. Furthermore, you may highlight parts of the sequence to locate peptides and to calculate coverage.

The sequence window is the basis for most other functions in GPMW as indicated in the figure above and in Chapter 3.

# 1 - Introduction

## Protein cleavage



The protein cleavage is based upon a sequence window and a cleavage specification in a particular GPMW notation (Ch. 9). A large number of proteolytic enzymes and chemical cleavages are pre-defined, and additional ones can easily be added. The cleavage can be modified in a number of ways like limiting the mass range, modifying the new terminals, combining cleavage specifications etc.

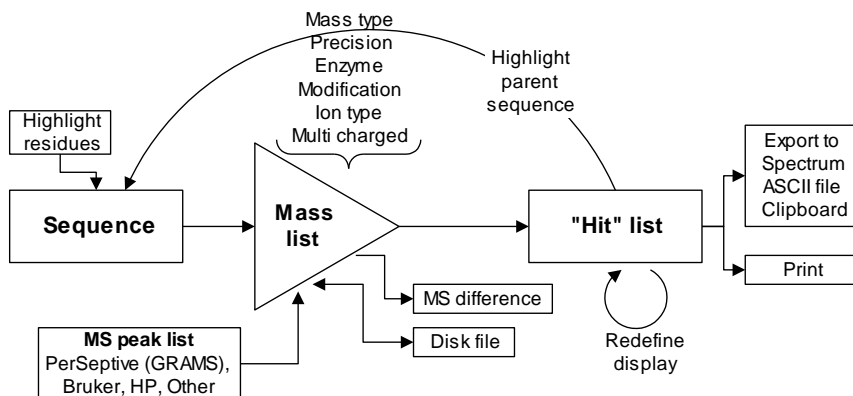
The results of a cleavage is a peptide list which is displayed in a separate window. The peptide list is shown with a number of parameters and can be sorted by mass, charge, number etc., displayed in 1- or 3- letter code, printed and exported to other programs. You can even select to have a peptide from the list as the basis for a new sequence window.

In order to easily navigate the peptide list you can color residues or sequence parts of the parent sequence window which will be 'carried on' to the peptide window. In the other direction you may underline a sequence in the parent sequence based on a selected peptide in the peptide window.

Based on the peptide window (Ch. 9.4) a number of other windows can be generated. This comprises simulated reversed phase HPLC chromatograms, N-glycosylated mass values, MS/MS of selected peptides, simulate a mass spectrum and compare it to a recorded experimental spectrum, cross-linking mass spectrometric experiments etc.

## 1 - Introduction

### Mass search



The mass search function tries to answer the question: 'Where in this protein can I find this mass'. Like the protein cleavage above, this function is based on a sequence window. You then supply a mass list, either with manual input, read from a disk file or transferred through the clipboard. GPMW is able to read the peak files from a number of mass spectrometric analysis programs.

The search parameters can be modified in a number of ways to cater for different ms instruments (mass type, precision, ion type) and sample parameters (enzyme, modifications).

The results are displayed in a 'hit' list window. The results are shown with a number of parameters, and can be displayed in various ways. A potential hit can be linked to the parent window, and coverage can be calculated.

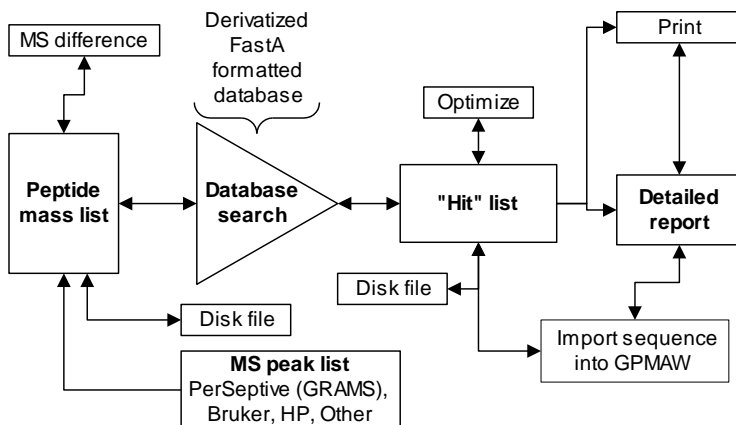
The 'hit' list can be exported to a mass spectrum (GRAMS only) and saved to disk or to clipboard.

### Digest mass search

The digest mass search has received its name because it is based on the masses of the peptides obtained from a proteolytic or chemical digest of a protein. The protein is typically obtained from a SDS gel. Similar to the mass search above, the digest mass search is based on a mass list that can be entered and read in an identical way.

The mass list is then compared to a derivatized database based on a FastA formatted database. These databases are available on the web, and can also be obtained on CD-ROM from NCBI or EMBL. The derivatization of the database is performed in GPMW prior to first use (Ch. 8.2).

## 1 - Introduction



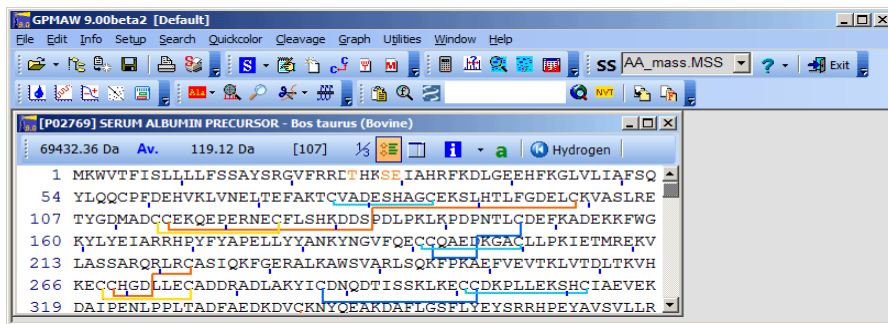
The result of a digest mass search is a list of proteins ranked by high scores (Ch. 5.5). The 'hit' list can be saved to disk for later analysis or comparison with other digest parameters. Individual entries in the list can be viewed as a detailed report (Ch. 8.5) and can also be retrieved into GPMaw for more detailed analysis.

## Starting GPMaw

1.4

### Main GPMaw window

You may start GPMaw either by double clicking on the icon on the desktop or by selecting **Start | Programs | Gpmaw | GPMaw** (Windows 95/98). You are then greeted with an empty window, unless you have enabled '**Autoload last sequence**' (Chapter 5.1) when the most recently loaded sequence will be loaded.



While the program loads a splash screen displaying a lighthouse will be displayed for five seconds. You can remove the picture sooner by clicking inside the picture. The reason for the splash screen is to give you something to look at while auxiliary files are loaded, and also to identify the actual version of GPMaw that is running.



## 1 - Introduction

The title bar of the program shows the name (GPMaw), the version (4.04), the program type (32-bit) and the currently loaded user (Default – see Chapter 5.7).

GPMaw is an MDI application (Multiple Document Interface) in a manner similar to a word processor. This means that you can load any number of sequences (documents) simultaneously, each in its own window and all bounded by the main GPMaw window. From each sequence window you can derive a number of daughter-windows, each with a specific function (e.g. peptide window, mass search results, HPLC graph etc.).

You normally control the program in by a combination of the following:

- Use the mouse or the keyboard to activate the menu options.
- Use the mouse to activate the speed buttons in the toolbar.
- Right-click the mouse in the relevant window (this opens a context-sensitive pop-up menu).
- Use the keyboard shortcuts (see page ii after 'Contents').



**Note:** The pop-up menu that is accessed by right-clicking in the given window is both context sensitive (that is the content depends on the window) and may vary for different parts of a given window. This will usually be the quickest way of accessing commands.

### Functions

Most functions in GPMaw work directly on protein sequences, which means that you have to get a sequence into the program. This is accomplished either by reading a sequence in GPMaw- or text format (Chapter 2.1), reading from a database (Chapter 2.6), pasting from the clipboard (Chapter 2.5), or entering a sequence directly into the sequence editor (Chapter 4.1).

Without a sequence loaded into the program, only a few options are available, like loading a sequence, performing a digest mass search (Chapter 8), a few peptide mass comparison functions (Chapter 6.2) and the utilities described in Chapter 12.

As soon as a sequence is loaded into a sequence window (Chapter 3), a number of menu options are enabled and the menu line is expanded

Whenever you switch between sequence windows and other daughter windows (peptide windows, graphs etc.), the content of the menu will change to reflect the options available.

### The main toolbar

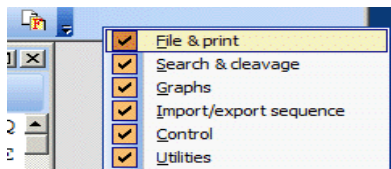
The main window toolbar consists of several smaller toolbars that can be moved around and rearranged. You move a toolbar by grabbing the vertical bars at the left end of the toolbar with the mouse and drag the toolbar to the position wanted. The other toolbars will rearrange to accommodate the new position.



**NB.** The presence and position of each individual toolbar is saved in the 'ini' file when the program closes. This means that they will appear in the same positions when the program re-opens.


## 1 - Introduction

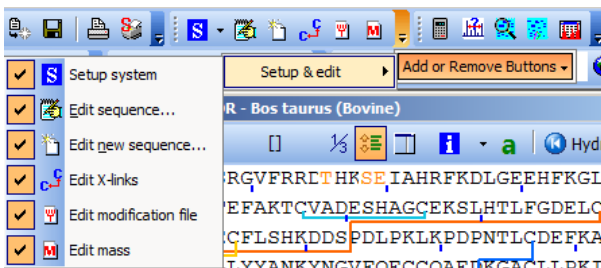
Each small toolbar can be turned on and off, either by clicking on the down-arrow next to the 'Help' question mark or by right-clicking in an unused part of the main toolbar window (thus calling on the pop-up menu). Each visible toolbar is indicated by a check-mark.



All toolbar buttons perform commands that are duplicated in the menu. Furthermore, some of the commands can be accessed through the pop-up menu (right-click the mouse) local to each individual window.

Seven toolbars are available. Some of the commands will only be available when a sequence window has the focus (e.g. 'Save', graph commands etc.). The actual buttons in each bar can be customized by clicking on the down-

arrow to the right of each tool-bar . This will open a menu that enables you turn individual buttons on and off.



The color scheme of the toolbar can be customized by clicking on the down-arrow to the right of the Setup-button. Currently eight different schemes are available.

By grabbing the left edge 'handle' of a toolbar, you can rearrange them freely inside the toolbar area (the area will resize automatically), or you can pull them out and have each bar free-floating on the desktop.

### Individual toolbar commands.

Please check individual chapter for more information.



**File and print**

Open sequence (file, chapter 2.1).

Protein Explorer (chapter 2.9).

Close sequence window.

Save sequence to file (chapter 2.3).

Print.

Printer setup.



**EDIT**

## 1 - Introduction

System setup (chapter 5).

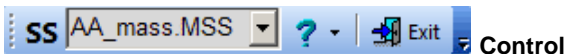
Edit current sequence (chapter 4.1).

Edit new sequence (chapter 4.1).

Edit cross-links of current sequence (chapter 3.5).

Edit modification file (chapter 4.3).

Edit mass files (chapter 4.2).



### Control

SS-button – enable/disable disulfide bridges (oxidized/reduced Cys). See below.

Mass file selection list (drop-down list). The list is gray when the standard mass file (AA\_MASS) is selected, white for all other mass files. See also chapter 4.2.

On-line help. The drop-down arrow presents a menu for menu colors.

Exit – close GPMW.



### Graphs

Hydrophobicity (chapter 11.3).

Secondary structure prediction (GOR - chapter 11.2).

Charge vs. pI (titration - chapter 9.4).

Dot-plot graph (chapter 11.6).

Protein sequence coverage tool (chapter 9.6).



### Search and cleavage

Highlight residues (motifs – chapter 3.3).

Search for mass (chapter 6.1).

MS/MS search (chapter 8.9).

Automatic cleavage (digest - chapter 9.1).

Ms/ms fragmentation (chapter 10.1).



### Import/Export

#### sequence

Import from clipboard (chapter 2.5).

Search FastA formatted database (chapter 2.6).

Search Web Entrez (chapter 2.7).

Internet based sequence retrieval based on accession number (chapter 2.7).

Highlight residues (chapter 3.3).

Copy/export to clipboard.

Export to clipboard in FastA format.



### Utilities

## 1 - Introduction

Composition calculator (chapter 12.2).

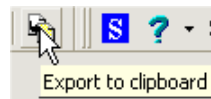
MS peak analysis (chapter 12.1).

Digest (peptide) mass search (PMS – chapter 8).

Simulated 2D gel (chapter 12.5).

Fragment analysis (chapter 12.3).

When using the mouse you can get additional information on a number of controls (particularly the toolbar buttons) by letting the mouse cursor rest on top of the control for a second or two. A **tool tip** will appear explaining the function of the control.



The **'Print'** button is shared for all MDI windows (sequence and daughter windows). Most local pop-up menus (right-click with the mouse) and the 'File' option of the main menu have a copy of the print command. This button cannot be accessed when displaying terminal dialog boxes, but these will have their own print button (if applicable). Most printing from GPMW will be black and white, even if you have a color printer connected to your computer. When you copy a graph to the clipboard (Chapter 11.1) it will be pasted in color enabling you to print graphs in color through other programs.

The **'SS'** button enables and disables cross-links. In the 'SS' state, Cys residues are calculated in the oxidized state (mass 102 Da), while in the 'SH' state Cys residues are calculated in the reduced state (mass 103 Da). If the mass defined in the current mass file (Edit mass file, Chapter 4) does not correspond to 102/103 Da., this button will be disabled (this will typically be the case when cysteine residues are carboxymethylated). The default state of the **'SS'** button is defined in 'Setup system parameters' (Chapter 5.1).

The **'Mass file drop-down list'** enables you to change the masses of any (or all) amino acid residues in all sequences opened on the GPMW desktop by selecting a new mass file. In addition to sequence windows, peptide windows are modified. Other daughter windows may not be changed – you have to check the individual window. In many cases you have to recreate the derived windows (like mass search) in order for the modified masses to be displayed. The most typical use of this function is when you modify an amino acid residue, for example when you carboxymethylate cysteine residues (for more details see Chapter 4.2, Edit mass files).

The background of the drop-down list will be gray when the standard mass file (AA\_MASS) is selected and white when another mass file is selected.

The **'Exit'** button closes all MDI windows and exits the program. In the 'System setup' (Chapter 5.1) you can enable a dialog box that asks you before closing down. If you have made changes to a sequence you will also be asked to save the sequence before closing down.

### Non-sequence dependent functions

A few functions are not dependent on a pre-defined sequence. This comprises **'Database mass search'** (Chapter 8), the **'Mass difference'**

## 1 - Introduction

functions (Chapter 6.2, Mass difference) and the functions of the Utilities menu (Chapter 12).

### Essential tables

1.5

A number of tables in GPMaw are essential for the function of the program. Premier among these tables is the **mass table** that defines the composition, name and abbreviation of each amino acid residue. The mass table works in concert with the atom mass table that defines the mass of each atom used for calculating the mass of compositions.

Modification tables specify the mass of chemical modifications. These tables are also user definable.

In addition a number of tables are hard-coded into the program (they cannot be changed by the user). These include tables for calculating HPLC retention times, pI values, etc.

More information about editing of the various mass and modification tables can be found in details in Chapter 3.

#### Atom mass table.

The atom mass table is edited through the **Edit|Edit mass file** command (Ch. 4.2). The file is global to all mass calculations carried out in the program. For this reason only one copy of the table that is saved in the 'ini' file (initiation file). The atom mass table can at present contain 10 atoms.



**Note:** Any change made to this file will affect all mass calculations carried out by GPMaw!

#### Mass file.

The mass files consist of a table of amino acid residues that are active in the current session of GPMaw. The table consists of 31 entries, each of which contains 1- and 3- letter abbreviation, name and the atomic composition of an amino acid residue.



**Note:** The mass of each residue is not saved as part of the table, but is calculated based on the values in the 'Atom mass table' (i.e. changes to the atom mass table will change the values of the current mass file).

The first entry denotes an unknown residue 'X' that has a mass close to 110 Da (the mass of an average amino acid residue). The following 20 entries contain the 20 'standard' residues, while the last 10 entries can contain any kind of modified residue (see also 'Modification file' below).

The mass file is also global to all mass calculation tables in GPMaw, but can be easily changed through the drop-down box in the main window toolbar.



**Note:** Although the mass file is global, already calculated values may be fixed. However, most windows like the peptide window will change on-the-fly.

Different mass files are usually created in order to accommodate modifications to amino acid residues that are global to a given sequence. A typical example is the carboxylation of cysteine residues. The standard

## 1 - Introduction

installation of GPMAW is supplied with a number of the most common cysteine modifications (pyridyl ethylation, carbamido methylation, cysteic acid etc.). These files can also be downloaded from the web site.



**Note:** If you use some of the 'extra' amino acid residues (after residue 20) you will have to include them in all the modification files you use (e.g. both in aa\_mass.mss and pe\_cys.mss (pyridylethylated cys) if you use both files).

### Modification file

Modification files are tables of amino acid compositions that represent potential modifications to amino acid residues. An example could be methylation of carboxylic acid residues (i.e. a change of +C1H2 valid for Glu, Asp and the C-terminus). As modifications can result in both the addition and removal of atoms, the compositions can be both positive and negative (see Chapter 4.4 for composition formulas).

Only a single modification file can be loaded at any time. The scope of a modification file is global (e.g. a single file covers all sequences loaded at a given time). This means that if you have different sequences loaded, you should construct your modification files to contain common modifications. However, in many cases (e.g. sequence windows) information is extracted from a sequence file and the information stored along with the sequence. This means that the information does not disappear or change when you load a different modification file.

The modification files are used in a number of cases: Modification of residues in sequences (Chapter 3.6), when comparing peptide masses (Chapter 12), and searching for mass matches in a sequence (Chapter 6.1).

Modification files can contain 30 entries each and are saved in the 'System' directory under \gpmaw\.

### N- and C-terminal modification

The N- and C-terminal modifications are similar to the entries in the modification file (above), but are restricted to modifications of the respective terminals. The state of the terminals can be set when editing the sequence (Chapter 4.1), but you can also set the state of the newly generated peptide terminals when you digest a protein.

The composition and modifications of the terminals are edited through the menu **Edit|Edit modifications** (Chapter 4.3) and are saved in a file called 'TERMINALS.TMS'. You are limited to 12 modifications (+ 1 normal state of the terminals, 'H' and 'OH' for the N- and C-terminal respectively).

### Window menu

1.6

The 'Window' entry in the main menu is only concerned with the overall control of MDI windows. The menu contains five commands:

# 1 - Introduction

## Cascade (Ctrl+F5)

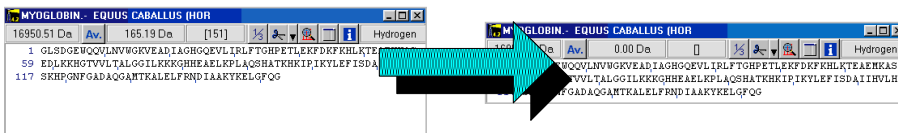
All open windows on the desktop will be resized to the same format, and each window will be positioned slightly below and to the right of the previous one in order to make the title bar of all daughter-windows visible.

## Tile (Shift+F5)

All open windows on the desktop will be tiled, that is resized, and rearranged so they cover the complete desktop

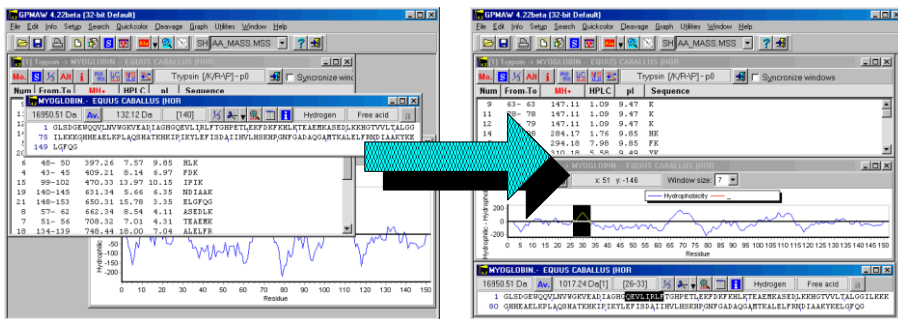
## Fit sequence window

The height of the sequence window will be changed (increased or decreased) to fit the sequence displayed.



## Tile sequence window

When you have multiple daughter windows open (e.g. hydrophobicity, peptide, mass search, secondary prediction etc.) the display can quickly become confusing and difficult to manage. By selecting this option, the currently selected sequence window will move to the bottom of the display (the height will be made to fit) and the daughter windows will be tiled above. Non-related windows will be buried under these windows.



## Arrange icons

All iconized daughter windows will be arranged at the bottom left of the GPMW main window.

## Minimize all

All daughter windows will be iconized and arranged from the bottom left corner of the GPMW main window.

## MDI windows on the desktop

Below the window arranging commands all open MDI (sequence and daughter) windows are listed. An underlined number precedes each window

## 1 - Introduction

(up to 9). This means that you can change the focus to a particular window by pressing <Alt>+W followed by the underlined number.

### Recent program changes

1.7

For a detailed description of the changes made to the program and their chronology, please see the on-line help and/or check online at <http://www.gpmaw.com>.

### Help

1.8

#### Help index

The menu command **Help | Help index** opens the on-line help on the index page. From this page you can either navigate to the help wanted or you can use the search function of the help program.

An alternative way of using the help system is to press the Ctrl + <F1> button to open the help system in a context sensitive way, that is the help will open on the page containing help relevant to the window/dialog that currently has the focus (active window). Most dialog boxes have a help button that likewise opens the context sensitive help.

The on-line help system will always be updated with the most recent features, unlike the manual which is more infrequently updated.



**NOTE:** Additional documentation for GPMAW can be found on the installation CD in the 'Document' directory. Pdf versions of the manual and the 'Dummies Guide' as well as all published newsletters are available. Most of this can also be found at the web site <http://www.gpmaw.com>.

#### About

**Help | About** displays the version and contact information for GPMAW. The top line displays the version number of the current installation. The version number, e.g. 5.11, means main version 5, minor version 1, update version 1. Both main and minor versions will usually have an accompanying manual while an update version only has an updated help file. If you want additional copies of the current manual, please contact Lighthouse data for price and availability (it is also available on your installation CD as a .pdf document).

Below the version name follows the license number and supplier of the program. Please use the license number whenever you contact Lighthouse data concerning GPMAW.

The **'Send mail'** button opens your default mail client with the address of Lighthouse data in the sender section, and GPMAW license information in the body. You can then add your information and send the mail.



## 1 - Introduction

At the bottom of the section, the license date (month and year) and the version number of the originally installed copy of GPMaw is displayed. The license date is useful information when you upgrade, as you can only upgrade for free for a certain time (minimum free upgrade is one year, contact Lighthouse data for the current actual time). When you use the 'Info' button (see below), the program will try to calculate whether your license is eligible for the current upgrade.

Each version of GPMaw (both major and minor versions) has its own associated lighthouse. A picture is shown in the start-up 'splash' screen and a picture is also presented in the left side of the 'About' box'. Below the contact information is a short description of where in the world the lighthouse is located.

Below the description of the lighthouse is the compilation date of the current version of GPMaw along with three buttons:

'System info' contains information on where in your system various GPMaw files are located. This is mainly to make it easier to troubleshoot the system.

'Web home' this will open your default web browser and display the GPMaw web site (<http://www.gpmaw.com>).

'Done' will close the dialog box.

If you have an Internet connection, you can press the 'Info' button at the bottom of the dialog box. This will make the program contact the GPMaw web site and retrieve the latest information about upgrades, bugs and other relevant information. It will also calculate whether your license is valid for the latest upgrade. The actual upgrade you will have to download from the web site using your favorite browser. See also **Auto upgrade notification** in Chapter 1.2.



**Note:** Beta versions of GPMaw are irregularly posted on the web site. As long as you have a license that is valid for upgrades (time limit is generally 18 month from date of purchase), you are free to download and install the beta versions, please see the web site for details on installation. The

## 1 - Introduction

improvements in the beta versions are often only sketchily described, but you are welcome to inquire about specific improvements.

### Lighthouse data

1.9

GPMAW is developed by

**Lighthouse data**

**Engvej 35**

**DK-5230 Odense M**

**Denmark**

**Fax: (+45) 66 19 33 96**

**E-mail: [php@bmb.sdu.dk](mailto:php@bmb.sdu.dk)**

**www: <http://www.gpmaw.com>**

If you cannot access the web site please contact Peter Højrup, Lighthouse data by e-mail at the above address.

Updates to the program will be available for download every 6-9 month (beta versions are posted in-between, please see note above). You are entitled to free upgrades for 1½ year from purchase date (actual update period vary).

In order to continue development of the program, feedback on the current versions as well as input of new ideas are always appreciated.



**Please note:** Unlike many other programs, GPMAW is being developed continuously with relatively frequent updates in the form of beta versions. This means that your reports and suggestions will be taken seriously.

**Technical support:** If you have any problems with the program, please send a fax or e-mail, and we will try to answer in a day or two. Please remember to include license number and version number of your copy of GPMAW when you contact Lighthouse data. Both items can be found in the 'About box' (see 1.8). If you use the '**Send mail**' button in the About box, this information will be included automatically in the mail.

The web site contains additional information pertaining to the program, links to other web resources and downloads of updated versions of GPMAW.

## Reading and saving sequences

How do I get a sequence into GPMW?

How do I save it afterwards?

Handling of protein sequences from/to disk and clipboard.

GPMW normally reads and saves protein sequences in its own format (see Appendix A), but is also able to read a number of other file formats as well as write in FastA format (Export). Furthermore, you can import sequences from almost any source through files and clipboard (Import) as long as the sequence is in standard 1-letter code (see appendix C.3). The sequences can be read from disk, CD-ROM and the Internet.

If you want to enter a sequence manually or edit an already entered sequence, please see Chapter 4.1 for details.

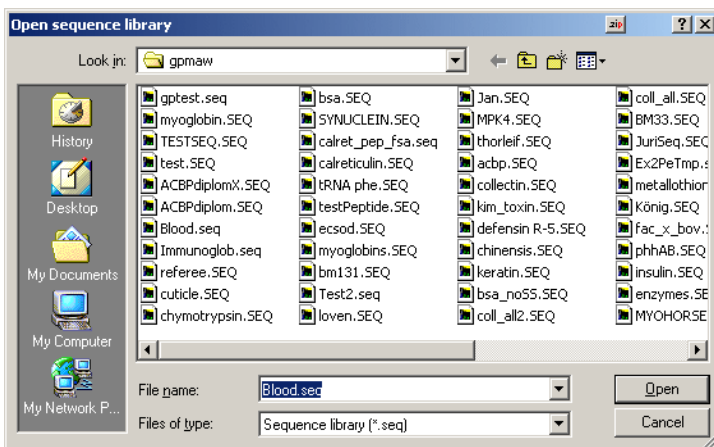
For information on how to work with sequences please read Chapter 3.

### Open sequence

2.1



The **File|Open** command is the standard way of reading protein sequences from disk. GPMW enables you to read sequences in GPMW and FastA format. For an explanation of the GPMW format please see the section below. For information on FastA format please see appendix B.



**Hint:** You should also read section 2.9 'Protein Explorer' for an alternative method of opening sequence files.

## 2 – Reading and saving sequences

In the 'Open sequence library' dialog box, only files with the extension '.SEQ' will initially be shown. Alternatively, you can select 'Old GPMW format' (no extension) or 'All files' in the 'File type' drop-down box. Selection of a file takes place either by double clicking on a library name (opens the file directly), selecting a library name or by entering any file name in the 'File name' field. In the last two cases you have to press <enter> or the 'Open' button. Alternatively, you can change to a different directory or disk drive in the usual File Manager/Explorer style.

The initial directory displayed will be the one entered as 'Default working directory' in the Setup system dialog (see Chapter 5.4). You may also check Appendix D on how to set up GPMW for multiple users.

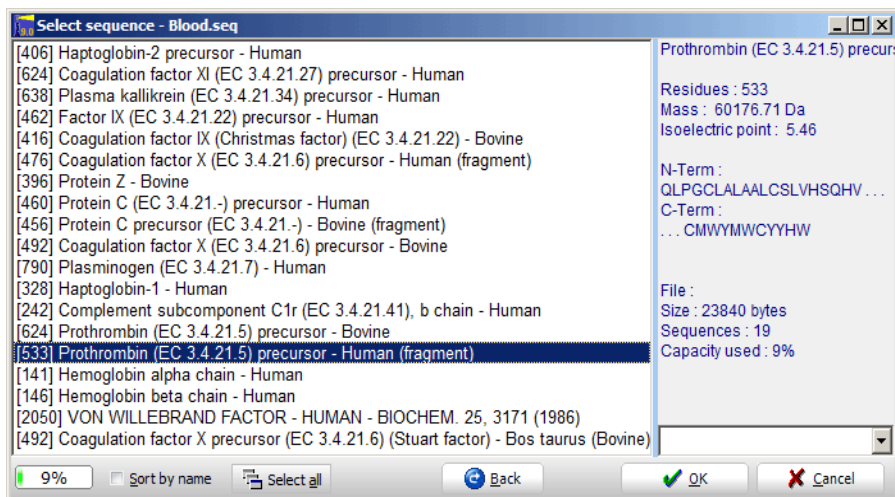
If the selected file only contains a **single sequence**, it will open on the desktop immediately.



**Hint:** The bottom of the file menu shows the nine most recently opened sequence libraries and these can be opened directly. The drop-down arrow next to the 'File open' icon opens a menu with the same five file names.

**Shortcut:** Alt + F followed by a number (1-9) will open the corresponding sequence library.

If the 'Select sequence' dialog box contains multiple sequences, it will list the name and length in residues (in square brackets) of all sequences contained in the file:



Open sequence options:

- You may select all files by pressing the '**Select all**' button.
- Alternatively you can select multiple files by holding down the Shift button when selecting (continuous selection) or holding down the Ctrl button for a discontinuous selection. All selected sequences will be

## 2 – Reading and saving sequences

opened when pressing the **'OK'** button, each sequence in a separate window.

- Pressing the **'Back'** button will close the 'Select sequence' dialog and reopen the 'Open sequence library' enabling you to immediately select a different sequence library.
- Press the **'OK'** button to open the selected file and close the dialog box.
- Press the **'Cancel'** button to close the 'Select sequence' dialog without opening further sequences. Any sequences already opened will stay opened.
- Check the **'Sort by name'** box to sort the contents of the sequence list alphabetically. Un-check the list to display in file order.

The first part of the currently selected sequence can be viewed just above the status line at the bottom of the dialog.

The **status panel** at the right side of the dialog box shows:

- Name of currently select sequence
- The size, mass, and pI of the currently selected protein
- N- and C- terminal residues
- The total file size and number of sequences
- The size of the current file as percentage of maximum file size.
- The bottom left box will be green when the file is less than 80% full, and will then turn red, to warn you that you cannot save many more sequences. The maximum file size is 512 kByte or 250 sequences, whichever limit is reached first.

Dragging the edges of the dialog with the mouse can expand it, if you need to see more sequences and/or more of the sequence names.



**Hint:** You can drag-and-drop sequence files onto the GPMW desktop from the Windows File Explorer.

### Information that are saved with a sequence

The primary information saved contains the name and sequence of the protein (in 1-letter code). In addition, you should also save the accession number of the protein when available (when reading a FastA indexed database the accession number is retrieved automatically).

In addition, you can enter the following information manually (described in detail in Chapter 2 and 3): **Cross-links** (usually between Cys residues, but can be between any residue), **modified residues**, **N- and C-terminal modifications** and **annotation**. The annotation is a free text field where you can enter any information you want saved with the sequence (see chapter 3.9). When reading SWISS-PROT sequences from the EMBL CD-ROM or PIR sequences from the Atlas CD-ROM, the complete database entry is saved in the annotation. Also when you import data (either from file or clipboard) you have the option of placing the entire record on the annotation page (see chapters 2.5-2.7).

## 2 – Reading and saving sequences

### The GPMaw file format

The sequence files of GPMaw can contain up to 300 sequences or 512.000 bytes (characters) whichever is greatest. The files are saved in a proprietary format that in addition to the name and the sequence can contain information on accession number, modified residues, cross-links and annotation.

The sequence files are ASCII (text) files that can be edited in a text editor. However, you should only consider opening the sequence files in a text editor if the file gets corrupted (i.e. unreadable by GPMaw). All input and output is more easily (and safer) handled by GPMaw.

When the file size gets to be more than 90% of capacity you should either delete sequences or create a new file.

For a more detailed description of the file format, please refer to appendix A.

### Drag-and-drop

You may drag files from the Windows Explorer onto the desktop of GPMaw in order to load sequences. The standard rules for opening files apply, i.e. single sequence files open immediately, multiple sequence files open with the 'Select sequence' window. If you drag multiple files simultaneously, the files will open sequentially.

### How to acquire sequences

2.2

Protein sequences can be acquired from a large number of sources; the most important of these will be discussed.



**Note:** GPMaw only accepts sequences in 1-letter code. If your sequence is available only in 3-letter code you have to enter the sequence manually using the **Edit|Edit new sequence** command.

### Journals, own data and other data in print:

You have to enter the sequence manually using the sequence editor as detailed in Chapter 4.1. If you have a scanner with OCR software you can input through your text editor, or a disk file (use **File|Import ASCII**). Beware; as most OCR software have a tendency not to translate 100% correctly you will have to be careful in checking the entered sequence.

If data are in print, they will usually also be available on-line through the Internet. Please see sections below.

### Data in a disk file.

If the file is in FastA format and smaller than 30000 bytes, GPMaw can read it directly using the **File|Open** command. If the file is in an ASCII (text) format you can use the **File|Import ASCII|From file** command (see below). If the file is in a proprietary format (word processor, html etc.) you have to open the sequence in the relevant program and transfer the sequence using 'cut and paste'.

### Transfer of data from a word processor.

If you have your sequence as 1-letter code in a word processor (Word, WordPerfect etc.) the easiest way of getting your sequence into GPMaw is

## 2 – Reading and saving sequences

to copy to the clipboard and select **File|Import ASCII|from clipboard** (see below). Alternatively, you can paste into a new sequence (**Edit|Edit new sequence**). If the sequence is in lower case and/or contain extra formatting characters you remove these using the relevant buttons in the editor (see Chapter 4.1).

### Transfer of data from a web browser (Internet).

Most sequences can be found through the **File|Web Entrez search** (see section 2.7), or, if you have the accession number from the input box, in the main toolbar. However, if you obtain a sequence elsewhere you may proceed as follows:

When you have a sequence loaded into your browser, you can most easily transfer it to GPMaw using 'copy and paste'. The fastest way of transfer is:

1. Highlight the complete record (**Edit|Select all** or <Ctrl-A> if the record takes up the whole page).
2. Change focus to GPMaw and select **File|Import ASCII|from clipboard**
3. Proceed as detailed in the 'Import ASCII' section described below.

Alternatively you can highlight and transfer name, accession number and sequence individually.

You may also download complete databases by FTP. Databases in either FastA or Swiss-Prot format can be indexed it with the 'DBIndex' program (Chapter 12.4). This utility can be downloaded from the GPMaw web site if not present in your installation (see also Appendix B).

### Internet.

You can type the accession number of your protein of interest into the web retrieval field in the main toolbar to retrieve a sequence directly from Expasy (UniPro) or NCBI (Entrez). If you need to retrieve multiple sequences, you can make a list of accession numbers and retrieve all sequences in a single operation (Retrieve Sequence List from Web). You may also query the Entrez database, for details see Chapter 2.7.

### CD-ROM

The GPMaw installation CD from Lighthouse data contains the entire Swiss-Prot database and some other databases (details on the CD). The databases are already indexed but need to be installed on the user's hard drive before access, see chapter 2.6. All other databases (local or downloaded from the internet) can be indexed and accessed from GPMaw (see Appendix B for details).

### Nucleotide sequences.

If you have a nucleotide sequence it can be imported and translated to protein sequence through the **File|Import ASCII|From file** or **File|Import ASCII|From clipboard** command (see below, ch 2.5).

When you have changed the information of a sequence window (e.g. sequence, name or post-translational modification), you can save the information in a GPMW sequence library using either of two commands (below). The information saved contains at least the name and the protein sequence and may additionally contain chemical modifications on terminal or individual residues, cross-linked residues and annotation. Information on multiple peptide chains is saved as part of the protein sequence.

For a complete list on information that can be saved with a sequence please see Appendix A.

#### Save



The save command saves the currently selected sequence and all modifications to the file and position occupied by a previous instance of the sequence. This means that the command only works when the sequence has been read from a GPMW sequence library. The program looks for a sequence with the same name and position in the library. If you have changed the name of the sequence or if you have just entered or imported the sequence file, you will automatically be transferred to the '**Save as**' command (below).

#### Save as

The '**Save as**' command is used when you want to save your sequence to a new file/position or when you have just entered or imported a new sequence and want to save the information in a GPMW sequence library.

- 1) Select the relevant sequence window.
- 2) Select **File | Save as**.
- 3) The '**Save sequence**' dialog box will open in the currently selected user directory. You can change to a different directory. By default only sequence libraries (files with the .seq extension) will be displayed.
- 4) You now either select an existing file or enter a new name. The .SEQ extension is automatically added to the filename.
- 5) If you select an existing file, the sequence in the active window will be appended to the ones already present in the file.



**Note:** If you save a sequence with a name that is already present in the sequence file, the name will get a 'rev1' (or 'rev2' etc.) attached to the end of the name.

#### Save all seq. as

This command will save all sequences opened on the desktop to a single file. This is a convenient shortcut when you have retrieved a number of sequences e.g. from a database search or a retrieval list.



## 2 – Reading and saving sequences

### Save w. highlights

This command is similar to the 'Save' command above, but information on the underlined areas of the sequences (chapter 3.4) is saved along with the other information.

**Note:** This information is not saved dynamically; you have to save it specifically when you make changes. Neither are you informed about changes in underlining that need to be saved.

### Delete sequence

2.4

When you want to remove a sequence from a file you:

- 1) Select **File|Delete** sequence.
- 2) From the '**Select file**' dialog box you select the file containing the sequence to be deleted. By default only sequence libraries (files with the .seq extension) will be displayed.
- 3) In the '**Select sequence**' dialog box you select the sequence to be deleted. If you have multiple occurrences of the same sequence name you should remember that new sequences are always appended to the end of the file.
- 4) From the 'Delete sequence' dialog box you select 'Yes' when you have verified that the selected sequence is correct.

When you remove the last sequence from a file, the file will be removed completely.

When you delete a sequence, the previous sequence file is saved with the same name but with the extension '.BAK'. This ensures that if you delete a sequence by accident, you will be able to retrieve it (until the next save or delete operation on the file).

### Import ASCII (sequence in text format)

2.5

If you are unable to read a sequence file using the normal **File|Open** command, you will be able to import the sequence using the **File|Import ASCII|From file** command. The only limitations are:

1. The sequence has to be in 1-letter amino acid or nucleotide code.
2. The file has to be an ASCII (text) file.
3. The file is smaller than 30000 bytes (characters).

### From file

The '**Open text file**' dialog box is identical to the '**Open sequence file**' for opening standard sequences except that files with the extension '.TXT' are displayed by default. In the drop-down list 'List files of type' you can select 'All files' or you can type any name into the 'File name' box. If the file selected is not an ASCII file, an error message will appear and you will be unable to proceed with the import. If the requested file is a text file, the following dialog box will open showing the file contents:



**Note:** The 'Import ASCII' dialog recognizes the **FastA**, the **Swiss-Prot** and the **GenPept (Entrez)** formats. This means that the name, sequence and

## 2 – Reading and saving sequences

accession numbers are pasted directly into the respective fields of the dialog.

However, remember to include only the complete record and not any extra lines of text or graphic (e.g. when you copy from a web page).

**Import ASCII file**

ID CP2B4\_RABIT Reviewed; 491 AA.  
AC P00178; P00177;  
DT 21-JUL-1986, integrated into UniProtKB/Swiss-Prot.  
DT 21-JUL-1986, sequence version 1.  
DT 13-JUL-2010, entry version 102.  
DE RecName: Full=Cytochrome P450 2B4;  
DE EC=1.14.14.1;  
DE AltName: Full=CYPIIB4;  
DE AltName: Full=Cytochrome P450 isozyme 2;  
DE Short=Cytochrome P450 LM2;  
DE AltName: Full=Cytochrome P450 type B0;  
DE AltName: Full=Cytochrome P450 type B1;  
GN Name=CYP2B4;  
OS Oryctolagus cuniculus (Rabbit).  
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
OC Mammalia; Eutheria; Euarchontoglires; Glires; Lagomorpha; Leporidae;  
OC Oryctolagus.  
OX NCBI\_TaxID=9986;

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

Name: Cytochrome P450 2B4; - Oryctolagus cuniculus (Rabbit)  
Sequence: MEFSLLLLLAFLAGLLLLFRGHGPKAHGRLLPPGPSPLVGLNLLQMDRKGLLRSFLRLREKYGDVFTVYLGSRPVVVLGGTD  
Length: 255 Access. no.: P00178

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

To import a sequence you have to carry out the following steps:

1. Highlight the name and press the **'Name'** button. This will copy the highlighted text into the name field at the bottom of the dialog box (only the text that fits into the dialog box will be displayed).
2. Highlight the accession number and press the **'Access no.'** button.
3. Scroll to the sequence.
4. Highlight the sequence part of the record.
5. If the sequence is written in lower case press the **'Caps'** button to transform lower case letters to upper case. The sequence will stay highlighted.
6. If you are importing a nucleotide sequence press the **'DNA'** button (see import nucleotide sequence below) otherwise press the **'Sequence'** button. Do not worry about numbers, space characters or line breaks - these will not be imported.
7. The first part of the sequence will be displayed in the 'Sequence' line below. The length of the sequence read will be shown in the 'Length' line.

## 2 – Reading and saving sequences

- Press **'OK'** to open a sequence window containing the imported sequence. If necessary, you can edit the sequence later by selecting **Edit|Edit sequence** (Chapter 4.1).
- If you check the **'Save all as annotation'** checkbox, the entire content of the import box will be saved in the annotation page of the sequence. For more information on the annotation see Chapter 3.9, 'Annotation'.



**Hint:** The text field in the top part of the dialog box is an edit control. This means that you can edit the name and sequence before importing into GPMW. You can also use 'cut and paste' from the pop-up menu or you can use the standard keyboard shortcuts (Ctrl-X, Ctrl-C, Ctrl-V).



**Note:** Only amino acid residues that are defined in the currently selected mass file will be imported as part of the sequence. If you need to import unusual 1-letter codes please make certain that you have the appropriate mass table loaded before import.

### Importing a nucleotide sequence

If, in step 6 above, you press the **'DNA'** button you will be presented with the **'Convert DNA sequence'** dialog box:

The six green lines along the top of the dialog box represent the translated nucleotide sequence (three forward reading frames and three backward) with the red dots representing stop codons. If a name has already been selected for the sequence it will be shown above the green lines.

The six buttons below control the display and selection of the reading frame. The protein sequence of the currently selected reading frame will be shown in the large list-box. Stop codons are shown as 'X' and will be imported into GPMW as chain terminators (a maximum of six chains can be imported). To the right of each reading frame button the number of ORFs is shown (open reading frames) along with the size of the largest ORF in the current frame.

## 2 – Reading and saving sequences

If you check the **'Longest ORF only'** checkbox, only the longest ORF will be displayed in the list-box and translated into GPMW.

Usually, the longest open reading frame (in this example frame 3) is the correct one. Checking the **'Longest ORF only'** and pressing **'OK'** will return you to the 'Import ASCII' dialog box above with the translated protein sequence displayed in the sequence line (step 7).

### Import from clipboard

The **'Import from clipboard'** dialog box is identical to the **'Import from file'** except that the content of the clipboard is pasted directly in the text box (top part of the dialog box).

If the sequence on the clipboard is in FastA format, it will be parsed immediately and the name and the sequence will be entered directly into their respective lines (i.e. you only need to press the **'OK'** button).



**Note:** The 'Import ASCII' dialog box is also used as transit station when reading protein sequences from a number of other sources like BLAST search (Chapter 7.2), Internet accession number retrieval (Chapter 2.7) etc..

### Reading sequences from a database

2.6

You can read a sequence from a local database (this section) or directly by accessing certain databases on the Internet (next section, 2.7).

GPMW can access local databases in two formats:

- General FastA format indexed with the database-indexing tool (DBindex – Chapter 12.4) available from Lighthouse data. If the DBindex program was not part of your installation, you can download it free of charge (see Chapter 1.9).
- Swiss-Prot, EMBL, IPI. These databases are annotated and you have to make a FastA version of the databases, which you then search. However, if all the files are contained in the same directory, GPMW will load the full Swiss-Prot style record in stead of the limited FastA record.

See Appendix B for how to acquire the.



**Note:** The Swiss-Prot database is not public domain (freeware). If you are part of a commercial company or institution you need a license agreement. In this case, please contact Swiss-Prot ([www.ebi.ac.uk](http://www.ebi.ac.uk)).

### FastA formatted databases

Protein databases in FastA format are available from a number of sources, particularly on the Internet. Some of the available databases that have been tested with GPMW are:

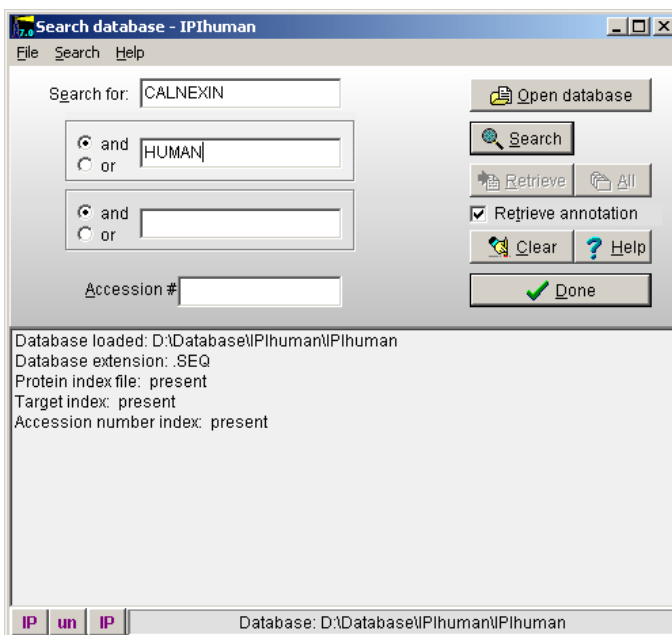
- PIR (Protein Identification Resource),
- Swiss-Prot,
- UniProt,
- the IPI databases

## 2 – Reading and saving sequences

- EMBL non-redundant.
- TREMBL (translated EMBL),
- GenPept (translated GenBank),
- OWL (non-redundant combined database),
- NCBI nr (non-redundant).

Appendix B contains more information on the databases and how to obtain them. Before you can access any of these databases from GPMW you have to create index files. For this purpose the utility 'DBIndex' is included with GPMW. If it is not present in your installation, you can download it from [www.gpmw.com](http://www.gpmw.com). In addition to indexing straight FastA databases, 'DBIndex' can perform the following functions (see Chapter 12.4):

- Convert the Swiss-Prot database to into FastA format before use.
- Rewrite the NCBI nr and EMBL nr databases to a simpler FastA format before indexing.
- Extract sequences with a certain composition to a new database.



Selecting **File|Open FastA database** opens the 'Search FastA database' dialog box.

To search an indexed FastA database:

1. Select **File|Open database** or press the '**Open database**' button. In the 'Open file' dialog box select an index file with the **.trg** extension. The dialog 'remembers' the five most recently opened databases, which are shown at the bottom of the **File** menu. Three buttons at the bottom of the dialog show the initial letter of the three most recently

## 2 – Reading and saving sequences

opened databases. If you let the mouse cursor rest on a button for a few seconds, the fly-by help shows the full name of the database connected to the button. The full path and name of the database opened will be shown in the bottom status line.

2. You can enter up to three search words and combine them with 'and'/'or' as appropriate. Alternatively you can enter an accession number. If you enter an accession number, this field will take precedence over search words.
3. Press the **'Search'** button and any matches to the search criteria will be displayed in the bottom yellow result box. When you enter search words, you should enter the least common word first in order to speed the search and not overload the search.



**Note:** The words 'protein', 'sequence', 'temporary', and 'tentative' have been removed from the search list as being too common and not providing enough information. Furthermore, the following symbols have been removed and replaced by word delimiters: '[",.,:()']

4. Highlight any sequence you want to retrieve from the database and press the **'Retrieve'** button to load the sequence into GPMaw. Alternatively, you can load the sequence by double-clicking on the name.  
As FastA formatted databases contain no annotation, only the name, the accession number, and the sequence will be loaded. You can highlight several sequences while holding down the <ctrl> key. Pressing the **'All'** button will load all highlighted sequences into each own sequence window.



**Note:** If you have several databases installed on your system you do not have to reenter search words when switching between the databases.

The Swiss-Prot and the GenPept databases are published as infrequently updated primary releases and frequently released updated databases (SwissNew and GpNew). When both the main and the updates are installed in the same directory, GPMaw will recognize the updates and the **'Upd.'** button will be enabled when loading the main database. You can now search the main database by pressing the **'Search'** button and the database update by pressing the **'Upd.'** button.

**Swiss-Prot:** If you are searching the FastA version of the Swiss-Prot database ('SPROT.SEQ') and you have all the accessory files in the same directory, you can read the full database record into the annotation page of the sequence window if the **read annotation** box is checked (see Chapter 3.9 for more information on the annotation page).

In order for the annotation retrieval to work you need the following: The full Swiss-Prot database release 37 (the file has to be called 'SPROT.SEQ'). The FastA version of Swiss-Prot ('SWISS.SEQ') generated by the **'DBIndex'** utility (version 1.02 or later) and the corresponding index file 'SWISS.IDX'. The FastA index files 'SWISS.TRG', 'SWISS.NDX', 'SWISS.ACC', and

## 2 – Reading and saving sequences

'SWISS.FAC' also generated with the 'DBIndex' utility (see Chapter 12.4). The '**DBIndex**' program is included in the standard GPMW distribution and can be freely downloaded from web site (Chapter 1.9).

### Retrieve sequences from the Internet

2.7

If you know the **accession number** of your protein of interest you can enter it into the web access field in the mail toolbar:

A screenshot of a software toolbar. It features a text input box containing the sequence 'P01308'. To the right of the input box are three buttons: a 'Web' button with a globe icon, an 'Ala' button with a red square icon, and a 'Mark' button with a black square icon.

When you press the '**Get**' button GPMW will either access the Swiss-Prot database on the Expasy server (if the accession number starts with O, P or Q) or the Entrez server hosted by NCBI. If the first query is unsuccessful, the other server will be queried. The result will open in the 'Import ASCII' dialog box (section 2.5). Both the Swiss-Prot and the Entrez (GenPept) format are recognized by GPMW so you can import the sequence just by pressing the '**OK**' button. The complete database entry will be saved in the 'annotation' page (section 3.9).



**Note:** When searching for NCBI gi numbers you should include the 'gi|' in the accession number; e.g. gi|49522055 not just 49522055



**Note:** The Web search actually works on any unique identifier in the sequence record. However, using anything but the accession number may result in an unpredictable result and should thus be checked carefully. If multiple hits are found, only the first is retrieved.

The results from the search will open in the 'Import ASCII' dialog box with the various sections parsed into their respective fields. In most cases you just have to press the '**OK**' button to import the sequence, but you should, just in case, check that the fields have been interpreted correctly. For more details on the import ASCII dialog, see chapter 2.5.

If GPMW **does not start** searching the Internet, it is most likely caused by a blocked Internet connection, and you have to go through a proxy. Please see section 5.9.

**Note:** The text input box has a second function as you can enter any residue or part of a sequence and press the '**Mark**' button to highlight the residue/sequence in sequence windows. For more information see Chapter 3.2, 'Highlight'.



**Note:** If you are unable to retrieve sequences through the Internet, you may need to set up access through a proxy. Please see section 5.9.

## 2 – Reading and saving sequences

### Retrieve sequence list from web

If you have a list of accession numbers and you would like to retrieve the related sequences into GPMW you select the **File | Retrieve | Retrieve accession number list**. This opens a dialog box where you either enter the accession numbers, load a list from disk ('Load' button) or paste the list from the clipboard – one accession number pr line.

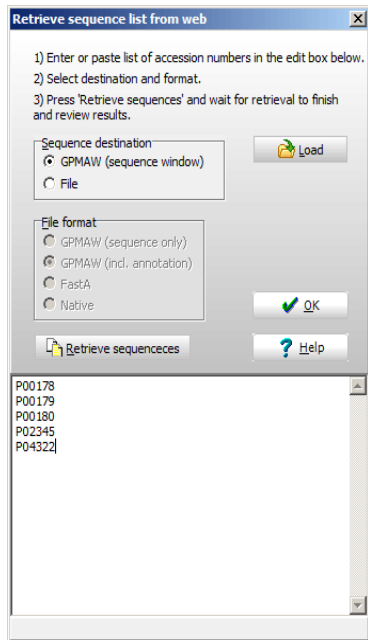
You then select the destination of the retrieved sequences, either as individual sequence windows in GPMW or saved to a file on disk. The 'File format' option will only be enabled when the last option has been selected.

If you retrieve into GPMW you can later save all the sequences into a single file using the **File | Save all seq. as** (see below).

If you select 'File' as destination, you can select file format from the list below: GPMW without annotation, GPMW with annotation, FastA or Native (i.e. as received from the Internet).

When you press the '**Retrieve**' button, GPMW will search the Internet using the SRS server at EBI, UK, if the accession number is recognized as a Swiss-Prot or TrEMBL number, otherwise the search will be done on the Entrez server at NCBI.

The state of the retrieval process is displayed in the bottom status line. As the sequences are retrieved from the web, each accession number is appended by ' – Sequence retrieved'. **Note:** A short interval may pass after the status line reports 'Done' and until all sequences have been retrieved.



**Note:** Gene index numbers (NCBI) have to be entered with a gi| before the number, i.e. gi|18858381.

### Searching Entrez

Through the **File|Web Entrez search** you have the possibility to search for protein sequences on the web using the Entrez search engine. The search by GPMW only implements part of the Entrez search engine and is not meant as a replacement for searching the WWW, but rather as a quick way to retrieve protein sequences into GPMW. This option is particularly useful if you do not have access to protein databases on CD-ROM or do not wish to download a complete database yourself.



## 2 – Reading and saving sequences

**NCBI Entrez web search**

Select database and display format.  
Enter terms and select fields.  
Press 'Search' to search the Entrez database

Database  
☒ Protein  
☐ Nucleotide

Display  
☒ FastA  
☐ GenPept

Terms: Fields:  
 1 CD91 All fields  
 2 All fields  
 3 All fields  
 4 human Organism

Search Cancel Print  
 Max. number of items: 50 No web display  
 Highlight name and press 'Retrieve' to load sequence into GPMW  
 Retrieve Done

prolow-density lipoprotein receptor-related protein 1 precursor  
 LRP1 protein [Homo sapiens].  
 LRP1 protein [Homo sapiens].  
 LRP1 protein [Homo sapiens].  
 LRP1 protein [Homo sapiens].  
 mannose-binding protein C precursor [Homo sapiens].  
 mannan-binding lectin serine protease 1 isoform 3 precursor [Homo sapiens].  
 mannan-binding lectin serine protease 1 isoform 2 precursor [Homo sapiens].  
 mannan-binding lectin serine protease 1 isoform 1 precursor [Homo sapiens].  
 PTB domain-containing engulfment adapter protein 1 [Homo sapiens].  
 RecName: Full=Prolow-density lipoprotein receptor-related protein

LOCUS NP\_002323 4544 aa linear PRI 20-MAR-2011  
 DEFINITION prolow-density lipoprotein receptor-related protein 1 precursor [Homo sapiens].  
 ACCESSION NP\_002323  
 VERSION NP\_002323.2 GI:126012562  
 DBSOURCE REFSEQ: accession NM\_002332.2  
 KEYWORDS -  
 SOURCE Homo sapiens (human)  
 ORGANISM Homo sapiens

Sequence Web  
 11 sequences http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=protein&rettype=gp&retmax=50&id=126012562,47940659,33879163,30

The Internet address <http://www.ncbi.nlm.nih.gov/> contains a full implementation of an Entrez search engine and information on the databases. You can also download a client/server version of the search engine with an enhanced functionality and greatly increased speed. Appendix A and B contain further information on database formats and content.



**Note:** You need an Internet connection in order to use these functions! The Internet connection through GPMW may not work through a firewall. If you have problems, please consult your local network specialists before contacting Lighthouse data.

The Entrez web dialog allows you to enter up to four search terms, search in two different databases and display in two formats.

The results of the search are shown in the bottom part of the dialog box. This is a 'tabbed notebook' where the first page, **'Sequence'**, is made up of two list boxes and the second page, **'Web'**, shows the web page of the Entrez search. Of the list boxes on the 'Sequence' tab, the top one shows the proteins found, while the bottom one shows the database entry of the currently selected protein. Double-click on the entry or press the 'Retrieve' button to retrieve the sequence into a GPMW window. The 'Web' page works like Internet Explorer, except

Enter terms and select fields.  
Press 'Search' to search the Entrez database

Display  
☒ FastA  
☐ GenPept

Terms: Fields:  
 2 All fields  
 3 All fields  
 4 human Organism

NCBI

Protein Nucleotide Protein Genome Structure  
 Search Protein for cd91[ALL] AND human[ORGN]  
 Limits Preview/Index History  
 Display FASTA Show 50 Send to  
 Item 1 - 10 of 10  
☐ 1: NP\_002323. Reports low density lipop... [gi.4738686]  
 >gi.4758686|ref|NP\_002323.1| low density lipoprotein-related r

## 2 – Reading and saving sequences

you do not have an address bar (more information see below).

The '**Max. number of items**' field can be set between 25 and 150. If you retrieve more than 150 sequences, only the first 150 items will be displayed, and you will have to narrow your search by using more strict options.

### Database

**Protein:** This is the NCBI non-redundant protein database containing over 705.000 proteins (July 2001). The database is compiled by combining Swiss-Prot, PIR, GenPept and additional databases. The NCBI non-redundant database is updated almost daily.

**Nucleotide:** The NCBI non-redundant nucleotide database, based on GenBank supplied with other databases. This database is not normally used for retrieval of protein sequences but can be used for cross-references.

### Display

**FastA:** A compact format mostly used for storage of sequences used in homology searches. The first line starts with '>' and contains accession number and name of the sequence. The following lines (usually formatted with 60-character per line) contain the sequence in 1-letter code.

**GenPept:** Considerably more information is available in the GenPept format. In addition to name and species, information on species, literature reference, post-translational modifications etc. are included.

### Terms

Up to four search terms can be entered. The terms are always 'AND'ed, meaning that both term 1 and term 2 have to be present in a database entry in order for Entrez to retrieve it. The '**Fields**' drop-down list boxes enable you to narrow the search to specific fields of the database entries.

**Fields:** The default selection of the 'Field' box is '**All fields**' where the complete database is searched, except for the last term field where '**Organism**' is selected. The '**All fields**' option is usually sufficiently specific in most cases, but as the database entries often contain cross-references to other database entries you will often retrieve a number of homologous proteins.

Other 'Fields' options are **Protein name**, **Keyword**, **Organism**, **Author name** and **Accession #** (number).



**Note:** Whenever you change database the 'Field' boxes are reset to '**All fields**' / '**Organism**'.

### Retrieving a sequence into GPMW

Selecting a sequence name in the top list box and pressing the '**Retrieve**' button retrieves a sequence. Alternatively, you can double-click on a sequence name.

If the display format is '**FastA**', the sequence is directly copied into GPMW as a sequence window. If another format is chosen (e.g. '**GenPept**'), the complete sequence record is copied into the '**Import ASCII**' dialog box (see

## 2 – Reading and saving sequences

2.5) from where you can select whether you want just the sequence read into GPMaw or you want to save the annotation as well.

### Proxy

If you need to go through a proxy to get access to the Internet, you have to press the **'Proxy'** button first and fill out the form with **Server name**, **Port number**, **User name** and **Password**. Check the 'Use proxy settings' to tell the internet connection to go through the proxy. Once entered, the proxy settings are used throughout the session, but you have to enter the password when you start a GPMaw session next time.

### Print

Pressing the **'Print'** button will print the complete search results unless part of the results is highlighted. In this case, the program will ask whether to print only the selected part. If you answer **'No'**, the complete result list will be printed.

The printout will also list the search terms, fields and database.

### Local menu

The dialog supports two local menus:

Right clicking in the top part of the dialog displays a dialog with the options: **Clear Terms**, **Retrieve sequence**, **Copy to clipboard** and **Print**. 'Clear terms' clears the two term input boxes, 'Retrieve sequence' and 'Print' duplicates the corresponding buttons and 'Copy to clipboard' copies the selected protein to the clipboard.

The bottom edit box supports a local pop-up menu that enables you to copy and paste part or all of the contents.

### Session example

Search for 'Surfactant protein A':

Select Database: Protein; Display: FastA; Term 1: Surfactant protein A; Field 1: All fields (make certain to get all hits);

The result is 67 hits, several of which do not have relation to SP-A.

To limit the search you enter 'human' in Terms field number 4 and press 'Search' again.

The result is now 15 proteins

In order to get the annotation you switch to 'GenPept' format and press 'Search'.

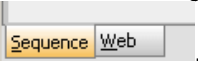
Highlight the sequence 'PSPA\_HUMAN' which originates from the Swiss-Prot database and has the best annotation. Press 'Retrieve' to open a new sequence window in GPMaw with human SP-A.

## 2 – Reading and saving sequences


Although the FastA and the GenPept search show the same results, the search in FastA format is usually much faster (semi-automatic).

### Web view

The bottom of the dialog box shows that you have two pages to the bottom

part:  .

Clicking on the ‘Web’ page shows the web entry from which the ‘Sequence’ page has been extracted.



NCBI Entrez Protein

Search: Protein for [Go] [Clear]

Display: FASTA Show: 100 Send to: File

Items 1-67 of 67 One page

☐ 1: Q9NPY3 Complement compon...[gi.21759074]

>gi.21759074|sp|Q9NPY3|CD93\_HUMAN Complement component Clq receptor precursor (Complement component 1, q subunit)  
MATSMGLLLLLLLLTQPGAGTGADTEAVVCYTACTYTAHSGKLSAAEAQNHCHNQGONLATVKSKEEAQ  
HVQRVLAQLLRREAALTARMSKFUIGLQREKKGKCLDPSLPLKGFQSVYGGEDTPYSNWHKELRNSCISK  
CVSLILDLSOPLLPRLPKWSEGFQCGSPGSPGSNIEGFVCKFSFKGMCRPLALGGPGGVITYTTPFOTTSS

This is a standard web interface, and you are able to follow links and enter information into input boxes etc. However, you only have the commands available in the pop-up menu (right-click in the window).

You may highlight copy and paste from this window if the ‘Sequence’ view is not sufficient. However, it is usually faster and easier to navigate through the ‘Sequence’ tab window.

### Export sequence

2.8

The Export sequence command yields five options. The first two options save the sequence to a file on disk and the last three copies the sequence(s) to the clipboard:

### As basic GPMW file

You can save your sequence to a disk file in the basic GPMW format, that is name and sequence only, without information about cross-links, modified residues, annotation etc. See Appendix A for details.

### as FastA file

The sequence is saved to disk in FastA format containing name and sequence only (see Appendix A and B for details). This format is very useful for interchange with other programs, transfer to the Internet, etc. For transfer to input boxes in other programs (e.g. the Internet) the ‘to clipboard’ function is usually more convenient, see below.

## 2 – Reading and saving sequences

### to Clipboard

It is often required that a sequence is formatted in a special way in a report, and for this purpose you can choose the **File|Export sequence|to clipboard**.

The **'Export sequence to clipboard'** dialog box displays the sequence name in the top part (for verification) and presents a number of options below:

**Residues per line:** In 1-letter code 60 is the default; in 3-letter code 20 is the default. Range is 10 -100.

**Residue type:** 1- or 3-letter code. Default is like the current sequence window.

#### Numbering:

**On** - each line ends with the number of the last residue.

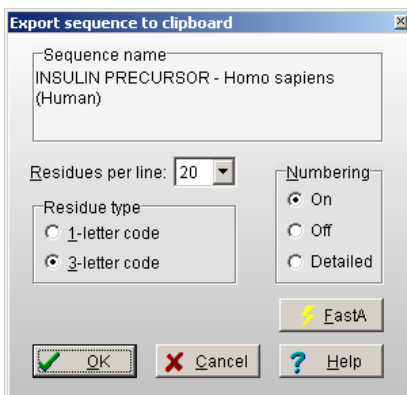
```
Haptoglobin-2 precursor - Human (406 res.)
MSALGAVIALLLWGQLFAVDSGNDVTDIADDGCPKPPEIA 40
HGYVEHSVRYQCKNYYKLRTEGDGVYTLNDKKQWINKAVG 80
```

**Off** - no numbering.

```
Haptoglobin-2 precursor - Human (406 res.)
MSALGAVIALLLWGQLFAVDSGNDVTDIADDGCPKPPEIA
HGYVEHSVRYQCKNYYKLRTEGDGVYTLNDKKQWINKAVG
```

**Detailed** - sequence residue numbers appear beneath every 10th residue of the sequence (remember to display the sequence in a monospaced font, like Courier, for this function to be effective).

```
Haptoglobin-2 precursor - Human (406 res.)
      10      20      30      40
MSALGAVIALLLWGQLFAVDSGNDVTDIADDGCPKPPEIA
      50      60      70      80
HGYVEHSVRYQCKNYYKLRTEGDGVYTLNDKKQWINKAVG
```



The **'FastA'** button will put a '>' in front of the name, switch residues per line to 60, and put numbering 'off' in order to present a FastA formatted sequence

## 2 – Reading and saving sequences

to the clipboard. This is a common way of presenting data in input-boxes on the Internet, and most other sequence analysis programs will recognize this format.



**Note:** When you transfer a sequence to a report, remember to print it in a monospaced font (e.g. *Courier New*) in order for the numbering and amino acid residues to line up correctly. You can also copy your sequence to the clipboard (copy and paste) for transfer to other programs by selecting **Edit | Copy** (or press Ctrl + C), which places a copy of the sequence in the currently selected format (1- or 3-letter code) on the clipboard.



**Note:** When copying this way only the sequence, not the name is copied to the clipboard. If you need both name and sequence use the Export option.

### to Clipboard as FastA (<Ctrl-F>)

The currently selected protein will be copied to clipboard in standard FastA format (60 residues/line). By using the shortcut <Ctrl-F> you can quickly export sequences when you need it for transfer to other programs or to the web.

### all sequences as FastA

All protein sequences open on the desktop (i.e. content of all currently opened sequence windows) will be copied to the clipboard. The order of the sequences will be in the Z-order of the respective sequence windows (i.e. the topmost sequence window will be first).

This option can for example be very handy to copy all sequences to a multiple alignment input box on the Internet.


## Protein Explorer

2.9

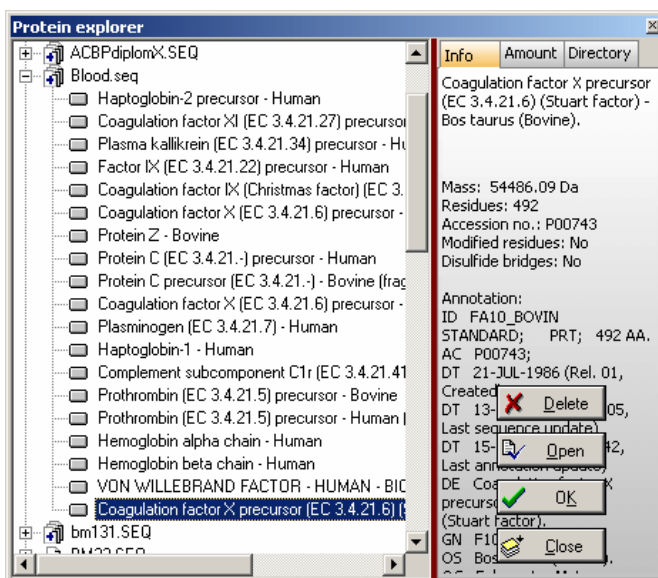
The Protein Explorer is an alternative way of managing your protein sequences. The sequences are stored in the same files on disk as detailed in Chapter 2.1 and 2.3, but they are managed differently.

The main advantage of the Protein Explorer over the standard 'File open' command (Ch. 2.1) is that you can check and open sequences from multiple files without having the dialog close on you. Furthermore, you can delete sequences, check the content, and calculate amounts, all from the same dialog box.

The Protein Explorer is opened from the main toolbar by pressing the second

button in the file section of the toolbar . The initial display will always be the sequence files in the default working directory (see Ch. 5.4).

## 2 – Reading and saving sequences



The Protein Explorer displays the content of a single directory at a time. The directory name and path are shown at the top, and the files are shown as the first level in a tree structure.

If you click on the '+' sign ( ) next to a sequence file name, the tree will expand to show the name of the sequence contained in the file. If the file icon shows multiple pages ( ), the file contains multiple sequences. Once a file name is expanded, you may close the view by pressing the '-' sign ( ). Expanding and closing a file can also be performed by double-clicking on the file name.

When a sequence name is selected (highlighted), the name, mass, size, accession number, modifications, and annotation will be shown in the right-hand panel.

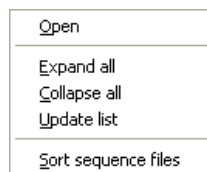
You may open the sequence either by double clicking on the sequence name, press the 'Open' button , or press the 'OK' button .

Pressing the 'Open' button will leave the dialog open; ready for additional manipulation, while pressing the 'OK' button will close the dialog. Alternatively, you can right-click on the name and select **Open** from the pop-up menu.

A sequence may be deleted by selecting it in the tree view (left side panel) and press the 'Delete' button .

### Local menu

A few additional commands are available in the local pop-up menu:



## 2 – Reading and saving sequences

**Expand all:** All sequence files are expanded to show the contained sequence names.

**Collapse all:** All sequence files are collapsed and will only show the file name.

**Update list:** The file list is updated with new and removed files and sequence names.

**Sort sequence files:** The file list is sorted according to name. **Note:** The sequence names in each list is not sorted, they are listed in the order they occur in the sequence file.

### Picomole calculator

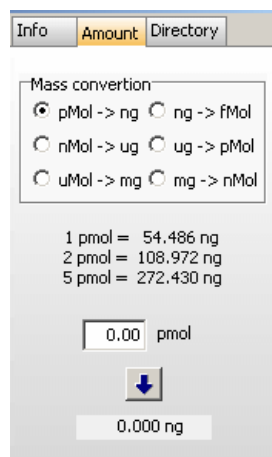
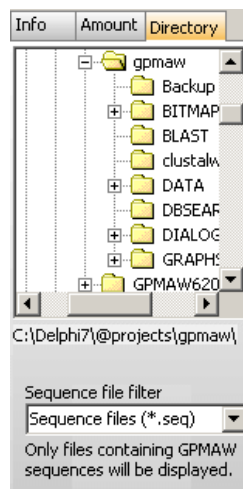
If you select the '**Amount**' tab in the right-hand panel, you switch the panel to the **Picomole calculator**. Based on the protein currently selected in the Protein Explorer, you can on this page calculate the weight of a given amount (pMol/nMol/ $\mu$ Mol) or vice versa. Start by selecting the conversion type among the **Mass conversion** radio buttons, then type in the amount to be converted in the top edit box, and finally read off the converted amount in the bottom edit box. The conversion is calculated for every character entered.

As the value in the top edit box is conserved, you may change between different **Mass conversion** types without re-entering your values.

### Directory

The Protein Explorer always starts up in the default directory (see Ch. 5.4). On the last tab in the right-hand panel you can change to a different directory.

You can also change the file display filter. Options are \*.seq, \*.dat and \*.\* (all files).



**Note:** The Protein Explorer is permanent in the way that you do not close it just hide it. This means that the next time you call the Protein Explorer, it will open with a view identical to when you closed it.



## The sequence window

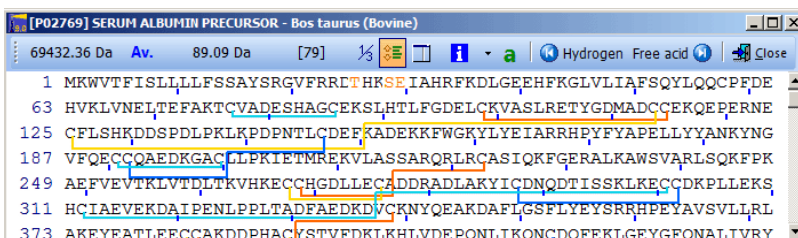
Displaying protein sequences in GPMW. The sequence window forms the basis for most other windows and operations.

The protein sequence window is the parent window for most other windows (peptide windows, graphs, results of searches, etc.). These windows will be daughter windows derived from the sequence window, and when the sequence window (the parent) is closed, all derived daughter windows will also close. Exceptions from this rule are the database mass search window (Chapter 8), fragment windows (which are parent sequence windows themselves, see below), and all dialog boxes, most of which are terminal windows, which means that they have to be closed in order for the workflow to continue.

### Sequence window display

3.1

When a sequence is read from disk or clipboard (Chapter 2) or has been entered as a new sequence in the sequence editor (Chapter 4.1) it will be displayed in a sequence window:



Depending on the default values in Setup (Chapter 5.1), the sequence will be displayed in either 1- or 3-letter code. The sequence wraps around the right edge of the window, and each line starts with the number of the first residue of that line. If the sequence is too long to be displayed completely, a scrollbar will appear along the right edge, allowing the remainder of the sequence to be brought into view.

The number of residues on each line depends on the width of the sequence and whether the 'Display modula 5' option has been set in the setup (Chapter 5.6). With the option set, the number of residues on each line is limited to numbers that can be divided by 5 (e.g. 20, 25, 30 etc.), the rest of the right-hand side of the display will be empty. If the 'Display modula 5' option has not been set, the window will be filled to the maximum number of residues that can be fully displayed.

### 3 – Sequence window

#### Navigating the sequence

The first level of navigation is the numbering of the first residue on each line. Then it is possible to label every 10<sup>th</sup> residue by setting the 'Number 10<sup>th</sup> residue' option in the setup (Chapter 5.6). If you are viewing the sequence in 3-letter code, the numbering will show as subscript with the number divided by 10 (e.g. residue 120 will have the subscript '12').

```
1 Ala-Gly-Ser-Tyr-Leu-Leu-Glu-
19 Trp-Glu2-Glu-Ile-Cys-Val-Tyr-
37 Thr-Thr-Asp-Glu4-Phe-Trp-Arg-
55 Pro-Cys-Leu-Asn-Asn-Gly6-Ser-
```

If you are viewing the sequence in 1-letter code, the subscript will only be a small vertical line due to limited space.

```
1 AGSYLLEELFEGHL
75 PGYEGPNCAFAESE
149 QNLLPFPWQVKLTN
```

In addition to direct numbering of residues, GPMW has a number of tools to help you analyze your sequence and you should read each of the following paragraphs to make certain you understand the differences.

The primary tool is the mouse. When you move the mouse pointer across the sequence, the mass and position of the residue pointed at will be shown in the toolbar.

Often you are interested in specific residues, residue types, or short sequences (motifs). GPMW enables you to **color** these in three different background colors.

- Individual residues can be colored by double clicking on the relevant residue followed by selecting a color at the bottom of the dialog box (see **Highlight residues** below).
- Residue types and short sequences are colored through the **Search|Highlight residues (motifs)** command (see below)

This **background coloring** of residues is persistent and is carried along to peptide windows as well. Please read the detailed description on the limitations below.

**Highlighting** (inverting) sequences. This is a quick way of drawing attention to and getting the mass of a partial sequence. Highlighting is also the fastest shortcut when making Fragment windows (see below). Highlights made in this way are not persistent, and will disappear when you next click the mouse inside the sequence window. See also the discussion in the 'Highlight sequence' section 3.2 below.

**Underlining** is the third way of drawing attention to a peptide. The normal use of this function is to underline peptide sequences found when making a mass search (Chapter 6.1) or digest mass search (Chapter 8), but highlights can also be converted to underlines. This function is particularly useful when you want to calculate the coverage of a given number of peptides. Underlining is persistent (has to be deleted explicitly) but is not carried on to daughter windows. You can activate underlining either from the mass search results window or directly in the sequence from the menu (**Edit|Underline**) or pop-up mouse menu (**Underline**).

**Marked residues** are individually identified residues. They are displayed on screen with a colored line around each marked residues. You select marked

### 3 – Sequence window

residues in the **QuickColor|Marked residues** command of the main menu. The ‘marked residues’ are persistent if the sequence is saved after definition of the residues.

	Persistent	Sub-windows	Main function
<i>Color</i>	Yes	Yes	Navigating sequence
<i>Highlight</i>	No	No	Peptide mass / fragment
<i>Underline</i>	Yes	No	Calculating coverage
<i>Marked</i>	Yes	No	Attention to spec. res.

#### Main Toolbar

Each sequence window has its own status bar above the sequence containing five panels and five buttons.

46463.39 Da **Av.**

The first panel shows the total molecular mass of the protein. The button to the right of this panel toggles between average and monoisotopic mass display (see Appendix C). When showing average masses the button displays a blue ‘**Av.**’ and a red ‘**Mo.**’ when the mass is monoisotopic (mass types see Appendix C). If you right-click on the panel, a pop-up window will open with a list of the mass of cluster ions (2MH<sup>+</sup>, 3MH<sup>+</sup>...6MH<sup>+</sup>) and multiply charged ions (MH<sup>2+</sup>, MH<sup>3+</sup>.... MH<sup>6+</sup>) (see right). These values are also displayed in the **Info|Sequence info** dialog box (Ch. 3.8)

1M+ 46464.39 / M1+ 46464.39  
2M+ 92927.77 / M2+ 23232.69  
3M+ 139391.16 / M3+ 15488.80  
4M+ 185854.55 / M4+ 11616.85  
5M+ 232317.93 / M5+ 9293.68  
6M+ 278781.32 / M6+ 7744.90

from where they can be printed or copied to the clipboard.

147.13 Da [204]

The second panel in the sequence toolbar shows the mass of the residue pointed at by the mouse cursor, while the third panel shows the number of the residue. When highlighting part of the sequence (see below), the content of the second and third panel changes to reflect the selected peptide (i.e. mass and range of selection respectively) see below and right.

1319.40 Da[1] [81-92]



**Note:** If the current sequence has an offset number (e.g. the first residue is not counted as residue 1 – see Chapter 4.1) the numbering in the third panel will be shown in **red**.

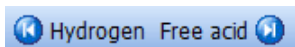


The next five buttons have the following functions:


- Toggle between 1- and 3-letter code.

### 3 – Sequence window

- Increase line spacing, in particular to give more space to drawing the Cys cross-links.
- Open/close the information frame (see below).
- Extended protein sequence information (see section later). The drop-down arrow opens a menu with direct access to the various pages in the information dialog box (chapter 3.8).
- The last button opens the annotation window. The button will be colored according to the status of the annotation:  
**Gray:** No information (i.e. annotation is blank).  
**Red:** There is information in the annotation.  
**Green:** Information in Swiss-Prot format (i.e. the feature table can be imported into the sequence).  
**Blue:** Information if GenPept (Entrez) format.



The next two panels show the status of the N- and C-terminals, respectively. The defaults are '*Hydrogen*' and '*Free acid*' for the unmodified polypeptide. The terminals can be edited in the '**Edit sequence**' dialog box. Resting the mouse cursor on either field for a few seconds will allow the fly-by hint to show the composition of each terminus. When the mouse cursor passes over a **modified residue** in the protein sequence, the two panels will turn yellow and show the name and composition of the modification respectively.

The last button  **Close** closes the sequence window and all derived daughter windows like peptide window or graphs.

Two toolbars in the main toolbar (chapter 1.4) are linked to the sequence window (i.e. the buttons will be grayed when the focus is on a non-sequence window):



Search and cleavage

- Highlight residues (motifs – chapter 3.3).
- Search for mass (chapter 6.1).
- MS/MS search (chapter 8.9).
- Cleave protein function (automatic digest, see Chapter 10).
- Open a drop-down menu enabling the fast selection of a cleavage (Quick cleave, see Chapter 10).
- Ms/ms fragmentation (chapter 10.1).

The arrow next to the '**Highlight**' and '**Cleavage**' buttons indicates that both buttons have drop-down menus.

The '**Highlight**' drop-down menu contains the '**Quickcolor**' menu that enables you to color a number of specific residues (e.g. basic, acidic, cysteine etc.) by a single mouse click (chapter 3.3).

The '**Automatic cleavage**' drop-down menu contains the top-most 10 enzymes defined in the enzyme list (chapter 9.1). Note that when you use the drop-down enzyme list, you simulate a 'straight' cleavage without overlaps (missed cleavages) and other options you can specify in the cleavage dialog.

### 3 – Sequence window



Graphs

- Hydrophobicity (chapter 11.3).
- Secondary structure prediction (GOR - chapter 11.2).
- Charge vs. pI (titration - chapter 9.4).
- Dot-plot graph (chapter 11.6).
- Sequence coverage map (chapter 9.6).

The commands in the 'Sequence' toolbar are available in the 'Search' menu, while the 'Graph' toolbar commands are in the 'Graph' menu.

#### Information frame

The information frame appears when you click on



the frame button in the sequence window toolbar. Alternatively, you can specify in the setup (Chapter 5.1) that the frame shall open along with the sequence window. The frame is a tree view control that initially opens with all secondary levels closed, only the headers are visible. By clicking on the '+' signs, the corresponding sub-level will open showing the relevant information. Clicking again on the '-' sign will close the sub-level.

The frame contains information on:

**Termini** – the status of the protein terminals.

#### Modified residues

- + Modified residues
- Cross-linked residues
  - Cys120-Cys146
- Net charge
  - 57.8 at pH 2.0
  - 58.7 at pH 7.0
  - 44.7 at pH 5.0

+ Molar Ext./Abs. @280 **Cross-linked residues** – typically disulfide bridges.

**Net charge** – theoretical charge of the protein at pH 2.0 and pH 7.0 and user-selected.

**Molar Ext./Abs. @280** – theoretical extinction coefficient and absorption calculated at 280 nm.

- Highlights
  - Inverted: 6 [1.5%]
  - Underlined: 0 [0.0%]
- Sel. mass
  - MH1+ : 675.759
  - MH2+ : 338.383
  - MH3+ : 225.925

Protein

- + Termini:
- + Modified residues
- + Cross-linked residues
- + Net charge
- + Molar Ext./Abs. @280
- + Highlights
- + Sel. mass

### 3 – Sequence window

**Highlights** – fraction of the sequence which is underlined or highlighted. This value is dynamically updated.

**Sel. Mass** – the mass of the currently selected region of the sequence. Both the singly, doubly and triply charged species are shown. These values are updated dynamically (e.g. changes as the sequence selection changes).

The information frame can be resized by grabbing the right edge of the frame with the mouse cursor and move it left / right.

If you have made a coverage map (Chapter 9.6) an extra field is added to the end labeled **Peptide hit**. Selecting this will display **First-last**, **Record #**, **Exp. mass**, **Accuracy**, and **e-value**, as you move your cursor across the highlighted parts of the sequence.

Edit	▶
Modify -Cys241-	▶
Search for mass	
Highlight motif	
Search	▶
Underline	▶
Automatic Digest	
MS/MS fragmentation	
Create Fragment Window	
Peptide info	
Fit window	
Print	
Help	

#### Pop-up menu

The pop-up menu, which can be accessed by right clicking in the sequence window, contains the following commands, most of which are copied from the **Edit**, **Search** and **Cleave** main menu options and are explained in the following sections.

**Edit**

- Edit sequence
- Edit cross-links

**Modify** [followed by the residue pointed at by the mouse cursor]

- Simple modification choice submenu, see chapter 3.6,

**Search for mass**,

**Highlight motif**

- Digest mass search,
- Mass difference,
- Mass X-links,
- Local BLAST),

**Underline**

- Underline range,
- Underline highlight,
- Clear underline,
- Clear highlighted underline) (Chapter 3.4),

**Automatic digest**,

**Ms/ms fragmentation**,

**Create fragment window**,

**Peptide info** (only enabled if part of the sequence is highlighted),

**Print**

**Help**.

The peptide info works only if you have part of the sequence highlighted. GPMW then extracts the highlighted part and treats it as a separate

### 3 – Sequence window

peptide. All the parameters that are calculated for peptides (mass, pI, HPLC index etc.) are then displayed in a peptide info window. For details see the peptide info window in the section on Automatic digest (Chapter 9.1). If multiple sequences are highlighted, only peptide information for the last highlight is displayed.

#### Copy

You can copy the sequence to the clipboard (ready for pasting into another application) by selecting **Edit | Copy** (<Ctrl+C>). This places a copy of the sequence in the displayed format (1- or 3-letter code) on the clipboard. If part of a sequence is highlighted, only this region will be copied to the clipboard.

If you press <Ctrl+F> the complete sequence will be copied in FastA format (appendix B.3). This command corresponds to **File | Export sequence | to clipboard as FastA**.



**Note:** The name of the sequence is not copied. You can copy the sequence name as well and get much finer control of the copying process by selecting **File | Export sequence** (see Chapter 2.8).

#### Display font

You can change the display font and size by selecting the **Info | Change sequence font** menu option. You are only able to select monospaced fonts (the default is Courier New), but you can select any font size and display type like bold, italic etc.

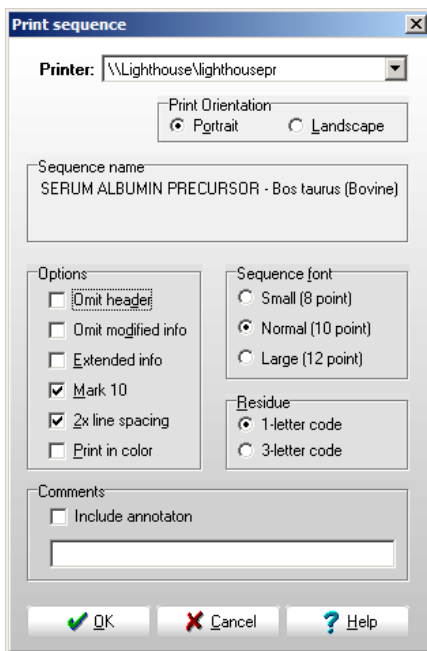
The selection of sequence display font is transient as the font name and size is not saved. Furthermore, it is limited to each individual sequence window.

#### Print

Printing can be selected either from the toolbar, the main menu (**File | Print**), the pop-up menu, or by pressing <F5>.

In the header the standard printout includes: The name of the protein, file origin and position in the file, sequence range, mass file, total mass, state of N- and C-terminus and cysteine. Then follows the sequence in 1- or 3-letter code.

### 3 – Sequence window



When you print the contents of a sequence window you have the option of setting a large number of parameters for the hardcopy, both for the header, the sequence and for additional information. Fortunately, most of the options can be preset in Setup, see Chapter 5.1/5.2.

#### Print options:

**Printer:** In the drop-down box you can select any printer installed on your system. The default printer will always be selected when the dialog box opens.

**Print orientation:** Enables you to switch between portrait (vertical) and landscape (horizontal) print mode.

**Omit header:** Does not include the standard header (file origin, size and mass of the sequence, mass file, state of terminals and Cys).

**Omit modified info:** Does not include information about modified amino acid residues. If this box is not checked the information will only be printed if at least one modification is present in the protein.

**Extended info:** Includes amino acid composition, cross-linked residues, elemental composition, and mass of residues.

**Mark 10:** Every tenth residue will be marked with '=' instead of '-'. This option is only effective when printing in 3-letter code.

**2x-line spacing:** Print the sequence with double line spacing.

**Print in color:** Includes color information (background color, modified residues, cross-links). When printing on a monochrome printer (e.g. laser



### 3 – Sequence window

printer) most of the colors will print as shades of gray. You will have to experiment as to how the colors translate.

**Sequence font:** Select between small (8 point), normal (10 point - standard) and large (12 point) characters. The sequence will be printed in the same font as the sequence display (default is Courier New). See '**Display font**' above for information on how to change the display font.

**Residue:** Select 1- or 3- letter code. The selection will always be default to the sequence window display selection.

**Include annotation:** If checked, the annotation will be printed at the end of the sequence report. If there are less than 10 lines on the annotation page, the annotation will be printed in the same font as the sequence. If there are more than 10 lines, it will be printed in 8 point in order to conserve space on the printed page (in this case it is likely to be a database annotation). See Chapter 3.9 for more information on the annotation.

**Comments:** Here you can write any comment that you want to include on the printout. The comment will be printed after the header, just before the sequence.

#### Highlight sequence

3.2

You can highlight part of a sequence by moving the mouse cursor to the first (or last) residue of the fragment you want to analyze, press and hold the left mouse button while moving the mouse cursor to the last (or first) residue you want highlighted. The status panel of the current sequence window will change to reflect the highlight status: the second panel will show the mass of the highlighted peptide (residue mass + water) and the highlight number in square brackets, the third panel will show the number of the first and last residue highlighted.



**Note:** Be careful not to confuse the '**Highlight sequence**' (a transient operation, this section) with the '**Highlight residues (motifs)**' command (persistent operation, see the section 3.3 below).

#### Persistent and Multiple highlights

The highlighted area will remain displayed until the mouse button is clicked again inside the sequence window. Up to three regions can be highlighted simultaneously by pressing the shift button when highlighting region two and three with the mouse. The status panels will show the **combined** mass of the **individual peptides** (not their residue mass) and the residue numbers reported will be that of the last highlighted region (number of highlights will be shown in square brackets).

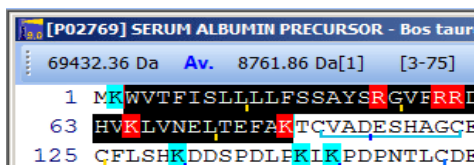
When you have a highlighted sequence you can extract, contract and move the highlighted region by using the keyboard arrow keys:

**Left/right key:** Extends and contracts the C-terminal position.

**Ctrl + left/right key:** Extend and contracts the N-terminal position.

**Shift + left/right key:** Moves the selected region towards the N-/C-terminus of the protein.

### 3 – Sequence window



**Hint:** If you highlight relevant residues (i.e. Lys and Arg when working with trypsin) it is much faster to locate relevant peptides.

#### Using highlights as input for other functions

When a region has been highlighted, you can use this part of the protein as the default input for a number of other functions. These are most easily accessed through the pop-up menu (right-click the mouse in the sequence window), but several of them can also be accessed through the menu.

If more than one region is highlighted, the input will be the most recently defined highlight region.

**Peptide info:** This is similar to the peptide info for peptides in the '**Automatic digest**'. This window contains physical/chemical information on the highlighted part of the protein, as if it existed as a peptide. For more information, please see 'Peptide window', chapter 9.3.

**Ms/ms fragmentation:** The input for the ms/ms fragmentation will be the highlighted region, if any. If the total sequence length of the protein is less than 50 residues and no region is highlighted, you will get ms/ms of the whole sequence (see chapter 10.1 for more information on ms/ms analysis).

**Edit | Copy to clipboard:** If part of the protein is highlighted, this region will be copied, if no part is highlighted the complete sequence will be copied.



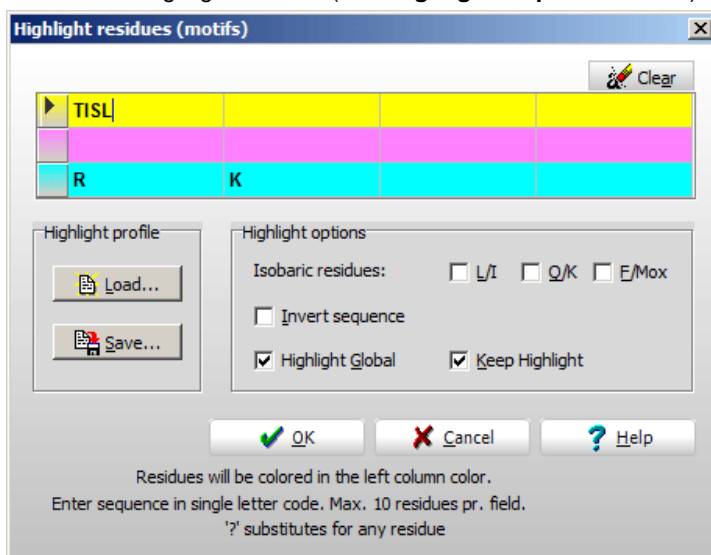
**Note:** The sequence will be copied in either 1- or 3-letter code depending on the setting of the display.

**Make fragment window:** The default selection range for the fragment window will be the currently active highlighted area. The range can be modified before creating the fragment window (see chapter 3.7 below).

**Underline residues:** A highlighted sequence can be used as the input for underlining a range in the sequence (use the pop-up menu). See chapter 3.4 below.

**Highlight residues (motifs)****3.3**

The menu command '**Highlight residues (motifs)**' is not to be confused with the mouse-driven highlight function (see '**Highlight sequence**' above).



The '**Highlight residues**' command enables you to color short sequence stretches that can be up to 10 residues in length. As you can also include 'wildcards' (e.g. any residue) you can also use the function as a motif search function.

In each of the 4 x 3 cells of the table you can enter any sequence motif (up to 10 residues) using 1-letter code. The highlight that will color each motif is shown in the left-most column. The highlight colors are defined in 'Setup color' (Chapter 5.3).

A question mark ('?') may substitute for any residue. In the above example the all sequences are searched for all occurrences of the typical N-glycosylation motif: Asn followed by any residue followed by Ser, Thr or Cys. The basic residues Lys and Arg (tryptic cleavage sites) are colored in a different color (yellow).

The Highlight residues (motifs) command is typically used to get a quick overview of the distribution of specific residues, occurrence of a specific sequence or to search for sequence motifs. As the coloring is persistent, that is it 'follows' the sequence into the peptide window (Chapter 9.4), it can also help you get a quick overview in daughter windows.

You may select the '**Highlight residues**' command from any sequence or daughter window. This will color all windows or just the selected sequence and related daughter windows depending on the state of the '**Highlight global**' option.

### 3 – Sequence window

#### Options:

**Isobaric residues:** When either of these options is checked, the corresponding isobaric residues are counted as identical. You may, typically through ms/ms analysis (Ch. 10.1), acquire a sequence tag where you have the mass difference of 113. This can be either Isoleucine or Leucine (with the same chemical composition as their mass values are identical). Instead of entering all possible combinations of L and I you can just check the L/I box and both residues will count as one.

**Invert sequence:** If checked, all sequences will be searched in both directions. I.e. the sequence DVTL above will also highlight LTVD.

**Highlight global:** All sequence windows on the desktop are searched for motifs to highlight even if they are not selected.

**Keep highlight:** The highlight motifs are saved between each highlight call. If not checked, the highlight dialog will be cleared upon exit.

**Highlight profiles:** The contents of the highlight table can be saved to disk as a highlight profile (in .PRF files, Appendix A).


**Clear:** Clears the table.

#### Quickcolor menu

The 'Quickcolor' command in the main menu is a fast way of coloring the most common residues or combinations:

Basic (R/K)	<i>Tryp</i>
Acidic (E/D)	<i>Endo Glu-C (wide range)</i>
Aromatic (W/F/Y)	<i>Chymotrypsin</i>
N-glycosylation	<i>NxT, NxS, NxS</i>
Cysteine (C)	<i>Disulphide bridges</i>
Methionine (M)	<i>CNBr cleavage</i>
Lysine (K)	<i>Endo Lys-C</i>
Arginine (R)	<i>Endo Arg-C</i>
Glutamic acid (E)	<i>Endo Glu-C (narrow range)</i>
Aspartic acid (D)	<i>Endo Asp-N</i>

This command is also available as a drop-down menu next to the 'Highlight' button in the main toolbar

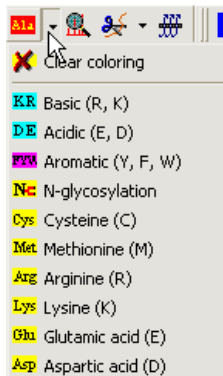
(chapter 1.4) 

The quickcolor command inserts the appropriate residue in the 'Highlight residue' table (see above). The program looks for the first free 'row' in the table to use for coloring. If all rows are full, the last row is replaced by the selection.

**Clear coloring.** Clears all color highlighting, both 'quickcolor' and 'highlight residues', and redraws the sequence window.

#### Color individual residues.

If you double-click on a residue, the 'Insert modification' dialog box opens



### 3 – Sequence window

(see Ch. 3.6). At the bottom of the dialog box there are four panels, one white and three colored ones for highlighting residues.

Clicking on the left-most panel clears any coloring for the selected residue. Clicking on any of the colors will change the background color of the selected residue.



**Note:** When you select Highlight motif or QuickColor, the coloring of individual residues (chapter 3.6) will be removed.

#### Marked residues.

The marked residues are a special case of coloring a residue. From the main menu you select **QuickColor|Marked residues**, a dialog box informs you to left-click on residues to mark, right-click in the sequence window to end 'marking' of new residues. If you want to delete an already marked residue, just left-click it again.

Marked residues are persistent if the sequence is saved after 'marking'. You are not warned that the sequence has been changed after marking.

Marked residues are shown with a line around each marked residue. The color of the line is the 'Aux1' color in the 'System colors' setup (Ch. 5.3).

42 1001 1002 1003 1004 1005  
:n-Lys-Pro-Phe-Leu  
11-Asn-Phe-Gln-Asp  
:u-Ser-Lys-Thr-Pro  
10-Tyr-Thr-Ile-Met

#### Underline residues

3.4

The underlining of residues is a way of emphasizing part of the sequence in relation to the rest. Underlining is a permanent feature, meaning that it remains as part of the display when changing between different display modes, but it is not saved along with the sequence. Underlined residues are displayed in red in addition to being underlined.



**Note:** Underlining residues may obscure the coloring of modified residues that are also colored red.

The most common use of underlining is to calculate the **coverage** of a digest or mass search, but the feature should be flexible enough for other purposes.

From the **digest mass search** result window (Chapter 8.5) you can retrieve sequence 'hits' from the database and display the protein sequence in a new sequence window. The peptides identified in the digest mass search will be underlined in this window.

When performing **mass searches** (Chapter 6.1) you can emphasize peptide 'hits' by double clicking on the relevant line, or select underlining from the local menu of a selected peptide. This will redraw the line in bold letters in the search results and send a message to the parent sequence window resulting in underlining of the corresponding peptide(s). A special command in the pop-up menu will bold all peptide 'hits' and underline the corresponding peptides in the parent window. This command is most useful when the **'Fit to enzyme'** option has been turned on and there are only relatively few hits.

From the sequence window you can modify underlining from the child windows or you can manually set underlines through the **Edit|Underline**

### 3 – Sequence window

or the corresponding item in the pop-up menu. Four sub-menu items are available:

**Underline range:** Through a dialog box you enter the first and last residue to be underlined.

**Underline highlight:** If you have highlighted part of a sequence, you can turn the highlighted part into underline through this command.

**Clear underline:** Removes all underlines from the sequence.

**Clear highlighted ul:** Clears a highlighted area for all underlines.

When open a **coverage window** (chapter 9.6) from the ms/ms search you will also open a normal sequence window with all identified peptides underlined. In the left-hand information frame (see above, 3.1) you can get information on each peptide when you move the mouse cursor above the sequence.

#### Cross-links

3.5

-Gla-Gla-Ala-Arg-Gla-Val<sub>3</sub>-Phe-Gla-Asp-Ala-Gla-Gln-  
-Cys-Tyr-Lys-Asp-Gly-Asp-Gln-Cys<sub>5</sub>-Glu-Gly-His-Pro-  
-Cys-Lys-Asp-Gly-Ile-Gly-Asp-Tyr-Thr-Cys<sub>7</sub>-Thr-Cys-  
-Lys-Asn-Cys-Glu-Phe-Ser-Thr-Arg-Glu-Ile-Cys-Ser<sub>9</sub>-  
-Asp-Gln-Phe-Cys<sub>10</sub>-Arg-Glu-Glu-Arg-Ser-Glu-Val-Arg-

Cross-links are displayed in the sequence window as red lines going from one residue to another. Up to 30 cross-links can be defined for each sequence. In order to differentiate between different cross-links, the lines are in three different shades of red. Furthermore, each color has a different vertical offset.

The **'SS'** button in the status bar controls the display of cross-link lines and how the mass of Cys residues is calculated. When depressed (the button shows **'SH'**) the cross-links are shown as gray lines, and Cys residues are calculated in the reduced form. When the button is in the up state (legend shows **'SS'**) the mass of Cys residues is calculated as the oxidized form and cross-links are shown.



**Note:** Cross-links are not restricted to Cys residues (see below). When there are cross-linking residues other than Cys, the display color is still controlled by the **'SS'** button.

Cross-links are edited by selecting **Edit | Edit cross-links** from the menu or right-click on the relevant sequence window and select from the pop-up menu.

Links are defined as coming from Cys-1 going to Cys-2. In the main sequence window, all residues to cross-link are highlighted. The residue can be selected in the drop-down box labeled **'Residues to highlight'** (default is cysteine, but any residue can be selected). Press the **Update** button to change coloring and refresh the linkage lines in the sequence window.

### 3 – Sequence window

X	Cys-1	Cys-2
1	77	86
2	99	115
3	114	125
4	147	192
5	191	200
6	223	269
7	268	276
8	288	302
9	301	312
10	339	384
11	383	392
12	415	461
13	460	471
14	484	500
15	499	510
16	581	590
17	537	582

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

Met 1

SS profile:  
Load Save

OK  
Cancel  
Help

Line colors:  
[Color swatches]

If you click on a residue in the sequence window, the position of this residue will be entered in the next available slot in the table. This makes it fast to enter all necessary linkages. The value of the residues can also be entered directly in the table.

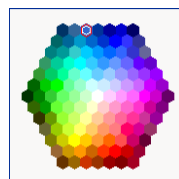
**SS profile:** If you work with multiple sequences with the same cysteine linkage pattern (e.g. IgG's) you can save the pattern as a file to disk. It is the pattern, and not the actual locations that are saved, i.e. linkage of Cys-1 to Cys-4 etc. and not residue 23 to residue 44. This allows you to load a pattern and GPMW will then assign residue position based on the occurrence of Cys residues.

If the chemical composition of the cross-link alters the mass of the cross-linked residues, you should combine the cross-link with the '**Amino acid modifications**' option discussed below. The mass difference of Cysteine vs. Cystine cross-links (i.e. 2 Da.) is catered for automatically through the '**SS**' button (Chapter 1.4). The '**SS**' button also determines whether drawing of the red cross-links in the sequence window takes place.

In printouts the linked residues are listed when the '**Extended info**' options have been selected. The links are shown as colored lines when the '**Color print**' option is selected.

The easiest way of specifying disulfide bonds is to import a record from the Swiss-Prot (UniProt) database where they are specified. Please see Chapter 3.9 for details on how to import disulfide bonds.

In the sequence window, the linkages are drawn with



### 3 – Sequence window

colored lines with a repeat of 4 in order to make it easier to differentiate different links (i.e. link number 1 will have the same color as link number 5). The color of the links can be defined by clicking the colored boxes at the bottom of the dialog and select a new color from the color hexagon.

#### Amino acid residue modifications

3.6

In GPMW you can specify chemical modifications for individual amino acid residues. This function is in addition to the modified residues that may be defined in the residue mass files (chapter 4.2). These works on all residues of a given type (i.e. each is defined by a single letter code). You may thus define up to 31 residues, where the first 20 usually are the standard amino acid residues. The chemical modifications are additions to the individual residues and are linked to a position in the sequence. Each sequence may contain up to 20 chemical modifications. In addition you can define changes to the N- and C-terminus (Edit sequence – chapter 4.1). You may also define cross-links, but only for disulfide bridges do the cross-links by themselves contain information like mass and composition. Cross-links can be combined with individual modifications in order to cover all kinds of cross-linking. If you perform cross-linking experiments and search for the resulting cross-linked peptides after digestion, you should check Chapter 6.3.

An amino acid modification (individual, N- and C-terminal) is defined by a name, an elemental composition and pKa value (optional). The name is not obligatory either, but entering a name makes it easier to navigate the sequence. When you define an amino acid modification in the modification file, you can optionally specify that the modification is restricted to specific residues. For the mass files you need to define 1- and 3-letter code in addition to mass and elemental composition.

#### Reporting modifications

Modifications are reported on most printouts, and can be seen in the sequence as:

*Single residue modification:* The residue is painted red and when the mouse points to the residue, the name and composition is shown in the toolbar of the sequence window (right-hand panels with a light yellow background).

*N- and C-terminal modifications:* The name is shown in the sequence window toolbar by default. The fly-by help shows the elemental composition.

*Residue mass file:* The name of the currently loaded residue mass file is shown in the main window toolbar. **Note:** The residue mass file is global for all sequence windows (i.e. when you change residue mass file, all mass and composition values are re-calculated in all windows).

#### What modification type to use

When you have a few residues modified or many different modifications, you should choose single residue modification. If on the other hand, you have a large number of identical modifications (e.g. hydroxylated proline in collagen) it makes more sense to specify a 'new' residue to replace the ones that are modified.

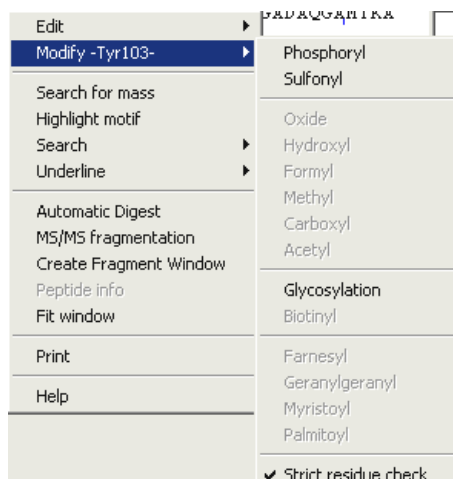


### 3 – Sequence window

If you have residues that 'change' during the course of your experiments you can make two mass files both with the 'extra' residues, but one file without modifications and the other with. This is the idea behind cysteine modifications where you have a mass file for each type of cysteine modification (e.g. aa\_mass for the default values, pe\_cys for pyridylethylated Cys, ae\_cys for amino ethylated Cys etc.).

#### Insert simple modification

The '**Simple modifications**' are a number of modifications that are hard-coded into GPMAW. When you right-click the mouse in a sequence window, you bring up the pop-up menu that presents you with a number of context-sensitive menu options. The second option on the list is to modify the given residue. The actual residue that is to be modified is shown after the '**Modify**' command, e.g. Modify – Lys228-. Select this option to open a submenu with a list of the modifications that are possible for this particular residue type. Selecting an option will insert the selected modification into the chosen residue.



You can enable all modifications on the list for all residues by removing the checkmark from the last option on the submenu '**Strict residue check**'. When you next access the '**Modify**' command, all modification options are available for any residue.

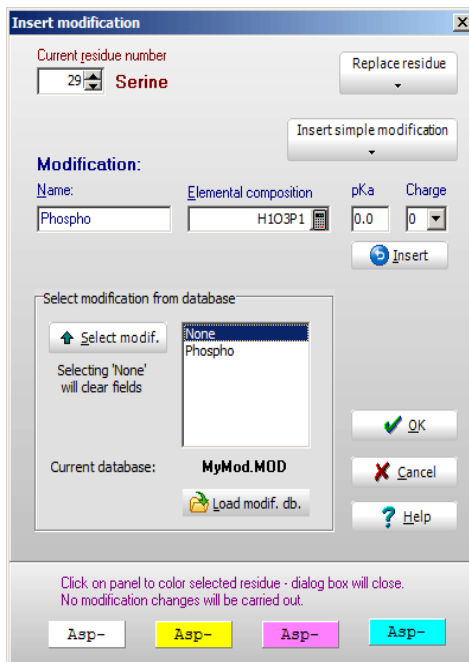


**Tip:** If you hold down the <Ctrl> button when you right-click on a residue, you will open the '**Insert simple modification**' menu directly, thus making insertion of multiple modifications faster.

If you select the **Glycosylation** option, you will open the **Glycosylation wizard**, which eases the insertion of complex carbohydrates. For more details of this wizard, please see Chapter 4.3.

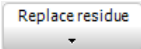
### 3 – Sequence window

#### Single residue modifications



If you **double-click** on a residue, you will bring up the ‘**Insert modification**’ dialog box. The same dialog can be accessed by selecting **Edit | Edit modification** from the menu. Here you can either specify individual modifications or color a residue. If the residue selected is already modified, the modifications will be entered in the ‘**Name**’ and ‘**Elemental composition**’ edit boxes:


**Residue:** The number of the residue to be modified will be shown in the ‘**Residue number**’ field at the top. To the right of this number, the full name of the residue will be displayed. If the ‘**Edit modification box**’ was activated by a double-click on a residue, the number displayed will be the residue clicked upon, otherwise it will be number 1 (the first residue). The residue numbers can be changed directly by editing the number or by clicking the up/down arrows next to the number. The residue name will change to show the sequence residue.

**Replace residue:** This is a drop-down menu  that enables you to replace the currently selected residue with any of the 20 standard residues. The insert modification dialog box will not close upon selection thus enabling you to change to another residue for changes/modifications.

**Modification:** Modifications can be entered directly into the ‘**Elemental composition**’ box or, if a modification file has been opened, you can select from the modifications available in the drop-down box at the bottom of the

### 3 – Sequence window

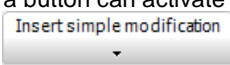
dialog box, followed by **'Select modif.'**. The **'Name of modification'** is

optional, but useful for reference. Clicking on the calculator  opens the 'Elemental composition calculator'; see Chapter 4.4 on how to enter elemental compositions. The **pKa** and **charge** are optional and you need only enter them if you plan on using the pI or charge of the protein or derived peptides. For the pI you can enter values between 0.1 and 14 while the charge can be either -1 or +1. Both values have to be different from 0 for the fields to be active. See also definitions in Edit mass file and Edit modification file (Chapters 4.2 and 4.3).




**Note:** When entering an elemental composition, be sure that all the atoms are defined in the currently active mass file (chapter 4.2).


Below the edit boxes a button can activate a drop-down menu for selecting

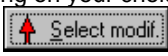
 Insert simple modification

simple modifications. This menu item is identical to the **'Insert simple modification'** from the pop-up menu described above. Selecting an item from this list will insert the selected modification in the edit boxes.

If you press the  **Insert** button, the modification will be inserted, but the dialog box will not close, pressing the **'OK'** button will insert the modification and close the dialog.

By using modification databases (see Chapter 4.3 - **Edit | Edit modifications**) you can have any kind of modification readily available for changing a sequence.

Start by loading a modification database file . The modifications available in the file will be listed in the selection list. Here you can select a modification either by double clicking on your choice or by

making a selection followed by **'Select modif'** . The selection will be entered into the two edit boxes above where they can be modified before entered into the sequence by selecting **'OK'**.

Modified residues are displayed in the Highlight 4 color (Chapter 5.3).



**Note:** Be sure to define the first three highlight colors different from highlight 4 (in **Setup | Setup system | Colors**), as you will be unable to see the residue, if it is part of a colored motif (Chapter 3.3).



**Tip:** For the first and last residue it is better to modify the N- or C-terminal in the **Edit | Edit sequence** as these modifications do not count as one of the limited twenty individual modifications.

#### Color selected residue

Clicking on one of the bottom colored panels enables you to color the background of the selected residue in the indicated color. **Note:** This action

### 3 – Sequence window

results in closure of the **'Modification'** dialog box and any modifications entered will not be executed.

Clicking on the white left-most field will clear background coloring for the selected residue.

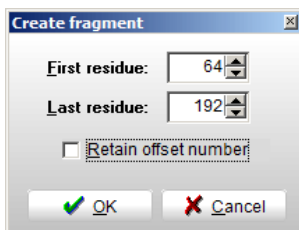
The color single residue command can be useful for drawing attention to a single residue. The coloring is persistent and will be carried on to peptide windows.

#### Fragment window

3.7

The 'Fragment window' option is a fast way of creating a new sub-sequence based on an existing sequence. A common usage of this function is when a pre-sequence has been loaded from a database and you want to work with the active protein. Alternatively, you may need to work with a smaller peptide, but find it inconvenient to work through a sequence window.

Select **Cleavage|Create fragment window** (or right-click and select from the local menu) and the following dialog box opens:



If an area is highlighted, the first and last residue of the highlighted area will be displayed as default. If no highlight exists, the value will be 1 and the last residue of the sequence. The values can be edited directly or you can click the up/down arrows. When selecting **'OK'**, a new sequence window will open containing the selected part of the original sequence. The name of the new window will be 'Fragment 51-90 of ' + the original sequence name.

If you check the **'Retain offset number'** box, the first residue of the sequence fragment will start with the number in the **'First residue'** box + the original start value - one. The offset number can be edited in the usual edit sequence dialog (Chapter 4.1). When an offset number has been specified, the color of the number in the residue number label of the sequence window will change to red.



**Note:** The new window will be an independent parent window that can be saved as a sequence, searched for mass, etc. When the original sequence window is closed, the fragment windows will remain on the desktop.

You have to edit the sequence name to remove the automatic addition of 'Fragment...'.

#### Sequence information

3.8

The **'Sequence information'** dialog box contains three pages labeled **'Sequence info'**, **'Composition'** and **'Masses'**, respectively. You switch

### 3 – Sequence window

between the various pages by clicking on the tabs. The first two options have their own menu entry (under **Info**) while in the last option you have to select from open dialog box.

Some of the information in this dialog is also present in the information frame of the sequence window (Chapter 3.1).

#### Sequence info.

The **Info | Sequence info** menu entry opens a multipage dialog box on the 'Sequence information' page.

Sequence information - SERUM ALBUMIN PRECURSOR - Bos taurus (Bovine)

Sequence info | Composition | Mass values | Calculations | Isotope

Name: SERUM ALBUMIN PRECURSOR - Bos taurus (Bovine)

bsa.seq Pos: 1

Properties		Sequence	
Average mass :	69431.4376	Residues:	607 Offset: 0
Monoiso. mass:	69387.1154	Chains:	1
Molar ext. coeff.(280nm):	49915	Modifications:	3
Molar absorbance:	0.719	Crosslinks:	17
Hydrophobicity (GRAVY):	-0.429	Inverted:	0 [0.0%]
Theor. pI (SS/SH):	(1) 5.78 / 5.77	Underlined:	0 [0.0%]
	(2) 5.80 / 5.79		
	(3) 5.95 / 5.95		

Print Copy OK Help





The page shows statistics for the currently selected protein sequence:

- The full name of the sequence.
- The origin file and position in the file.
- Average and monoisotopic mass (four decimals).
- Molar extinction coefficient at 280 nm. The values are based either on Gill and von Hippel or on Pace et al. as selected in the Setup dialog (see section 5.1).
- Molar absorbance based on the molar extinction coefficient calculated above.
- Theoretical pI. Values for both oxidized (SS - disulfide bonded) and reduced (SH) cysteine are shown. Three different values are reported, each based on a different table (1 – Skoog & Wichmann; 2 – Free amino acids; 3 – Rickard, Strohl & Nielsen). In the Setup (Chapter 5.6) you can set which tables to use in general calculations of the pI. See also Appendix C.7.
- Number of residues.
- Number of chains.
- Number of modified residues.
- Number of cross-links.
- Number of residues and percentage of highlighted residues.
- Number of residues and percentage of underlined residues.

### 3 – Sequence window

#### Composition

Sequence information - SERUM ALBUMIN PRECURSOR - Bos taurus (Bovine)								
Sequence info Composition Mass values Calculations Isotope								
Res	#	%	Res	#	%	Res	#	%
Xxx	0	0.00	Pro	28	4.61	Leu	65	10.71
Asp	40	6.59	Gly	17	2.80	Tyr	21	3.46
Asn	14	2.31	Ala	48	7.91	Phe	30	4.94
Thr	34	5.60	Val	38	6.26	Lys	60	9.88
Ser	32	5.27	Cys	35	5.77	His	17	2.80
Glu	59	9.72	Met	5	0.82	Trp	3	0.49
Gln	20	3.29	Ile	15	2.47	Arg	26	4.28
C	3072	31.81	O	933	9.66	Br	0	0.00
H	4795	49.65	P	2	0.02	Cl	0	0.00
N	816	8.45	S	40	0.41	F	0	0.00

 Print  Copy  OK  Help

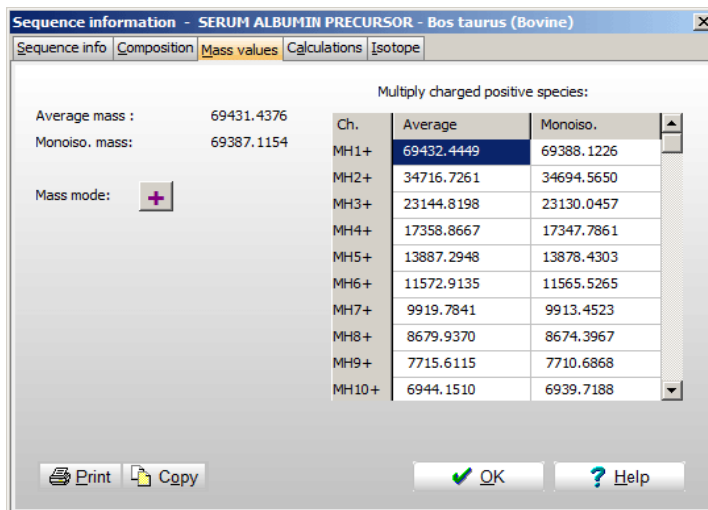
The **Info | Composition** menu entry displays the amino acid and atomic composition of the currently selected protein sequence.

The table at the top lists all amino acid residues with a 3-letter code, number of residues (#), and percent (%) counted as number of residues. All x's and 'extra' defined residues (see Edit mass file, Chapter 4.2) are grouped under the Xxx entry.

The bottom part of the dialog lists the composition of the atoms defined in the current mass list (Chapter 4.2) as number of atoms (#) and percent (%).

### 3 – Sequence window

#### Masses - multiply charged ions



The 'Masses' section of the 'Sequence information' dialog box shows a table of the multiply charged ion species for both monoisotopic and average masses up to 30 charges. Although the monoisotopic mass will most probably not be very useful, it is included for completeness.

The '+/-' button will toggle the table between positively charged and negatively charged ion species.

#### Print

Print will make a hardcopy of both the '**Sequence info**' page and the '**Composition**' page when the focus is on either of the two first pages. When the focus is on the last page, '**Masses**', you will get a hardcopy of the multiply charged ion species.

#### Copy to clip

Pressing the '**Copy to clip**' button will copy the information in the displayed page onto the clipboard (not the 'Calculations' page).

### 3 – Sequence window

#### Mass conversion.

Sequence information - SERUM ALBUMIN PRECURSOR - Bos taurus (Bovine)

Sequence info Composition Mass values **Calculations** Isotope

Mass conversion

☒ pMol -> ng    ☐ ng -> fMol  
☐ nMol -> ug    ☐ ug -> pMol  
☐ uMol -> mg    ☐ mg -> nMol

1 pMol = 69.431 ng  
2 pMol = 138.863 ng  
5 pMol = 347.157 ng

0.00 pMol

→

Molar ext. coeff. 49915 cm<sup>-1</sup> M<sup>-1</sup>  
Molar absorbance 0.719 cm<sup>-1</sup>  
A: 0.00 →

Calculations are based on average mass

OK Help

The dialog also contains a Picomole to mass converter. The aim of the calculator is to quickly calculate the conversion of pMol to ng (nMol to  $\mu$ g;  $\mu$ Mol to mg) and vice versa.

You start by selecting the correct mass range (e.g. pMol or nMol) and then enter the exact amount to be converted in the left-hand edit field. Values are immediately converted. 'Pre-calculated' amounts of 1, 2 and 5 are listed for a quick overview.

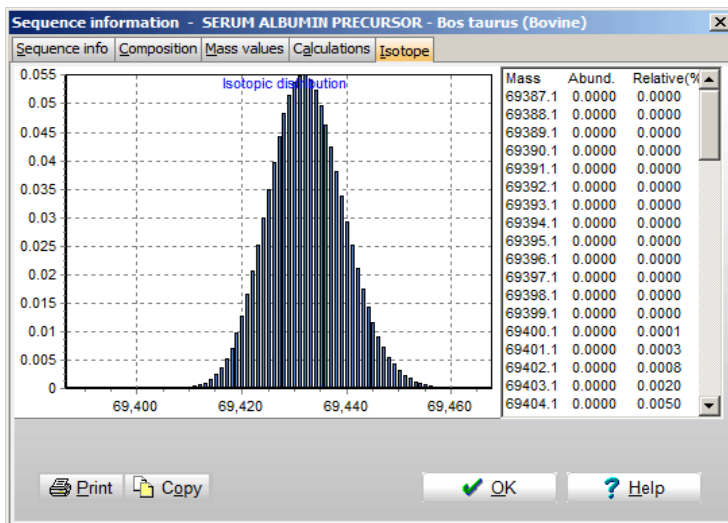
The mass conversion feature is similar to the one present in the 'Protein explorer', Chapter 2.9.

The molar absorption calculator works by entering the absorption and pressing the arrow button to read the concentration in mg/ml (=  $\mu$ g/ $\mu$ l).



### 3 – Sequence window

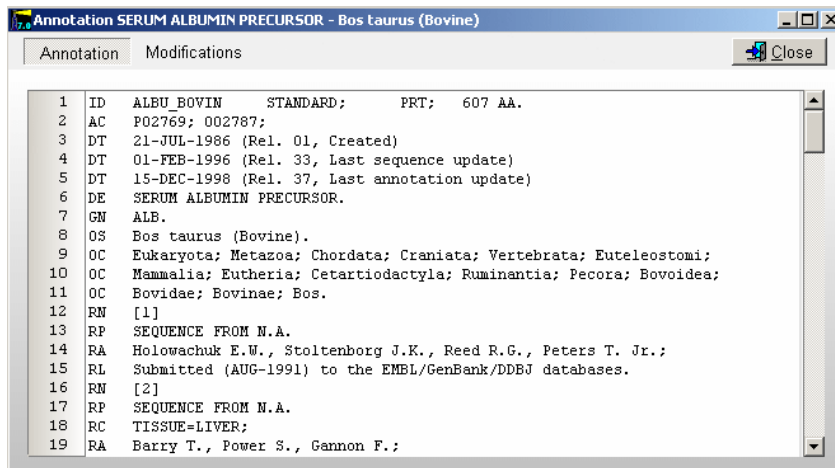
#### Isotopic distribution




The last page in the dialog box is the isotopic distribution of the protein. The main part of the window is taken up by a graph showing the isotopic distribution. This graph can be zoomed in the standard way for GPMW graphs (click and drag down and right to zoom in, click and drag up and left to zoom out).

The right-hand part of the dialog is taken up by a list of the individual isotopic peaks with absolute and relative abundance. Please note that the mass value is only approximate.

The isotopic distribution only works up to proteins with a size of less than 700 residues.



The annotation window is an editable text window that can contain any kind of text. The annotation page will be saved along with the sequence. You can view the annotation by selecting **Info|Annotation...** or pressing the

**'Annotation'** button  in the sequence window toolbar.

The color of the **'Annotation'** button changes with the content:

- Gray:** There is no content on the 'Annotation' page.
- Red:** There is text on the 'Annotation' page.
- Green:** The 'Annotation' contains a Swiss-Prot entry (with an accompanying 'Feature table'.
- Blue:** The 'Annotation' contains a GenPept (Entrez) entry.

## Annotation page

Although you can put any kind of text on to the annotation page, it is particularly useful when you read a sequence in Swiss-Prot or GenPept format. When you import a sequence via the **File|Import ASCII (from file or from clipboard)** command (chapter 2.5) you are given the choice of saving the intact record to the annotation window. If you read a sequence from the indexed Swiss-Prot database (chapter 2.6), the complete entry will automatically be placed in the annotation (if the full database is present).

Records in Swiss-Prot format are parsed onto the 'Feature table', see below. Records in GenPept format (e.g. Entrez) will be recognized in the near future.

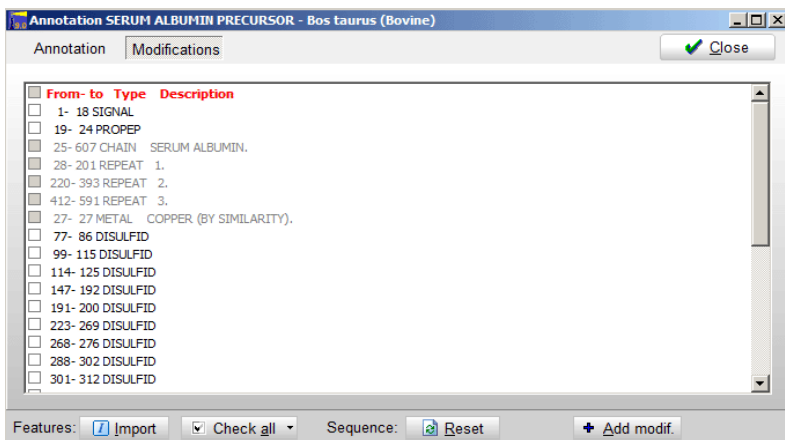


**Note:** When changes have been made to the annotation page, you have to save the sequence in order to save the annotation. You are not warned

### 3 – Sequence window


about losing information on the annotation page when you close the window!

#### Feature table

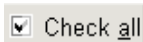


The feature table is a translation of the FT section of the Swiss-Prot record. The main function of the 'Feature table' is to allow easy import of posttranslational modification into GPMaw sequences. Most of the well-defined modifications defined in Swiss-Prot can be imported (i.e. modifications like 'phosphorylation' can be imported as there is only one of this kind in a given situation, however, a modification like 'glycosylation' cannot be imported, as the actual modification is not defined – and may further be variable).

To import modifications you check the appropriate boxes followed by

pressing the **'Import'** button . This action transfers the modifications to the sequence record and closes the 'Annotation' window.

You can check all recognized features by pressing the **'Check all'** button



The following features are recognized:

SIGNAL, PROPEP – this part of the sequence is deleted.

DISULFID – cross-links are created.

The following secondary modifications are recognized. The GPMaw translation is shown in upper case in brackets, followed by the name and composition as inserted by GPMaw. On the following lines are the UniProt names.

[ACETYLATION], Acetylation, C<sub>2</sub>H<sub>2</sub>O<sub>1</sub>,

N-acetylalanine, N-acetylaspartate, N-acetylcysteine, N-acetylglutamate, N-acetylglycine, N-acetylmethionine, N-acetylproline, N-acetyls erine, N-acetylthreonine, N-acetyltyrosine, N-acetylvaline, N<sup>2</sup>-acetylarginine, N<sup>6</sup>-acetyllysine,

[AMIDATION], Amidation, -C<sub>1</sub>+N<sub>1</sub>,

### 3 – Sequence window

Alanine amide, Arginine amide, Aspartic acid 1-amide, Asparagine amide, Cysteine amide, Glutamic acid 1-amide, Glutamine amide, Glycine amide, Histidine amide, Isoleucine amide, Leucine amide, Lysine amide, Methionine amide, Phenylalanine amide, Proline amide, Serine amide, Threonine amide, Tryptophan amide, Tyrosine amide, Valine amide,

[FORMYLATION], Formylation, C1O1,

N-formylmethionine, N-formylglycine, N6-formyllysine,

[HYDROXYLATION], Hydroxylation, O1,

3-hydroxyasparagine, 3-hydroxyaspartate, 3-hydroxyproline, 3-hydroxytryptophan, 4-hydroxyarginine, 4-hydroxyproline, 5-hydroxylysine, Hydroxyproline,

[PHOSPHORYLATION], Phosphorylation, P1O3H1,

4-aspartylphosphate, Phosphoarginine, Phosphocysteine, Phosphohistidine, Phosphoserine, Phosphothreonine, Phosphotyrosine, Pros-phosphohistidine, Tele-phosphohistidine,

[SULFATION], Sulfation, O3S1,

Sulfotyrosine, Sulfoserine, Sulfothreonine,

[GAMMA-CARBOXYGLUTAMIC], Gamma-carboxyglu, C1O2,

4-carboxyglutamate,

[METHYLATION], Methylation, C1H2,

2-methylglutamine, 5-methylarginine, Cysteine methyl ester, Glutamate methyl ester (Gln), Glutamate methyl ester (Glu), Leucine methyl ester, Lysine methyl ester, Methylhistidine, N4-methylasparagine, N5-methylarginine, N5-methylglutamine, N6-methylated lysine, N6-methyllysine,

Omega-N-methylarginine, Omega-N-methylated arginine, Pros-methylhistidine, S-methylcysteine, Tele-methylhistidine,

[DEAMIDATION], Deamidation, -H1N1+O1,

Deamidated asparagine, Deamidated glutamine,

[CITRULLINE], Citrulline, -H1N1+O1,

Citrulline (Keyword: Citrullination),

[N-METHYLATION], N-methylation, C1H2,

N-methylalanine, N-methylisoleucine, N-methylleucine, N-methylmethionine, N-methylphenylalanine, N-methyltyrosine,

[DIHYDROXY], Dihydroxy, O2,

3,4-dihydroxyphenylalanine, 3,4-dihydroxyarginine, 3,4-dihydroxyproline, 4,5-dihydroxylysine,

[DIMETHYLATION], Dimethylation, C2H4,

Asymmetric dimethylarginine, N,N-dimethylproline, N4,N4-dimethylasparagine, N6,N6-dimethyllysine, Symmetric dimethylarginine,


[ADP-RIBOSYL], ADP-ribosyl, H21C15N5O13P2,


ADP-ribosylasparagine (Keyword: ADP-ribosylation), ADP-ribosylarginine (Keyword: ADP-ribosylation), ADP-ribosylcysteine (Keyword: ADP-ribosylation), ADP-ribosylserine (Keyword: ADP-ribosylation)

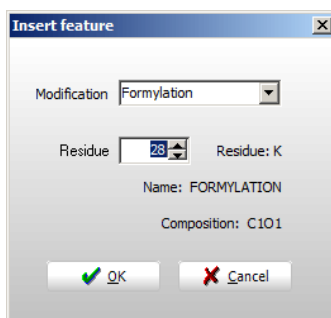
Features like chain, domain, site, glycosylation, and lipid, are not recognized as they either do not have a counterpart in GPMaw or they do not represent a specific chemical modification.

### 3 – Sequence window

**Important:** As GPMW does not check the ‘correctness’ of the assignment of imported modifications, it is important that the import is carried out on the intact protein. When importing signal and propeptides (i.e. removing the peptides) together with secondary modifications, the removal of residues is carried out last, so the chemical modifications go to the ‘correct’ residues.

The **‘Reset’** button  **Reset** reloads the sequence from the annotation page (i.e. removes all changes to the sequence).

The **‘Add modification’** button  **Add modif.** opens a small dialog, which allows you to enter a modification into the annotation page, from where you can easily select/deselect as described above:



The 'Insert feature' dialog box contains the following fields and controls:

- Modification:** A drop-down menu currently showing 'Formylation'.
- Residue:** A numeric input field showing '26'.
- Residue:** A text label showing 'K'.
- Name:** A text label showing 'FORMYLATION'.
- Composition:** A text label showing 'C101'.
- Buttons:** 'OK' (with a green checkmark icon) and 'Cancel' (with a red X icon).

Here you select the modification in the drop-down box where you can select from the standard UniProt ones described above. The composition will be shown in the ‘Composition’ field (cannot be edited). You then select the residue number in the number field below. As the number changes, the corresponding residue will be shown to the right. The **‘OK’** button only becomes active when the residue number has changed. The modification will be inserted into the annotation as an FT field assigned as MOD\_RES and the appropriate modification.

When the ‘Annotation’ window opens, the sequence from the sequence window is checked against the sequence in the annotation window. If there are discrepancies between the sequence lengths, first or last residues, you are given a warning in the top part of the annotation notebook.

Annotation	Feature table	Warning: Sequence length differs from annotation!
------------	---------------	---

In this case you should reset the sequence before importing modifications.



**TIP:** As the ‘Feature table’ is calculated from the annotation you can enter your own modifications in the ‘Annotation page’. Start a new line with FT followed by three spaces before entering the location, a space, MOD\_RES, and the modification. Remember to save the sequence. When you next open the annotation window, the new modification will be present on the ‘Feature page’.

### **3 – Sequence window**

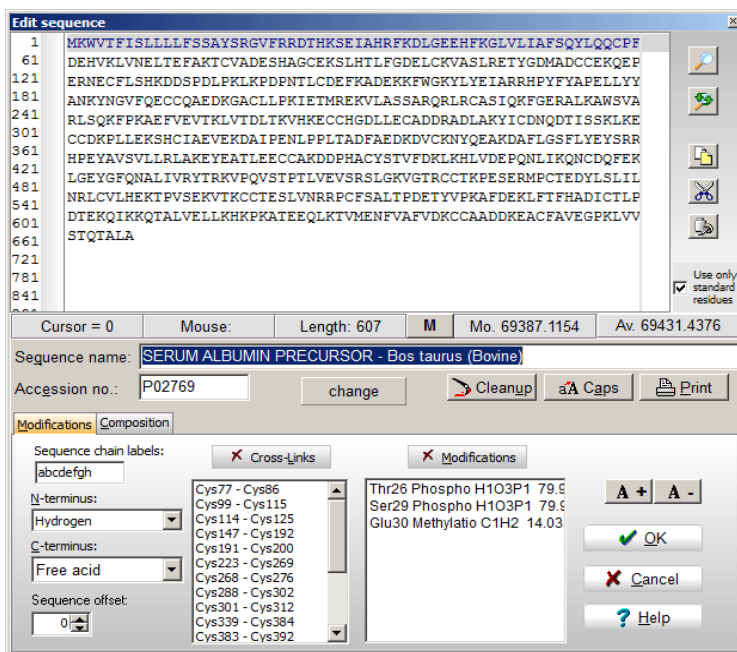
If you save the sequence 'intact', the Feature table can thus be a quick way of looking at your sequence with and without modifications.

## Edit

Editing protein sequences, mass and modification files

### Edit sequence

4.1



You can edit the sequence of the currently selected sequence window by

selecting **Edit | Edit sequence**, pressing the **'Edit sequence'** button in the toolbar or by right-clicking on the window and select **'Edit | Edit sequence'** from the local pop-up menu.



**Note:** The currently active window has to be a sequence window before you can edit the sequence. However, you can always start editing a new sequence – this will create a new sequence window, see end of section.

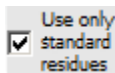
The **sequence** is edited in the large multi-line editor in the top part of the dialog box. The sequence can only be edited in **1-letter code**. You have to exit to the sequence window in order to view the sequence in 3-letter code. The editor supports cut and paste, meaning that you can copy sequences to

## 4 - Edit

the clipboard from other applications and paste them into the editor. You can also highlight and use cut, copy, and paste inside the editor.



**Note:** You have to use keyboard shortcuts, e.g. Ctrl-X or Sh-Del for **cut**, Ctrl-C or Ctrl-Ins for **copy**, Ctrl-V or Sh-Ins for **paste**. Alternatively, you can use the pop-up menu (right-click in the edit box).



The residues you can enter (or paste) in the edit box are controlled by the check-box in the right-hand margin below the paste button. If the box is **checked**, the sequence editor will only accept the 'standard' 1-residue notation (i.e. 'A' to 'Y'), if the box is **un-checked**, the sequence editor will accept all 1-letter codes defined in the currently selected mass file (see section 4.2). The default setting of this box can be done in Setup on the Systems page (chapter 5.1).

If you resize the dialog box, the sequence edit control will resize along with the dialog box. The rest of the dialog box controls will not change size or position.

The **name of the sequence** is edited in the edit line below the status panels. The maximum size of the name is 250 characters. If you need more information for the protein you can use the 'Annotation' page (**Info|Annotation** or the '**a**' button in the toolbar of the sequence window, see Chapter 3).

If you **paste** a sequence record in **FastA format** from the clipboard, the record will automatically be parsed into the sequence name (first line) and the sequence proper (the remainder of the record).

If you know the **accession number** of the protein you should enter it in the small edit box between the sequence and the name boxes. If you load from an indexed FastA database (Appendix B) the accession number will be loaded along with the sequence. If entered, the accession number will be shown in the sequence window title bar (eg. "[P35247] Pulmonary surfactant....").



**Note:** Although the accession number is not directly used by GPMW for identification of proteins you are strongly encouraged always to use the number as it is a unique identifier into the respective databases.

If you paste a sequence that is in single letter code but not in uppercase characters you have to convert it into upper case.

If the sequence contains extra non-sequence characters (e.g. numbers, spaces and carriage returns) just press the '**Cleanup**' button which removes all characters not defined as single letter characters in the current mass file. For more information on the mass files, please see the following chapter "Edit mass file", 4.2. The button is explained in detail below.

When importing sequences you can also use the **File|Import ASCII** functions, either as **import from clipboard** or **import from file** (Chapter 2.5).



## 4 - Edit

### Panels.

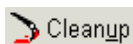
The panels just below the sequence editor show editing status and molecular mass information.

Cursor = 223	Mouse: 14	Length: 222	Mo. 24141.7769	Av. 24156.8206
--------------	-----------	-------------	----------------	----------------

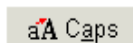
The first panel shows the position of the editor text cursor (the value is the number of the preceding residue), and the next panel the position of the mouse cursor. If part of the sequence is highlighted, the middle panel will show the first and last residue that is highlighted, otherwise the panel will show the total length of the protein. The last two panels show the monoisotopic and average masses of the intact protein as defined in the editor.

The bottom panel is a two-page notebook that shows either the modifications made to the protein or the amino acid and elemental composition of the protein. The composition panel is updated whenever a change is made in the sequence editor.

### Buttons and drop-down boxes



**Cleanup** Removes all characters in the edit box that are not defined in the current mass file (1-letter residue identifiers). This function is very useful when you paste a sequence from another application that contains numbers, space characters etc.



**Caps** Converts the text in the edit box to upper case (capital letters). This is necessary if you paste a sequence in lower case from another application. When you enter characters from the keyboard, they will automatically be converted to upper case.



**Print** Print the sequence. You can select 1- or 3-letter residue print-out. The printout is similar to printing from the sequence window (Chapter 3.1).



**Search** for a given sequence. A standard 'Search for' box opens, enabling you to locate a sequence of letters.



**Replace** residues. Opens a standard 'Search and replace' dialog enabling you to replace a given residue or sequence.



The three buttons on the right-hand side of the dialog box copies the following commands from the pop-up menu:

**Copy to clipboard** (Ctrl+C), **Cut to clipboard** (Ctrl+X) and **Paste from clipboard** (Ctrl+V). The keyboard shortcuts (in paranthesis) are only active when the focus is on the edit control (i.e. when you are actually editing the sequence).

Sequence chain labels:

**Sequence chain labels.**

abcdefghijkl

The edit box labeled 'Sequence chain labels' contains one character for each chain possible in GPMW

## 4 - Edit

(eight in total). The first character in this line will be the label of the first chain in the listed sequence. If you enter less than eight characters, GPMaw will fill in the rest automatically.

### N-terminus: / C-terminus:

You select the modification of each sequence terminal from either drop-down list box below. To edit the content of the drop-down list boxes you have to select **Edit|Edit mass file** from the main program menu and select the N-terminal or C-terminal tab (see below 4.2).



**Cross-links** Opens the 'Edit cross-links' dialog box (see Chapter 3.5) enabling you to modify cross-links. Cross-links are shown in the list box below the button.



**Modifications** Opens the 'Select modification' dialog box (see Chapter 3.6). Unlike when you double-click on the residues in the sequence window, the 'Select modification' dialog box always opens with residue 1 selected. Modifications are shown in the list box below the button.

Sequence offset:

The sequence offset enables you to specify that the numbering of the sequence should not start with one. This is typically used when you cut a sequence out from another sequence or when you are working with a pre- or pro-sequence. The offset number can be either positive or negative. When you have specified an offset, the residue number in the status panel will be shown in red numbers.



Enables you to change the font size in the edit sequence box in 1-point steps. The font is changed dynamically.

**Residues**

**Elemental**

Determines whether the table on the composition page shows amino acid residue composition or elemental composition. The table is updated for every change made to the sequence.

Xxx = 0	Pro = 13	Leu = 14
Asp = 13	Gly = 22	Tyr = 3
Asn = 7	Ala = 27	Phe = 7
Thr = 7	Val = 17	Lys = 5
Ser = 15	Cys = 6	His = 9
Glu = 15	Met = 2	Trp = 5
Gln = 11	Ile = 4	Arg = 20

### Edit new sequence

The **Edit|Edit new sequence** (toolbar button) is identical to the 'Edit sequence' command discussed above except that the name and sequence fields of the edit dialog box are initially empty. Furthermore, when the dialog box is closed, a new sequence window opens on the GPMaw desktop.

## 4 - Edit

The 'Edit new sequence' dialog box can be used as an alternative to the **File|Import ASCII|From clipboard** by pasting into the sequence edit box, removing all extra text and using the '**Cleanup**' and '**Caps**' buttons.



**Hint:** As the mass panels are updated for every entry in the edit box, you can use the editor to check the mass of a short peptide just by entering or pasting it into the edit box and modify it as appropriate. You only create a new sequence window when you select '**OK**'.

### Edit mass files

4.2

The **Edit|Edit mass files** command actually controls four different mass tables: The mass file, the N-terminal, the C-terminal, the atom mass table and the modification file. The mass tables are crucial for GPMaw to work. In addition to the mass value they also define the amino acid residues (name, 1- and 3-letter code). The mass files reside in the 'System' directory as defined in 'Setup – Directories', by default this is c:\gpmaw\system\.

Chapter 1.5 contains an overview of most of the essential tables used by GPMaw.

### How are mass values calculated by GPMaw?

The basic table is the *atom mass table*. This table defines the **average** and the **monoisotopic** mass of each atom used when calculating mass values. The table contains most of the commonly used atoms, but you may have to add specific ones.

Amino acid residues are defined in the *mass file*. Here you define the elemental composition of each residue along with name and abbreviations. This means that if you change the atomic mass, all residues will also change. The mass file is a separate file that you select through a drop-down box in the main toolbar. By selecting a new mass file, you can effectively change the mass of one or several amino acid residues in a single reproducible operation.

Modifications are also stored in separate *modification files*. Like the mass file, modifications are defined by elemental composition along with a name and other information. A modification file can only contain 30 different modifications, but you have as many as you want saved as different files. The modification file is loaded through the **Edit | Edit modification file** dialog. When you select a modification, the name and composition is stored along with the sequence, and the mass value is then added to (or subtracted from) the residue mass on-the-fly. As the modification information is copied to the sequence, changing modification file will not change modification already defined in the sequences.

*N-terminal* and *C-terminal modifications* can be defined separately. There is a single modification file that may contain up to 8 definitions of each. The N- and C- modifications are defined like normal modifications, but are stored along with the sequence as modifications separate from the residue modification.

## 4 - Edit

In most cases mass values are calculated by GPMW *as needed*. (e.g. most sequences are calculated residue by residue every time it is displayed). This means that if you change a mass value, the change will be immediately reflected in the sequence and peptide windows, while a database search is static, and you will have to re-load the window for the changes to be active.

You should be careful when editing a sequence containing modifications as the modification may 'jump' to a different residue.

### Mass file

GPMW always needs a mass file in order to work. The default file loaded at startup is called AA\_MASS.MSS.



**Important:** If the AA\_MASS.MSS file is not found during start-up, or if errors are encountered, a default mass file is constructed internally which you are recommended to save as AA\_MASS.MSS.

Each mass file contains 32 entries. The first entry is for unknown residues, usually called 'X'. The next 20 residues are the standard 20 amino acid residues, while the last 11 residue are user-definable and can be given almost any name (be careful not to use punctuation marks, \$ or \* as single residue character).


For each residue you have to enter:

1-letter code

3-letter code

Name (<=10 characters)

Composition (the atoms have to be defined in the atomic masses table, see below and end of chapter). When you are in a field in the

'Composition' column and in edit mode , you can click on the 'Composition' button to open the Calculator (see Chapter 12.2) for easier input.

The 'Average mass' column is only for verification of the mass and cannot be edited.

The pKa and 'Charge' columns are optional and have to be used in conjunction. The pKa can be defined between 0.1 and 14, while the charge has to be either -1 or +1 (for acids and bases respectively – you cannot define modifications with two charges). If either value is zero, both values are ignored.

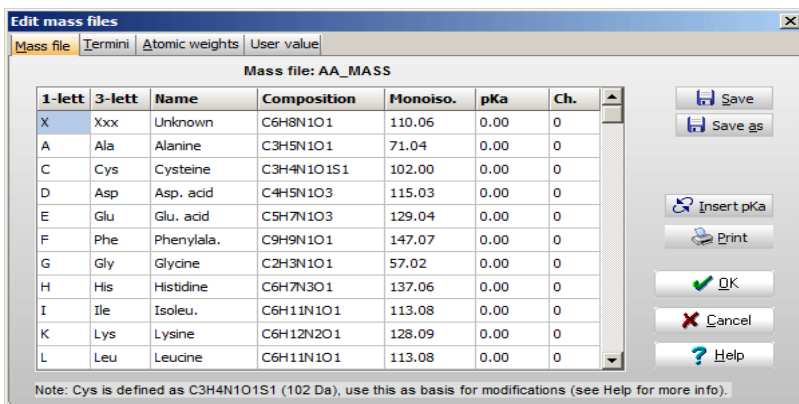
The 'extra' residues available in the mass table are best used for modified residues that are present in many copies or across several sequences. Modifying the individual residue (see Chapter 3.6, Amino acid modifications), best caters for single residue modifications. If you modify a residue type (e.g. carboxymethylate all cysteine residues) this is best carried out by changing the mass and full name of Cys (do not change the 1-letter code) and saving the mass file under a new name (e.g. pe\_cys for pyridylethyl cysteine). You can then modify cysteines just by selecting a new mass file in the toolbar of the main window.

## 4 - Edit

The **pKa** and **charge** are most commonly used for modified amino acid residues like carboxymethylated cysteine. When you want to use user-defined pKa values, you can transfer the pKa values from the built-in tables (appendix C.7) by pressing the '**Transfer pKa**' and selecting the appropriate table from the menu.



**Note:** When you want to use user-defined pKa values, you have to set the 'pl calculation' table in Setup to 'User defined' (Chapter 5.1).



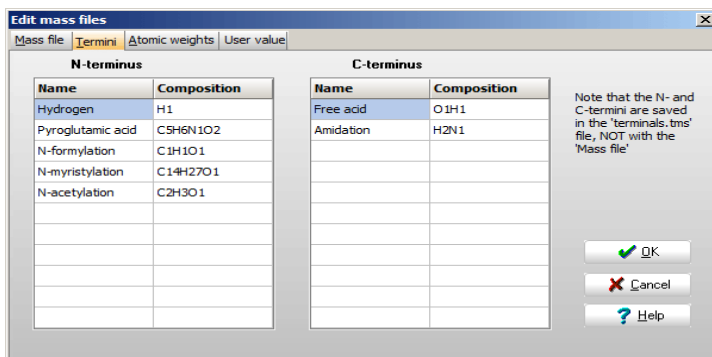
1-lett	3-lett	Name	Composition	Monoiso.	pKa	Ch.
X	Xxx	Unknown	C6H8N1O1	110.06	0.00	0
A	Ala	Alanine	C3H5N1O1	71.04	0.00	0
C	Cys	Cysteine	C3H4N1O1S1	102.00	0.00	0
D	Asp	Asp. acid	C4H5N1O3	115.03	0.00	0
E	Glu	Glu. acid	C5H7N1O3	129.04	0.00	0
F	Phe	Phenylala.	C9H9N1O1	147.07	0.00	0
G	Gly	Glycine	C2H3N1O1	57.02	0.00	0
H	His	Histidine	C6H7N3O1	137.06	0.00	0
I	Ile	Isoleu.	C6H11N1O1	113.08	0.00	0
K	Lys	Lysine	C6H12N2O1	128.09	0.00	0
L	Leu	Leucine	C6H11N1O1	113.08	0.00	0

Note: Cys is defined as C3H4N1O1S1 (102 Da), use this as basis for modifications (see Help for more info).

The '**Save**' button saves changes to the current mass file (shown at the bottom of the dialog while the '**Save as**' button enables you to save the whole list to a new mass file.

### N- and C-terminal

The tables for composition of the N-terminal and C-terminal are identical in setup, they differ only in the terminal they define and are presented on the tab called '**Termini**'.



N-terminus		C-terminus	
Name	Composition	Name	Composition
Hydrogen	H1	Free acid	O1H1
Pyroglutamic acid	C5H6N1O2	Amidation	H2N1
N-formylation	C1H1O1		
N-myristylation	C14H27O1		
N-acetylation	C2H3O1		

Note that the N- and C-termini are saved in the 'terminals.tms' file, NOT with the 'Mass file'

For each modification you enter a name for the modification and elemental composition (see Ch. 4.4).

## 4 - Edit



**Important:** In the N-terminal table the first entry has to be 'Hydrogen', H1, and for the C-terminal table the first entry has to be 'Free acid', O1H1. The first entry is automatically chosen whenever you load a new sequence, start editing a new sequence, perform cleavages etc.

All compositions are calculated relative to amino acid residues, not the free amino acid, see end of chapter for composition (formula) input.

Terminal modifications are saved in the system directory as a file called 'TERMINALS.TMS' when the program is closed.

### Atomic mass values

Atom	Name	Ave. mass	Mono mass
C	Carbon	12.010700	12.000000
H	Hydrogen	1.007940	1.007825
N	Nitrogen	14.006700	14.003074
O	Oxygen	15.999400	15.994915
P	Phosphor	30.973760	30.973763
S	Sulphur	32.066000	31.972072
Br	Bromide	79.904000	78.918390
Cl	Chloride	35.452700	34.968853
F	Fluoride	19.998400	18.998402
I	Iodide	126.904470	126.904660

Mass of a proton  
1.00727647

Default

OK

Cancel

Help

The 'Atomic mass table' is saved in the .ini file, NOT the mass file!

The atomic masses are the basis for all mass calculations carried out in GPMW. All atoms used in compositions in mass files, modifications etc. have to be defined in the atom mass table. The table can contain 32 atomic masses, and the values are saved in the GPMW.INI file and are always loaded upon startup. If the INI file is not found, default values are loaded.

The mass of a proton can be edited separately. This is the mass which is added to or subtracted from charged species, either a single proton ( $MH^+$ ) or multiple protons ( $MH^{2+}$ ,  $MH^{3+}$  etc).

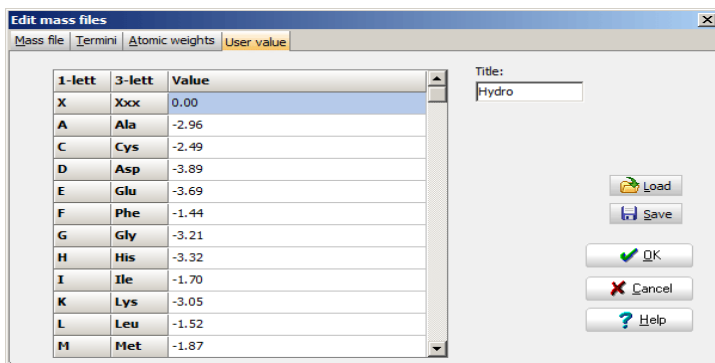
The '**D**' button resets the table to default values.

If you want to experiment with different values, remember to note down the previous values or make a copy of the GPMW.INI file.

### User value

The 'User value' tab enables you to enter a value for each residue, this will then be used for calculation in the peptide window (chapter 9.4), when the 'User' property has been chosen.

## 4 - Edit



The values that are entered can be saved in a .gpu file by using the 'Load' and 'Save' buttons. The format is a simple ini style text file, first line is [USERFILE], second is FILEID=GPMW USER AA VALUE FILE, third line is TITLE=name, then follow each residue in 1-letter code in the format A=-296 (i.e. real value multiplied by 100). This means that the resolution of the values is 0.01 and the highest number to be entered is 21000000.

The 'user' value is currently only used in the peptide window.

### Edit modification files

4.3

The modification files are used as a quick way to select a modification when modifying a residue (Chapter 3.6, 'Amino acid modifications'). The other function of modification files is when you perform a mass search of a protein. By including a modification file in the search, you can check whether any of the search masses could contain a modification as specified in the modification file.

The 'Edit modification database' dialog box works on the currently loaded modification file. If no file is currently loaded you have to load one through the **'Load'** button. Changes have to be saved to a file after modifications through the **'Save'** button. The file loaded will continue to be the 'active' modification file when the dialog box is closed. The modification files are saved in the 'System' subdirectory of the 'GPMW' directory (see Chapter 5.4).

Each modification file can contain up to 30 entries. Each entry consists of a name, a formula, and the residues for which the modification is valid. If no valid residues are specified, the modification is taken to be valid for all residues.

## 4 - Edit

**Edit modification database**

Modification file: adducts.MOD

Clear: Table Row

Name	Formula	Valid residues	OK	Charg	pKa	Term.
Oxygen	O1	M	<input checked="" type="checkbox"/>	0	0.00	-
Methylato	C1H2	DE	<input checked="" type="checkbox"/>	0	0.00	-
Phospho	H1O3P1	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"/>			
			<input type="checkbox"/>			

U Unimod

If no residues are entered in "Valid residues" column, then all residues are valid.

Only checked modifications ("OK" column) are active.

Buttons: Load, Save, Save as, Add sugar, Mass only, Selected modif. (Oxygen, 15.999 Da (av), 15.995 Da (mo)), OK, Cancel, Help

Whenever the focus changes to a new row, the name and mass of the current line will be shown to the right of the table. When editing a line, you have to move to another line and back again, before the mass of the line is recalculated (the program needs a complete formula in order to calculate the mass). Click on the **'Formula'** button Formula to open the **'Composition editor'**, for details check the following Chapters 4.4 and 12.2. The **'Add sugar'** button opens the **Carbohydrate editor** for easy input of carbohydrates (see below).

**Note:** An atom has to be defined in the atom table (see above Chapter 4.2) in order to be included in a formula.

The **'Mass only'** button is only active when a **'Formula'** field is selected and either empty or 'mass only'. It enables you to enter a mass instead of a formula, e.g. when you know the mass of a modification, but is uncertain of the actual elemental composition. Click on the **'Mass only'** button to edit the field. The values of a 'mass only' entry can be read in the green right-hand fields when the selection is in the relevant row.

The **'OK'** column enables the entry when checked. Normally, this column is only used when performing a mass search (Chapter 6.1), as having all entries valid in a large modification file can give a very large result list - modifications that might be known not to be relevant under the current circumstances.

The **'Charge'** and **'pKa'** columns work together to enter a charge for the modification. Both fields have to be entered in order for the charge to be active. pKa values can be between 0.1 and 14 while the charge can be either -1 or +1. Multicharged modifications are not supported.








**Note:** If the residue already has a potential charge (e.g. a defined pKa value) the results are bound to be inaccurate.

**Term.:** This field, if selected, limits the modification to either the N- or the C-terminus of the protein/peptide. This selection works in concert with the 'Valid residues' field, meaning that both conditions have to be fulfilled for the modification to be valid.

## Unimod

The Unimod is a public domain database, located at <http://www.unimod.org>. The database can be downloaded from the web in XML format and is included in the GPMW distribution. The file 'unimod.xml' has to be located in the 'system' directory (see Ch. 5.4) in order for GPMW to correctly locate it. As the database is regularly updated, it is recommended that you download and replace the 'unimod.xml' file on a regular basis to stay updated with new modifications (see below).

When you click on the 'Unimod' button (lower left corner of the dialog window), the database will be loaded, parsed and displayed below the modification file table:

 Unimod		If no residues are entered in "Valid residues" column, then all residues are valid.			Only checked modifications ("OK" column) are active.		 Help	
Double-click or select and press "Copy to list"					 Copy to list			
#	Long name	Short name	Mass	Formula	Unimod	ID	Residues	
82	Carboxylation	carboxyl	43.99	C1O2	C O(2)	299	W	
83	Gamma-carboxylation	Gamma-carboxyl	43.99	C1O2	C O(2)	38	ED	
84	S-Ethylcysteine from Serine	S-Eth	44.01	-O1+H4C2S1	H(4) C(2) O(-1) S	327	S	
85	Ethanolation of Cys	EtOH	44.03	H4C2O1	H(4) C(2) O	278	C	
86	Oxidation to nitro	Nitro	44.99	-H1+N1O2	H(-1) N O(2)	354	YW	
87	Acetate labeling reagent (N-term)	Acetyl_heavy	45.03		H(-1) H2(3) C(2)	56	>K	
88	Beta-methylthiolation	b-methylthiol	45.99	H2C1S1	H(2) C S	39	D	
89	Methyl methanethiosulfonate	MMTS	45.99	H2C1S1	H(2) C S	277	C	

You may copy entries from the Unimod table to the modification file either by double-clicking on the entry or by selecting the entry and pressing the 'Copy to list' button. The entries in the Unimod table are listed by mass.

**Long name:** The long name from the Unimod database.

**Short name:** Short name from Unimod database. This is the name copied to the modification file.

**Mass:** Mass value taken from the Unimod database, not calculated by GPMW.

**Formula:** Composition calculated by GPMW based on the Unimod formula in the next column. You should check that the composition is calculated correctly before copying. If the Unimod formula contains atoms not known by GPMW (e.g. O18, H2, C13), the formula cannot be copied and you have to construct it yourself.

**Unimod:** Composition of the modification as given by Unimod.

**ID:** Database entry number in Unimod.

**Residues:** Valid residues for the given modification. N- and C-terminal modifications are shown as '>' and '<' respectively. If the modification contains both amino acid and terminal 'accepted residues', the entry will be copied twice: one entry will show amino acid residues and the other

## 4 - Edit

entry will show the terminus. E.g. Methyl ester modification (Unimod ID 14) has the following valid residues: 'SED<T'. Double-clicking on the entry will make one GPMW entry showing 'SED<' under 'Valid residues' and '<' under 'Term.' and another entry will be empty in the 'Valid residues' column while having a 'C' in the 'Term.' column.

**Downloading the Unimod database:** You have to be connected to the Internet in order to perform this operation!

Click on the down-arrow in the left side of the 'Unimod' button. This displays a drop-down list, where you select 'Download Unimod'. This usually takes a minute or two, after which it is recommended that you close the edit window and reopens it in order to reload the Unimod file.

### Carbohydrate editor

Through the carbohydrate you can easily calculate the mass of various carbohydrate structures linked to polypeptides. It is built like a small wizard that leads you to the final structure in two or three steps.

The carbohydrate editor is selected either through the 'Simple modification' menu (right-click on residue in sequence window), the 'Insert modification' dialog box or the main menu **Search | Glycosylation | Glycosylation wizard**.

On the first page of the wizard is you can select the linkage type of the glycosylation as a type of **N-linked**, **O-linked** or **Other**.

The '**Base mass**' is a mass value (peptide, derivatization) that is included in all glycosylation mass values calculated. If you have a peptide highlighted in your sequence when selecting the glycosylation wizard through the main menu, the mass of the peptide will be entered in this field automatically.

## 4 - Edit

When you select the '**N-linked**' option, the wizard moves to the next step, where you have to select carbohydrate branching type. The **Next/Previous** buttons are now enabled so you can move forward and backward in the wizard. Selecting '**O-linked**' or '**Other**' leads you directly to the last page of the wizard where you can enter any kind of carbohydrate structure.

On the 'Branching type' option page, you select the branching structure, fucosylation and sialylation of the N-linked sugar. The fucose and sialic acids can also be entered on the last page.

Click 'Next' to go to the final page of the wizard. The left hand panel shows a schematic diagram of the sugar, the mass of the structure and, in the bottom panel, the sugar residues presently contained in the carbohydrate moiety. You can now add additional sugar residues by clicking on the respective buttons or you may alternatively select a residue from the drop-down box and press the '**Add**' button. Different sugars with the same mass are all grouped under the same button, i.e. galactose, glucose and mannose are all grouped under 'Hexose' as they are isobaric.

- Select branching type:
- ☐ Complex mono-antennary
  - ☒ Complex bi-antennary
  - ☐ Complex tri-antennary
  - ☐ Complex tetra-antennary
  - ☐ High mannose
  - ☐ Build-your-own (core unit only)
- ☐ Core fucose unit
- ☐ Sialylated end groups

**Glycosylation editor**

Add final sugar units:

Pentose  
Arabinose, xylose, ribose

Deoxyhex  
Fucose

Hexosamine  
Galactosamine, glucosamine

Hexose  
Galactose, glucose, mannose

HexNAc  
N-ac.gal.amine, N-ac.glu.amine

Sialic acid  
N-acetylneuraminic acid

Amino acid receptor residues  
N

Elemental composition  
C68 H112 N4 O50

Structure name:  
N-glycosyl

Mass:  
1785.634 Da (Av.)  
1784.634 Da (Mo.)  
+ base mass:  
1785.634 Da (Av.)  
1784.634 Da (Mo.)

Residues added  
Core unit  
Gal-GlcNAc  
Gal-GlcNAc  
Hexose

OK

Step back

Previous

Help

Cancel

Once you have added additional residues, the '**Step back**' button becomes enabled, enabling you to remove the last added residue. You can thus add and remove sugar moieties from your construct. You cannot 'subtract' residues beyond those added previous to this page of the wizard.

## 4 - Edit

Below the graphics on the left side is listed the mass values of the sugar moiety by itself and with the addition of the 'base mass'.

The **Amino acid receptor residues** are the residues that can accept the modification (corresponds to **Valid residues** in the modification file editor). These values can be edited later.

The **Elemental composition** is calculated based on the carbohydrate composition and is the structure only, without the acceptor amino acid residue. This field should not be edited.

Under **Structure name** you can enter a descriptive name for the glycosylation.

If you have chosen **O-linked** on the first page of the wizard, the graphic will show S/T with a single GlcNAc linked.

When choosing **Other** no graphic will be shown. Both the *O-linked button* and the *Other button* will skip the second page of the wizard.

When you press the '**OK**' button, the name, receptor residues and elemental composition will be transferred to the 'Edit modification file' editor if called from here. If called from the sequence window, the fields will be transferred directly to the respective fields in the sequence.

### Composition formulas

4.4

The composition in the mass file, the N- and C-terminals, and the modifications all follow the same rules.

The composition of the residue/modification is entered using the abbreviations specified in the atomic mass table (see above) followed by the number of atoms. If atoms are lost from the composition you put a minus sign '-' in front of the atoms lost (e.g. if you hydrolyze an amide you lose one nitrogen atom and two hydrogen atoms but gain an oxygen atom and a hydrogen atom, e.g. '-N1H1+O1'). Please note that negative numbers have to precede positive.

In several of the edit boxes you can activate the composition editor by clicking on the '**Formula**' button which appears in the right-hand part of the

edit field when in edit mode .

## 4 - Edit

Elemental composition H1N3P1

Carbon C	0	Fluoride F	0
Hydrogen H	1	Iodide I	0
Nitrogen N	3	Potassium K	0
Oxygen O	0	Lithium Li	0
Phosphor P	1	Sodium Na	0
Sulphur S	0	Selenium Se	0
Bromide Br	0	Zink Zn	0
Chloride Cl	0	Iron Fe	0

Composition:  Clear

Mass ave./mono. 74.00180 / 73.99081

OK Cancel Help

This opens the 'Elemental composition' dialog with the composition of the current selection. The composition can now be modified, either by directly entering the relevant numbers in the value boxes, or by using the up/down arrows next to the numbering boxes. Both positive and negative numbers can be entered. Negative numbers will only have meaning when editing post-translational modifications. Normally two columns of eight values are shown, but if more than 16 elements has been entered in the 'Atomic masses' page on the 'Edit mass file' page, three columns will be shown.

The **'Clear'** button resets all number fields to zero.

The **'Composition'** field shows the total composition and cannot be edited (but you may highlight and copy). The **'Mass ave./mono.'** field is for information only.

See also Chapter 12.2 'Composition calculator'.



## Setup

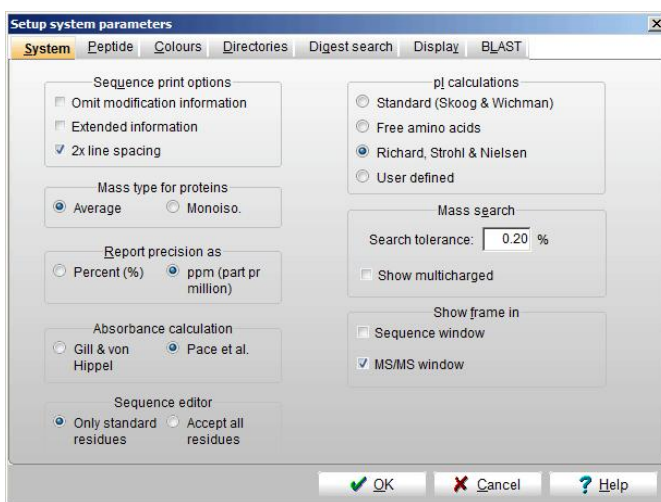
Setting up parameters for GPMW.

The **Setup | Setup system** dialog box contains most of the default data needed for GPMW. For setting up digest mass databases (**Setup | Make digest databases** please see Chapter 8, 'Database mass search'). All these data are saved in the GPMW.INI file (Appendix A).

Please browse this chapter carefully, as most of the layout and daily working of GPMW is dependent on the settings.

### Setup system parameters - System

5.1



#### Sequence print options:

These options will check/un-check the corresponding options in the '**Print sequence**' dialog box (Chapter 3.1).

**Omit modification info:** Information about sequence modifications and cross-links will be printed.

**Extended:** Print elemental composition and amino acid composition data. List cross-linked residues.

**2x line spacing:** Print sequence with double line spacing.

## 5 – Setup

### Mass type:

Defines the default display of masses to either average or monoisotopic (see Appendix C.1). This default value can easily be changed for each sequence window individually (Chapter 3.1) by clicking on the Av./Mo. button. Most other windows also enable the mass type to be changed on the fly. Please note that the mass type of the peptide window is set individually from the sequence window.

### Precision:

**% or ppm:** Determines whether the precision is reported as a percentage or as ppm (part per million). 0.01% equals 100 ppm. When working at high precision (e.g. better than 0.2%) you should use 'ppm' due to the better accuracy.

GPMAW will in general work with relative precision (% or ppm) and not absolute values (Da.) as most mass measuring instruments work in this way.

### Absorbance calculation:

Two scales for calculating concentrations based on absorbance at 280 nm are available:

Gill SC and von Hippel PH, Anal Biochem 182, 319-326 (1989)

Pace CN, Vajdos F, Fee L, Grimsley G, Gray T., Protein Sci. 1995, 4, 2411-23.

These values are used in the protein information window (section 3.8).

### Sequence editor

When editing a sequence in the sequence editor (section 4.1) you can set the editor to accept input (keyboard or paste) as either:

**Only standard residues:** Only the 1-letter residues defined in the currently loaded mass file will be accepted.

**Accept all residues:** All characters will be accepted as input whether or not they have been defined in the current mass file.

### pl calculations:

The advanced page allows you to choose between different pKa tables for the calculation of the pl of peptides and proteins. The options are:

1. B. Skoog & A. Wichmann, Trends in Anal. Chem. 3, 82-83 (1986)
2. Free amino acids
3. Rickard, Strohl & Nielsen, Anal. Biochem, **197**, 197-207, (1991)
4. User defined.

In all cases the algorithm of Skoog and Wichmann is used.

The pl calculations are used in a number of different places. When only a single value is reported, the value will be the one based on the option chosen here. Examples can be: Peptide window, peptide info (Chapter 9.4), Charge vs. pH graph (Chapter 11.8), DigestAlyzer (Chapter 11.9) and the Simulated 2D gel (Chapter 12.5). In a few cases, e.g. the sequence information dialog (Chapter 3.8), the three first pl values are displayed.



## 5 – Setup

### Mass search:

**Search tolerance:** Default tolerance for mass searches, digest mass searches, etc. The value can be changed before each search.

**Show multi-charged:** Displays multi-charged species as defaults while showing the results of the search for mass (Chapter 6.1).

### Show frames:

The frames are resizable parts of a window that shows information in addition to what is present in the main window. The 'frames' are available for the following windows:

1. Sequence window (Chapter 2). Contains information on modified residues, modified terminals, cross-links and pI.
2. MS/MS window (Chapter 10.1). Displays a sorted list of all masses displayed in the main window. The two displays are linked, so that clicking on a value in the frame will highlight the corresponding value in the main window.

## Peptide parameters

5.2

**Setup system parameters**

System **Peptide** Colours Directories Digest search Display BLAST

**Precision**  
☒ 2 decimals  
☐ 4 decimals  
☐ 6 decimals

**Residue display**  
☒ 1-letter code  
☐ 3-letter code

**Copy table to clipboard**  
☐ Copy table as text  
☒ Copy tab delimited

**Display mass**  
☐ Average  
☒ Monoisotopic

**Sort list by**  
Mass

**Table sequence**  
☐ Limited sequence  
☒ Full sequence

Calculate charge at pH: 8

Mass search switchover  
Mono/Ave 6000

Low mass cutoff: 610 Da

**Column layout**

Num	MH+	pI	HPLC	M2H-	Ch	User	Sequence

**Alternate column layout**

Num	M-H	M2H-	M3H-	M4H-	M5H-	Ch	Sequen

OK Cancel Help

The peptide parameters determine the initial display/print/copy parameters for the peptide window (the result of protein cleavage, see Chapter 9.4). Most parameters, except the copy parameters, can be changed on the fly.

### (Reported) Precision:

Determines whether peptide masses are reported with 2 or 4 decimals. The internal precision of the mass calculations as carried out by the program is not changed. The calculated precision is dependent on the values entered in the mass files (Chapter 4.2, default is 5 decimals for average masses and 6 decimals for monoisotopic masses).

## 5 – Setup

### Display mass:

Shows masses as either average or monoisotopic masses on the screen.



**Note:** The peptide window can have a different default mass setting than the other windows. E.g. the sequence window will usually have an 'average mass' setting while the peptide window will be in monoisotopic mass mode due to the higher resolution of peptides.

### Residues:

Displays amino acid residues in the peptide list using either 1- or 3-letter code. Can be changed dynamically in each peptide window.

### Sort by:

Sorts the peptide list by number (position in the protein), mass, HPLC index, Bull & Breese index [H.B. Bull & K. Breese, Arch. Biochem. Biophys., 161, 665 (1974)], charge (with a secondary sort by number) or sequence (alphabetical sorting based on 1-letter code).

### Copy table to clipboard:

When the peptide table is copied to the clipboard this option will determine whether delimiters are space characters (text) or tab characters. Use the **text** form when you copy to a report (e.g. Word), and **tab delimited** when you copy to a spreadsheet (e.g. Excel).

### Table sequence:

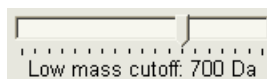
When you copy the peptide table to clipboard this option selects whether the peptide amino acid sequence is copied in total ('Full sequence:' EECSVPVCGQDR) or the central part of the sequence is replaced by ... ('Limited sequence:' EEC...QDR).


### Calculate charge at pH:

Determines at what pH will the charge of the peptides in the peptide list will be calculated. This setting is also used in various other places like the protein window frame, peptide info etc.

### Low mass cutoff:

The 'Low mass cutoff' determines the mass value below which peptides are hidden in the peptide list. The main use for this option is in MALDI mass analysis where low mass values are not determined



in the mass spectrum. The low mass filter is set with the low button  in the peptide window. The low mass cutoff can be set between 100 and 1000 Da by using the slider.

### Column layout:

As the information needed for different experiments varies, it is possible to specify a wide variety of parameters to be reported for each peptide. To make the setup more flexible, two different layouts are supported: **primary layout** (which is shown when the peptide window opens) and **alternate**

## 5 – Setup

**layout** (you can switch between the two modes by pressing the **'Alt'** button in the toolbar – see Chapter 9.4).

The layouts are shown in the setup dialog as two white lines displaying the selected column headers. Each line corresponds to the header of the corresponding peptide list. The column settings for each layout can be edited through the **'Setup'** button situated to the right of each line. Pressing either button opens a new dialog box that enables you to specify the settings.

The column layout is edited through a number of drop-down selection boxes. The left-most column is shown at the top of the dialog box. The first and the last columns are fixed as peptide number and peptide sequence, respectively. Additionally, up to six columns can be specified from the six drop-down boxes labeled 2 - 7. If you do not want to display a particular column, you set it to '(none)'.

The parameters that are available for display are:

**Number** (always present in the first column) – the number of the peptide in the peptide list. Overlapping peptides (e.g. missed cleavages) are numbered from the N-terminus, first all the peptides with 1 missed cleavages, then with 2 missed cleavages etc.

**Sequence** (always present in the last column) – sequence in either 1- or 3-letter code depending on the status of the **1/3** button.

**MH+** ( $MH^+$ ) to **MH8+** ( $MH^{8+}$ ) – singly charged ion to the ion with eight charges. Calculated as the neutral mass with the addition of the appropriate number of protons.

**MH-** ( $MH^-$ ) to **MH2-** ( $MH^{2-}$ ) – the singly and doubly negatively charged ions. Calculated as the neutral mass minus the stated number of protons.

**M** (neutral mass),

**From-to**, - first and last position of the peptide in the sequence.

**HPLC index** (reversed phase retention index),

**Ch.** – charge at the selected pH, see above for the main peptide parameters. – the actual value calculated depends on which pI parameter list has been chosen in the System setup (Ch. 5.1 above). See also App. C7.

**B&B** (Bull & Breese index),

**Add mass** – add a fixed mass to each peptide. The value is entered in the 'Add mass value' edit box. The value can be any real number, positive or negative.

**Alt. MS.** Here you have to specify the alternative mass file name (this list is identical to the drop-down list in the main toolbar, Ch. 4.2). In addition you can specify the charge state, the composition of the N-terminus and the C-terminus. The compositions of the termini are taken from the N- and C-terminal mass list (Ch. 4.2).

## 5 – Setup

**Av/Mo** This column shows the singly charged ion ( $MH^+$ ), but in the opposite mode of the **Av./Mo.** button, i.e. if monoisotopic mass has been chosen as the display of choice, this column will show the average mass.



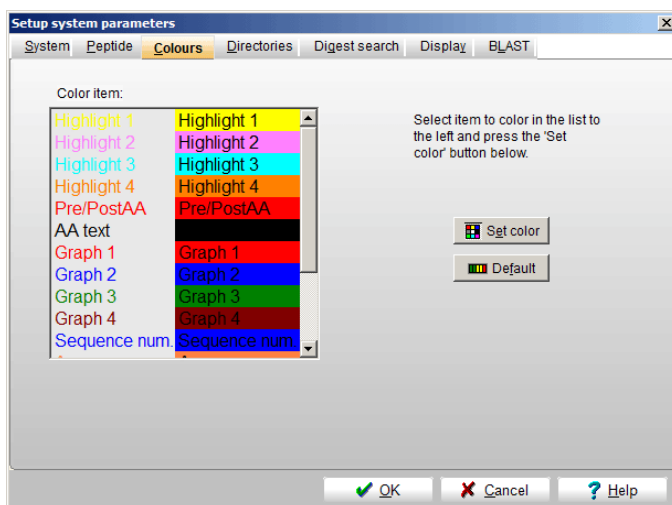
**Note:** The actual mass displayed will be either the monoisotopic mass or the average mass depending on the setting of the **Av./Mo.** button.

For a discussion of the individual parameters please see 'Protein cleavage', Chapter 9.1.

### System colors

### 5.3

The system colors determine how sequences, graphs etc. are displayed on the screen. By selecting appropriate colors, you can make the reading of information faster and safer. As computer monitors vary in clarity and color fidelity, you are encouraged to experiment with various color combinations.



The left hand list shows you the currently defined colors in GPMW. The list is divided into two columns, where the first column shows you text on white (actually light gray) background and the right hand column black text on colored background. The different colors are used either in one or the other mode, so the selection of color should reflect this.

You edit a color by selecting it in the table and either press the **'Set color'** button or double-click on the selection.

**Highlight 1-3:** These colors are used as background when highlighting sequence residues. You should use light but bright colors (e.g. check whether the black characters are easily read in the right-hand column).

**Highlight 4:** This is the color of modified residues and it is not a background color, so you should choose the color based on the left-hand column.

## 5 – Setup

**Pre/PostAA:** The color of the residues before and after the identified sequence in the mass search window (Chapter 6.1).

**AA text:** The sequence in the sequence window – usually black.

**Graph1-4:** The color of the four graphs that can be displayed in the various graphs. As the lines can often be thin and difficult to discern from the background, you should use bright and dark colors (e.g. check in the left hand column).

**Sequence num.:** The color of the subscript numbers in the sequence window.

**Aux:** This color is not used at present.

**Dot1-4:** The color of the dots in the ‘frame’ of the mass search window (Chapter 6.1).



**Note:** Several windows are able to print in color (e.g. sequence window, peptide window, and several graphs), and when printing to a monochrome printer (e.g. a laser printer) the different colors will be printed as various shades of gray. By experimenting with different colors, you will most likely be able to print in useful shades of gray. If you are using both a color printer and a monochrome printer you may have to change the color table when changing printer – this is most easily done by setting up different users (see Chapter 5.7).

The **‘Default’** button reverses all colors to the default colors for GPMW.

You can edit a color either by selecting the relevant line and click on **‘Set color’** or by double-clicking on the colored line.

In either case you will get the standard Windows ‘Color’ dialog box.

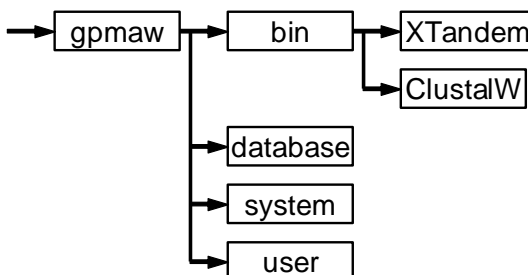
The current color will be selected (dotted line around the color). You can now select a new color from the ones displayed or press the ‘Define Custom Colors’ button to select a new color from the advanced dialog box layout.

Click on **‘OK’** to select the color.



**System directories****5.4**

During installation, the various components of GPMaw are installed into the following directory structure:



The main directory is set to C:\GPMaw, but can be changed by the user during installation. Although the program can be installed to any directory, it is recommended that you use the default c:\gpmaw\ as future updates will be much easier to perform.

Below this directory, four directories are created:

**\BIN** contains the main gpmaw3.exe executable program file, the gpmaw3.ini file (contains the initialization data between sessions), the validation file gpmaw3.chk, and the help file gpmaw3.hlp. Additional helper programs like DBIndex (Chapter 12.4) are also installed here.

Two directories can be created beneath the BIN directory:

**\XTandem** is created automatically, and the XTandem! executive and auxiliary files are placed here. These are automatically registered by GPMaw.

**\ClustalW** is a directory created by the user if you want to perform ClustalW multiple alignments. Due to licensing reasons, these files are not included with the GPMaw distribution, but can be downloaded by the end-user. Please see chapter 7.3 for more details.

**\DATABASE** default place for digest mass databases. Each digest mass database contains three files; a data file (.DAT), a name file (.NAM), and an information file (.INF). See Chapter 8.2 for a description of how to create digest mass databases. If your hard disk is partitioned into several drives or you work across a network, you are likely to place the protein elsewhere.

**\SYSTEM** contains various common files that are shared between different users/sessions: modification files (.MOD), mass files (.MSS), and highlight profiles (.HPR). The system file can be changed in the **Setup|Setup system|Directories**, but unless you have compelling reasons you should leave it in the default state.

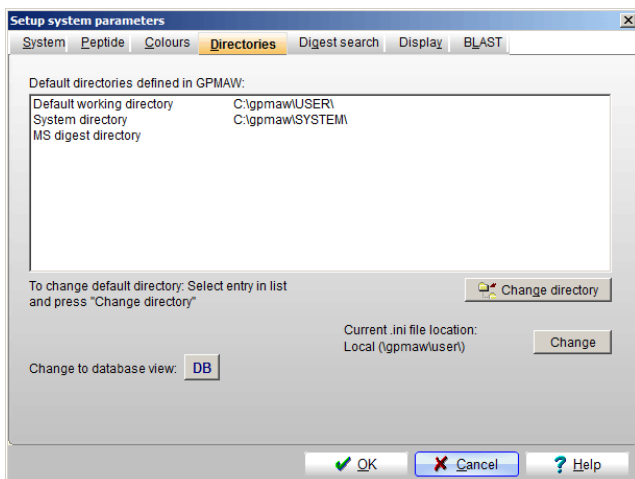
**\USER** contains the files that are individual for a session: GPMaw sequence databases (.SEQ), peptide mass files (.PEP), peak mass lists (.PKS), and peptide mass search result files (.PMS). You can create several different user directories and change between them in the **Setup|Setup system|**

## 5 – Setup

**Directories.** If you make different directories for your projects/users you should set up GPMaw for different users, see below in 5.7.

On the '**Directories**' page of the system setup you can specify a different working directory, different digest mass search directory ('MS digest directory'), and a different system directory. As the 'Database mass search' databases can be huge (> 20 MB), it can be advantageous to place them on a central server or another large shared disk in a network.

**Hint:** If you use multiple databases it can be advantageous to put each of them in a separate directory below a main directory (e.g. have the directory c:\gpmaw\database\SwissProt and c:\gpmaw\database\IPIhuman) and then let the 'MSdigest directory' point to c:\gpmaw\database. In this way each database will just be one additional click away.



If you want to change a directory you can either double-click on the corresponding line in the list or select the line and press the '**Change directory**'. Then you navigate to the correct directory.

### Protein databases referenced in GPMaw

Protein database	D:\Database\SwissProt\SWISS.seq
Date	12-04-2004 11:08:30
Text search indexed	yes - 12-04-2004
BLAST indexed	yes - 12-04-2004
Protein database	D:\Database\IPImouse\IPImouse.seq
Date	03-04-2004 14:19:32
Text search indexed	yes - 03-04-2004
BLAST indexed	no
Protein database	D:\Database\IPIhuman\IPIhuman.seq
Date	03-04-2004 14:20:28
Text search indexed	yes - 03-04-2004
BLAST indexed	no

Update DB list

## 5 – Setup

Pressing the 'Database view' will give you statistics on all the databases referenced by GPMaw (FastA search, digest search and BLAST).

**Note:** You can set up GPMaw for different users or projects by setting 'Users' (see 5.7) or using multiple icons on the desktop with different in-line parameters (see Appendix D).

### Digest mass search parameters

5.5

From the '**Digest src.**' page you control the initial settings of the search parameters for the digest mass search (see Chapter 10).

#### Search limits parameters:

The search limit parameters are also discussed in Chapter 8. The search limit parameters are set by pressing the **S Set Limits** button. The limits can also be set on-the-fly in the mass input dialog for the peptide mass search.

**Mass range:** The smallest and largest protein mass to search for mass hits. Usually, you know the approximate mass of the protein in question, but you should enter a wide margin in order to compensate for fragments, pre- and pro-proteins in the database. You should normally have a lower limit of 10 kDa (to exclude a large number of very small fragments) and an upper limit of 100 kDa (to exclude a small number of very large proteins that tend to give false positives).

**Precision:** The mass precision of the input search masses when searching the database. Can be listed either as % or ppm as defined on the System page (Chapter 5.1).

**Minimum prec.:** If you are unable to determine the low masses with absolute precision you enter the minimum attainable precision here, otherwise you enter 0.0.



## 5 – Setup

**Monomass:** When using high-resolution mass spectrometers, you are able to obtain monoisotopic mass values at low masses (e.g. below  $m/z$  3000). As these values are usually more precise than average data, it will be advantageous to use these. In order also to use high mass values, you set the monoisotopic crossover mass to the value below which you determine monoisotopic masses. As the digest database contains both sets of values, GPMaw can search both types simultaneously.

**Overlaps:** Determines how many un-cleaved potential cleavage sites should be allowed in the target peptides. E.g. a tryptic peptide like LIPKTGHNEDRKSVR contains two potential cleavage sites and will have an overlap value of 2. This value is also called missed cleavages.

**Min. hits:** The minimum number of peptide masses that have to fit in order to be entered into the final score list.

**Mass type:** Default ion type for the mass input table, determined by your mass spectrum.



**Note:** The program will do the fastest search when no overlaps have been specified. This is partly because each overlap adds a search overhead and partly because a slightly different algorithm is used.

### Scoring parameters

The scoring parameters determine how hits are evaluated. A hit is defined as a database value that falls within the search window defined by a search mass. You should feel free to experiment with various values, as most likely there is no universal magic setup for the search parameters.

**Overlaps:** The score for a given number of overlaps. Non-overlapping peptides are given the highest values. Peptides containing one or two overlaps are also common and should be given a high score. Overlaps of four and more are quite rare (at least they are rarely observed).

**Score type:** At present three different scoring types are supported: *Linear*, scores are not adjusted. *Score/NumPep*, the score is divided by the number of peptides present in the database protein. *Score/Square root*, the score is divided by the square root of the number of peptides of the protein found in the database. The last two scoring types compensate for the fortuitous hits of large proteins. *Score/NumPep* tends to overcompensate while the *Score/Square root* usually compensates satisfactorily for 'normal' proteins in the 20-150 kDa mass range.

**Precis./2 and Precis./4:** If the hit is closer than half/quarter of the given precision for the search peptide, an additional score is added to the total.

**Sequence and Compos.:** The score given for a match of a sequence or an amino acid composition.

**Optimization:** The 'hit' list from the peptide mass search can be re-searched using optimized parameters. You can here select the optimization to include:

- 1) an increased number of overlaps (missed cleavages)
- 2) a linear fit on the hits from the first search to be performed and used as a modified 'calibration' for a second search.

## 5 – Setup

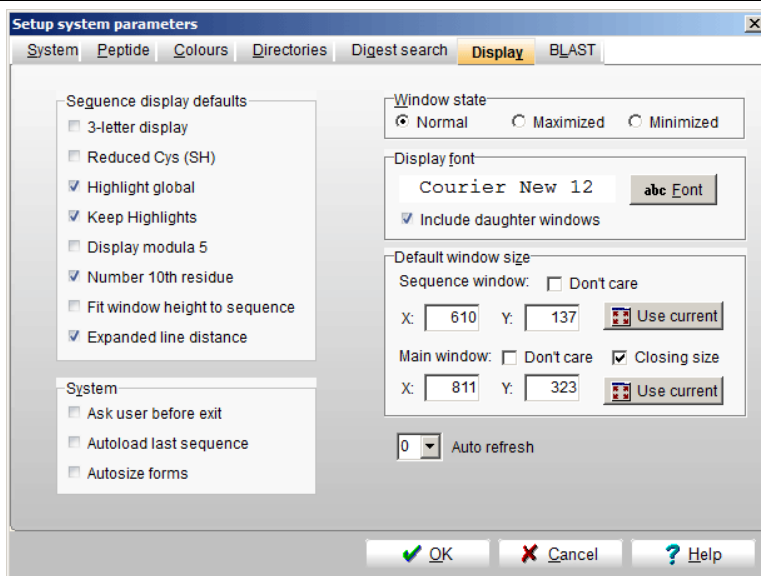
- 3) a compensation of the score by using tryptic PMS (peptide mass search) rules: if the basic residue is terminal or next to an acidic residue it is not counted as an overlap (and will therefore result in a higher score). You will get an additional score if the peptide starts with Gln and a mass is found at  $-15$  (corresponding to pyro-Glu). If the peptide contains Met and a  $+16$  mass is present an additional score will be added for oxidized Met.

**Autoload correct mass file:** When the program performs the second pass search, all mass calculations are redone. In order for this to function correctly the right mass file has to be loaded. If this option is checked, the file will be loaded automatically, otherwise you will be asked.

**Show pI in results:** If the original sequence database is available on-line, the program will calculate the pI of each result hit when presenting the result table.

## Display

5.6



### Sequence display defaults:

**'3-letter display':** If checked, the sequence window will show amino acid residues in 3-letter code. If not checked residues will be shown in 1-letter code.

**'Reduced Cys (SH)':** If checked, cross-links are not displayed or calculated. If unchecked, cross-links are displayed as red lines (Cross-links, Chapter 3.5). Cys residues are calculated as mass 103 Da when reduced (SH) and as 102 Da when oxidized (SS).

## 5 – Setup

**'Highlight global':** If checked, all sequence windows opened on the desktop will be highlighted whenever the highlight command is executed (Chapter 3, Highlight residues).

**'Keep highlight':** If checked the highlight dialog box will remember settings between executions. The two options **'Highlight global'** and **'Keep highlight'** can be changed at run-time (Chapter 3.2).

**'Display modula 5':** Sequence windows will display protein sequences only in multiples of 5, e.g. 55 residues pr. line, not 56 or 54 residues. Although most useful when displaying 1-letter code it also works for 3- letter code.

**Number 10<sup>th</sup> residue:** When checked every 10<sup>th</sup> residue in the sequence window will be labeled with a subscript number when displaying 3-letter code. The color of the number depends on the color setup (Chapter 5.3). In 1-letter mode every 10<sup>th</sup> residue will have a small vertical line.

**Fit window height to seq.:** When checked, all newly opened sequence windows will have a height that just fits the displayed sequence. See also 'Default window size' below.

**Expanded line distance:** In the sequence window you can choose to get a little extra distance between sequence lines. This is particularly useful when displaying multiple Cys cross-links.

### System:

**Ask user before exit:** If checked, GPMW will pop up a dialog box before closing asking whether you really want to close the program.

**Autoload last sequence:** GPMW will try to load the most recently accessed sequence automatically when the program is started next time.

**Autosize forms:** When the system font is changed, some dialog boxes also change in order to accommodate the new font size. Sometimes GPMW may have problems resizing correctly. If you experience this problem try to check this box to force the program to recalculate the size of dialog boxes.

**Auto refresh:** GPMW suffers from a problem with an as yet undetermined solution: in some cases, highlight information will be lost upon opening a new window. The problem is not reproducible, but by setting the 'Auto refresh' value > 0, all sequence and daughter windows will be refreshed whenever a new window is opened. This will in most cases counterbalance the above-mentioned defect.

### Window state:

This option enables you to determine the initial display state of GPMW:

- **Normal:** The program will open in a window that will take up approximately 1/3 of the screen.
- **Maximized:** The program will be displayed covering the whole screen area.
- **Minimized:** The program will be minimized to the task bar. This feature is most useful if you add GPMW to the Windows 'Startup' folder in order to automatically start GPMW whenever you start your computer.

## 5 – Setup

### Display font:

Click on the **'Font'** button to select any monospaced font installed on the system for the sequence window. You can also select a different font size. If you check the 'Include daughter window' box, the selected font will also be used for display in the peptide window (Chapter 9.4) and the mass search window (Chapter 6.1).

The default font is Courier New in size 10 point.

### Default window size:

This option, when enabled, defines the initial size of the main GPMaw program window and the initial size of the sequence window.

If the **'don't care'** box is checked, the values entered will have no effect.

Pressing the **'Current'** button will read the current size of the program window / the size of the topmost sequence window and put the values into the relevant boxes (X – width, y – height). The values can also be edited manually.

If you have checked the **'Fit height to seq.'** box above, the height parameter entered here will be ignored.

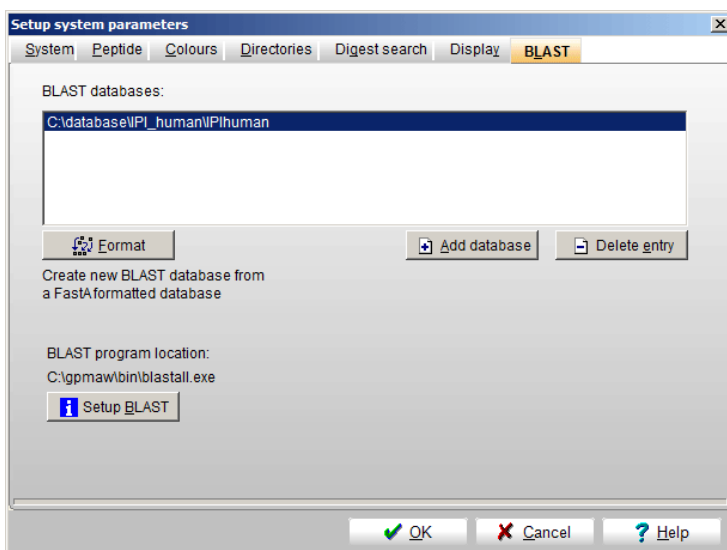
If you check the **'Closing size'** box GPMaw will open with the same size as the program had when it last closed.



**Note:** It is possible to enter values larger than the current window size. This will result in parts of the program / sequence window being inaccessible. In this case you should reopen the setup box and enter new values.

## BLAST

## 5.7



## 5 – Setup

The BLAST setup page works in concert with the 'Local BLAST homology search' see section 7.2.

The BLAST homology search uses the NCBI BLAST program called 'blastall.exe'. This file will in a normal GPMaw installation be installed in the C:\gpmaw\bin\ directory. Along with this file you will need the following files: formatdb.exe, blosum45, blosum62, blosum80, pam30, pam70 and seqcode.val.



**Note:** If you are unable to locate these files, you can download them from the NCBI FTP site as a compressed self-extractable file (<ftp://ncbi.nlm.nih.gov/blast/executables/blastz.exe>).

When you have decompressed the blastz.exe file, you can copy the files mentioned above to the \gpmaw\bin\ directory. The remaining files in the download are not used at present.

In order to use the local BLAST you need to tell GPMaw the location of the 'blastall.exe' file and you need a database in BLAST format.



Install BLAST

Pressing the 'Install BLAST' button will present you with a 'File Open' dialog box, which you use to locate the 'blastall.exe' file. By default this will be located to c:\gpmaw\bin\, but you can place it wherever you like. The other files mentioned above have to be placed in the same directory in order for GPMaw to locate them. If the 'blastall.exe' file is in the \bin\ directory GPMaw will usually locate it automatically.



Format

In order to run a homology search, you need a protein database to compare with. These can be generated from any FastA formatted protein database, please see Appendix B for how to obtain a database. If you have obtained your copy of GPMaw on a CD-ROM, you will usually find two databases (Swiss-Prot and EMBL-nr) on the disk ready for use. The databases can be the same as the ones used for retrieving sequences (Chapter 2.6) and peptide digest database search (Chapter 8).

When you press the 'Format' button, you will be asked to open the FastA formatted database to be converted. The actual formatting is carried out by an external program 'formatdb.exe' that is called by GPMaw. Do not close the black DOS box that opens when this function is called! It will close automatically when the database formatting is finished.

When finished with the conversion, GPMaw will ask whether you want it added to the list of BLAST databases. When you have done so, the database will be available from the 'Local BLAST' option (Chapter 7.2).

If you have a ready made BLAST formatted database, you can add it to the

list by pressing the



Add database

button. You will be asked to locate the '.psq' file of the database set.

You can remove entries from the list by pressing the



Delete entry

button. **Note:** This function will only remove the reference to the database, not the actual database.

## 5 – Setup



**Hint:** A BLAST database consists of three files with the extensions .phr, .pin and .psq. The total space required by the three files is slightly larger than the original FastA formatted database. If your only purpose is to perform BLAST searches (e.g. no sequence retrieval or peptide mass searches) you can delete the FastA database after the generation of the BLAST database.

### Users

5.8

The concept of 'Users' can be exploited in two ways:

- I. Multiple users can use the same installed version of GPMaw but have different preferences and directories to store individual data.
- II. A single user can have different projects localized to different directories. Furthermore, each project (or user) can have different preferences, very useful if you work with different mass instruments having different resolutions.

Selecting **Setup | User | New user** and entering a name of not more than eight characters creates a new user. The current .INI file is then saved with this name. Any preferences you have made or will make before closing the program will be saved to the new user.

Selecting an already existing user will load the preferences in the new ini file. Changed preferences in the current user profile will not be saved before loading the new profile.

You remove a user by selecting **Setup | User | Remove user** and entering the name of an existing user when asked in the dialog box.

The Default option loads the default gpmaw.ini file.

The currently loaded use profile is displayed in the title bar and after the 'User' option in the 'Setup' menu.

See also Appendix D on how to set up GPMaw for different users to start directly from a given shortcut.

### Setup proxy

5.9

Many companies do not allow direct access to the Internet, but directs all traffic through a proxy server. In order to access this, you need to tell GPMaw the specific settings.

In **Setup | Setup proxy** you can specify the relevant parameters. Remember to check the 'Use proxy settings' to enable the Internet connection.

You only need to enter the parameters once in a setting, but GPMaw will not remember your password between settings, but will ask upon start of a new session.

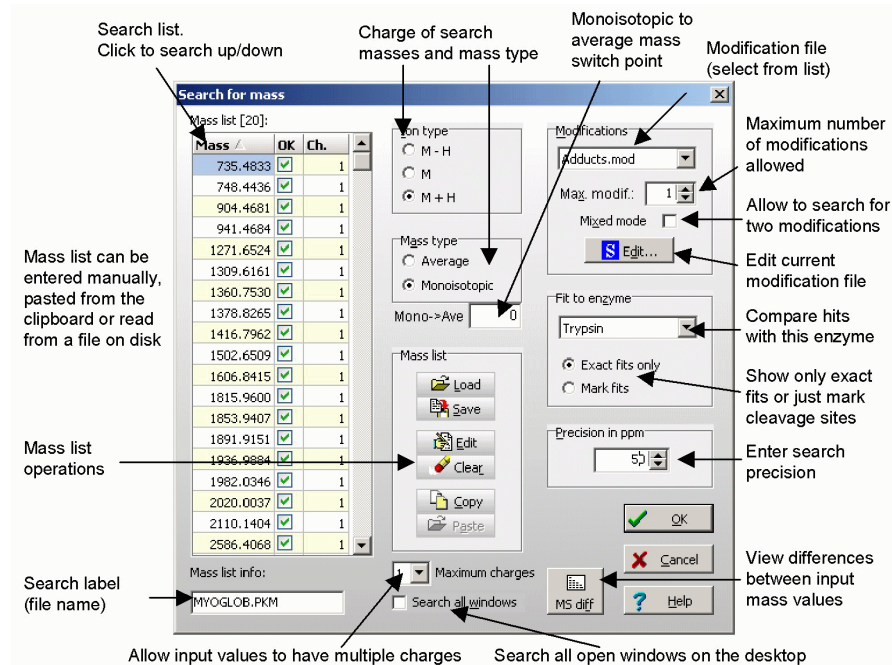
## Mass search.

How to determine whether a given mass is present in the protein.

The main concern in this chapter is to search a protein for a given mass (**Search|Mass search. .**). However, at the end three small utilities are presented that may help you identify mass differences (**Search|Mass difference. .**) as well as a function to help you identify cross-linked peptides.

### Search for masses

6.1



The search for mass function enables you to search a protein sequence for a list of masses. The result shows all peptides in the protein that fit within the given mass window. You can only perform one search at a time. If you perform new search on a sequence window that already has a mass search result window, this window will close and the new mass search results will be displayed instead.

## 6- Mass search / FastA files

**Mass list:** The list of masses can be entered manually, read from a disk file or pasted from the clipboard (Ctrl+V). The disk file can be in GPMW peptide mass format (Appendix A) or a peak file saved from a number of mass spectrometry acquisition software. The exact peak formats supported will continue to increase. If your current software is not supported please contact Lighthouse data for information.

Mass values can be entered with as many decimals as needed (but only four decimals will be displayed), but you should consider of the working mass precision that has been set. When a mass has been entered or read from disk it will be enabled (checked in the **'OK'** column). You can disable masses by un-checking the checkbox in the right-hand column. You can also

Mass list [17]:

Mass	OK	Ch.
732.47	<input checked="" type="checkbox"/>	1
748.38	<input checked="" type="checkbox"/>	1
993.40	<input checked="" type="checkbox"/>	1
1262.57	<input checked="" type="checkbox"/>	1
1271.63	<input checked="" type="checkbox"/>	1

enable/disable, delete or screen the mass table quickly by selecting the **'Edit'** button, see below. The last column in the table (labeled **'Ch'**) shows the charge state of each mass. The default charge state is set in the 'Ion type' box (top middle of the dialog. Changing the 'Ion type' resets all charge states in the table. To change a charge state, select field, enter edit mode (click twice, press F2 or enter a number) and enter value or use the up/down arrows in the active edit field.



**Note:** The maximum number of mass values in the list is 300.

Right-clicking on the mass list opens a local menu with the following options:

**Load table**, **Save table**, **Copy to clipboard** (Ctrl+C), **Paste from clipboard** (Ctrl+V), **OK all**, **Invert OK**, **Insert line**, **Delete line**, **Clear table**. 'Load table', 'Save table' and 'Clear table' duplicate the corresponding buttons of the **'Mass table'** buttons. **'Copy to'** and **'Paste from'** are standard clipboard routines that will copy a mass list either to or from the clipboard. **'OK all'** enables all mass values. **'Invert OK'** will disable all enabled values and enable all disabled values, thus enabling you to carry out complementary searches on a mass list. **'Insert line'** and **'Delete line'** will insert an empty line or delete the current line, respectively.

Load table	
Save table	
Copy To clipboard	Ctrl+C
Paste From clipboard	Ctrl+V
OK all	
Invert OK	
Insert line	
Delete line	
Clear table	

You can sort the list by pressing the column title bar.

### Options

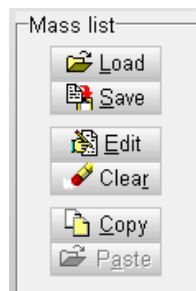
**Ion type:** Select the ion type that fits with the input. M+H subtract and M-H adds the mass of a proton before carrying out the search, while M uses the mass input list as given.



## 6- Mass search / FastA files

**Mass type:** Select average or monoisotopic as fits your data (Appendix C). You cannot mix monoisotopic and average mass values.

**Mass list table:** **'Load..'** loads and **'Save..'** saves the mass list to a disk file. **'Edit..'** opens the Enable/disable mass dialog box, see below. **'Clear'** clears the mass list. If you load a mass table from disk, the file name will be entered in the **'Mass list info'** edit line. **'Copy'** and **'Paste'** will write/read a mass list from the clipboard.



**Precision:** The mass window calculated around each mass value. The default value is defined in Setup (Chapter 5.1), as is precision unit as ppm (parts per million) or %. The up/down buttons shift the precision in 10/50 ppm units when in ppm mode and in 0.01% / 0.001% units when in % mode.


**Multicharged:** If checked, multicharged ions ( $M_2H^{2+}$ ,  $M_3H^{3+}$  etc.) will also be considered during the search (up to +5).

**Search all windows:** When checked, all sequence windows that are open on the desktop will be searched for the given mass values. The results of this search open in a window different from the usual results/report window. Please see end of section 6.1 for more details.

**Mass list info:** Here you can enter any text you want printed along with your mass search. The mass list info will also be shown in the title of the search result window. If the mass list has been read from disk, the edit field will show the data filename.

**Modifications:** The drop-down list box enables you to select a modification file to add to your search. If a modification file is added to the search, all masses in the list will first be used for a search as they are listed, and then the mass of each enabled modification will be added and the search will be repeated. Only peptides containing residues that are specified with the given modification will be considered and only up to the value specified in the **'Max. modif.'** field. Selecting a modification will automatically set the **'Max. modif.'** field to 1, if it is not already set at a higher value. The 'Modifications' drop-down list will show all modification files present in the 'system' directory.

When the **'Mixed mode'** box is checked, GPMW will search for one additional modification for each modification added to the search peptide. This means that when searching for 2 modifications (i.e. phosphorylation) then for both 1 phosphorylation and for 2 phosphorylations there will be an additional search for one modification from the selected modification file.

Selecting the  button opens the **'Edit modification file'** dialog box with the currently selected file (see Chapter 4.3) ready for editing. Remember to save the changed modification file in order for the changes to take effect.

**Fit to enzyme:** Selecting an enzyme in the **'Fit to enzyme'** drop-down box will compare the result list to the specificity of the given enzyme (as specified in the enzyme cleavage list, see Chapter 9.1). If the **'Exact fits only'** is

## 6- Mass search / FastA files


selected, only peptides that fit the specificity will be displayed. If the '**Check fits**' is enabled, all matching peptide terminals will be marked.



**MS diff.:** The mass difference button will open a dialog box displaying the mass table in an x/y difference table. By highlighting specific differences (i.e. amino acid residues, carbohydrate residues, modifications) you can quickly make a visual inspection for sequence tags, double basic residues in tryptic digests, identify modified residues, oxidations etc.

When the difference dialog closes you will return to the input dialog. For a more detailed description of the table please see Chapter 12.1.

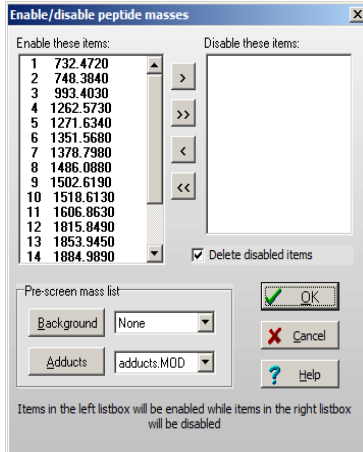
### Enable/disable masses

The dialog, accessed through the '**Edit**' button  of the 'Mass table', enables you to quickly enable and disable individual mass values in a mass list.

Masses are moved from the enabled to the disabled list (and vice versa) by highlighting the relevant masses and then pressing '>' or '<'. Alternatively, you can double click on a mass value to move it to the other list. Pressing '>>' will move all masses values to the other list.

Checking the '**Delete disabled items**' check-box will delete all disabled masses when accepting the dialog box, otherwise the masses will be disabled ('un-checked' in the mass list).

**Pre-screen mass list:** Selecting a mass list from the Background drop-down list and clicking the '**Background**' button will compare all masses in the given mass list against the current mass list and move all mass values, that fit within the given mass precision, into the disabled list. This facility is used to quickly screen for background and/or automatic digest mass values. Pressing the '**Adducts**' button works in a similar manner, but takes a modification file as input (also chosen in a drop-down list box) and compares all mass differences in the current mass list and moves any possible adduct ions to the disabled list. The file called ADDUCTS.MOD will be selected as default if present.



## RESULTS

The results of the mass search are displayed on two pages in a notebook-like window. You can change between the two views by selecting the appropriate tab at the bottom of the display.

## 6- Mass search / FastA files

☐ - / 1234.69

Analyze

Report

The analyze page gives you a complete view of all peptides that are potential 'hits' for the given mass list with the chosen parameters. The results page summarizes the results of your search and includes a view of the protein sequence. You can switch between the two views at any point. Notice that changes made in the 'Analyze' section will be reflected in the 'Report' section.

Peptide hits resulting from the mass search are displayed on the '**Analyze**' page in order of mass, with the closest fit to each mass first. If a modification file has been selected for the search, each modification up to the number specified will be searched for and, if found, will be listed immediately after each primary peptide hit.

**Result grid:** The results of the mass search will be displayed in a spreadsheet-like grid. The exact number of columns will depend on the search and the settings in the 'Customize' box (see below).

Mass search results - abrf_horsemyo.PEP									
Precision in ppm 40.00 E Trypsin Customize									
#	Input	Found	ppm	#Mod	Mod name	Mod mass	First	Last	Sequence
<input type="checkbox"/>	732.47								
<input type="checkbox"/>	748.38								
<input type="checkbox"/>	993.40								
<input type="checkbox"/>	1262.57								
<input checked="" type="checkbox"/>	1271.63	1270.66	23	-			32	42	IR#LFTCHPETLEK#FD
<input type="checkbox"/>	1351.57								
<input checked="" type="checkbox"/>	1378.80	1377.83	32	-			64	77	KK#HGTVVLTALGCIK#K
<input type="checkbox"/>	1486.09								
<input checked="" type="checkbox"/>	1502.62	1501.66	33	-			119	133	SR#HPGDFCADAQGCANTK#
<input checked="" type="checkbox"/>	1518.61	1501.66	34	1 Oxygen	-	15.99	119	133	SR#HPGDFCADAQGCANTK#
<input checked="" type="checkbox"/>	1606.86	1605.85	-5	-			17	31	GR#VEADIAHGQEVLR#
<input checked="" type="checkbox"/>	1815.85	1814.90	29	-			1	16	#GLSDGEWQQVNLNVGK#V
<input checked="" type="checkbox"/>	1853.94	1852.95	9	-			80	96	KK#GHHEAELKPLAQSHAT
<input checked="" type="checkbox"/>	1884.99	1884.01	17	-			103	118	IK#YLEFISDAIHVLR#
<input checked="" type="checkbox"/>	1982.04	1981.05	6	-			79	96	KK#KCHHEAELKPLAQSHA
<input type="checkbox"/>	2010.06								

The following columns can be displayed:

**#:** If a lin is checked, the corresponding peptide will be included in the 'Report' and the 'Mass/precision' graph.

**Input:** The mass from the input list. Notice that if the input list is M+H, this value will be 1 Da. Higher than the 'Found' column.

**Found:** The mass of the peptide identified as fitting within the mass precision.

**Delta:** Deviation of input and found value presented as either ppm (parts per million), dalton or percent.

**Charge:** Charge of input mass value.

**#Mod:** Number of modifications found.

## 6- Mass search / FastA files

**Mod name:** Name of modification.

**Mod mass:** Total mass of modification.

**First:** Number of the first residue of the hit peptide.

**Last:** Number of the last residue of the hit peptide.

**Sequence:** Sequence of the hit peptide. The two residues preceding and the two residues following the actual hit peptide will be displayed and shown in red. There is also a space to separate the pre- and post-residues from the hit peptide. If the 'Fit to enzyme' options was enabled with the '**Check fits**' option in the search dialog box, all peptide terminals that fit the enzyme specificity will be indicated by a green hash sign (#).

Notice that each column can be sorted by clicking on the header button. Click once more to reverse the sorting order.

### Check-boxes:

The check-boxes to the left-hand column of each peptide hit, enables you to select individual peptides for transfer to the clipboard (**Edit | Copy**) or highlight the corresponding sequences in the parent sequence

All lines that contain a 'perfect fit', i.e. where both the N-terminus and the C-terminus fits the selected enzyme cleavage specificity, will be checked by default.

The lines that are checked are the ones displayed on the 'Report' page.

### Toolbar:

The buttons in the local toolbar enables you to:



Toggle between 2 and 4 decimals



Toggle between 1- and 3-residue display



Redo the search using the same parameters. The results from the next search will open in a separate window (until 3 mass search windows are open).



Open/close the Mass vs. precision graph (graphical representation of the mass hits). See below.



Copy the table to clipboard. The first button copies in HTML format, while the second copies in text format.



Print table or report.



Color code the peptide line. Green: No missed cleavages. Yellow: One missed cleavage. Red: Two or more missed cleavages.



The SS button switches to the third page of the control, the SS linked. This page shows mass values that potentially link two peptides. This

## 6- Mass search / FastA files

button and the third page are only visible if the protein contains cysteines.



Close the mass search window.

Precision in ppm  ☒

The precision can be changed by entering a new value in the edit box and click on the **V** button (or just move the focus by clicking or pressing enter or tab). The value will be interpreted as determined by the first box. This will be either %, Da or ppm, with initial setting as determined in the System setup (Chapter 5.1).

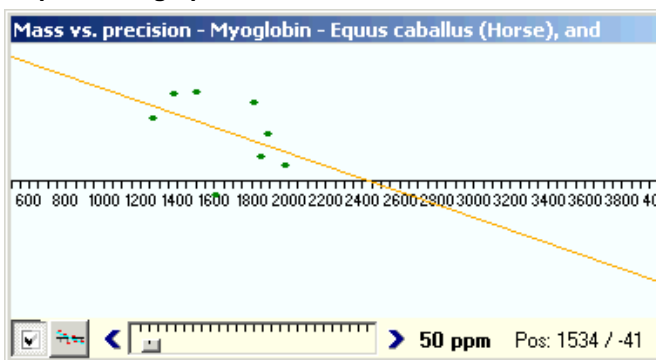
Trypsin


The enzyme cleavage specifications to match against the mass hits can be selected in the drop down list. The button to the right of the list toggles between '**Exact fits**' (a blue '**E**') or '**Check fits**' (a blue '**C**'). When exact fits are chosen, only peptides where both the N-terminus as well as the C-terminus fits the cleavage specifications will be shown in the result list. When check fits is selected matching cleavage specifications will be marked with a green '>'. If both terminals match, the sequence will furthermore be underlined.

Dev. 0.001% 6 ppm

The right-hand panel shows additional information (precision etc.) about the selected peptide.

### Mass vs. precision graph



When the **Frames** button  is depressed in the mass search window toolbar, the Mass vs. precision graph window will be displayed. Deactivating the button will hide the graph.

This window will show the precision of a number of hits from the mass search window in relation to the search mass. Each 'hit' in the mass search window will be represented by a dot in the graph. If the 'hit' fits with the selected enzyme specification the dot will be green, otherwise it will be red. An orange line will show the calibration line for a linear fit using all the points.

## 6- Mass search / FastA files



**Checked peptides.** When this button is in the down position, only peptides in the mass search list that are checked will be displayed in the graph. In this you can turn case points on and off in the mass list, and the calibration line will adjust dynamically. When the button is in the up position, all 'hits' will be displayed



**Calibration button.** Activating this button will transfer the linear calibration (as represented by the orange line) to the mass list and it will be redrawn. The orange line will move to the horizontal axis unless additional points (precision/mass) are entering the precision range. If this is the case, you can perform an additional calibration.



**Precision slider.** This slider control can be used to change the precision displayed in the graph. The arrow at each end will move the precision by 10 ppm in the down/up direction respectively. The precision range of the graph is shown to the right of the slider control.

The position of the cursor is shown in the bottom right position.

The window can be resized by dragging the window edges.

You may zoom the graph by left-click to the left and right on the area to be zoomed. A right-click will reset the graph to the maximum displayed area.

When the precision is set to a high value (e.g. 800 ppm) a gray curved line will show at the top and bottom of the graph, indicating the position of  $\pm 1$  Da. If points fall on this line, it may indicate assignment of the wrong isotope for the monoisotopic mass, or deamidation of an amide.

### Pop-up menu

The pop-up menu (right-click in the window) contains the following commands: **1/3 letter, Redo, Selected peptides, Export, Print, Select font, Help.**

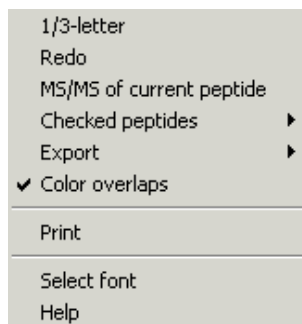
**1/3-letter:** Toggles between 1- and 3-letter peptide display.

**MS/MS of current peptide:** The currently selected peptide in the mass hit list will be transferred to an MS/MS cleavage window for further analysis. Please see chapter 10.1 for details.

**Redo:** Close the result window and opens the 'Mass input' dialog box with all input parameters intact.

**Checked peptides:** The following commands works, in general, only on those lines that have been selected by checking them in the left-hand column of the peptide 'hit' list.

1. **Highlight parent sequence:** The sequences corresponding to the selected peptides will be underlined in the intact protein sequence. If no



## 6- Mass search / FastA files

peptides have been checked, the currently selected line will be highlighted.

2. **Highlight parent sequence (toggle):** This underlines the checked peptides on if they are not underlined, and removes the underline if they already are underlined.
3. **Check perfect fits:** All peptides that have N- and C-terminal sequences that fit the specified enzyme cleavage parameters. If no enzyme has been specified, no action takes place.
4. **Check all hits:** All peptides in the list are checked.
5. **Clear all hits:** All check-boxes are cleared.
6. **Toggle selections:** All peptide selections will be inverted. E.g. checked peptides will be unchecked and vice versa.

**Export:** Two options are available under Export:

1. **ASCII:** Saves the complete 'hit' list to a file on disk. The file will be in text format (ASCII) with the individual columns separated by tab characters (#9). The file will get the extension .DAT.
2. **Copy to clipboard:** The complete result list is copied to the clipboard. This works like the standard 'Copy to clipboard' function (**E**dit | **C**opy or Ctrl-C). If peptides are checked, you will be asked whether you want the complete list or just the checked ones.  
The exact format for copying to the clipboard is determined by the setting in the 'Setup – Peptide' page (Chapter 5.2). In particular you should specify space or tab characters for output (text and spreadsheet analysis respectively).

**Color overlaps:** Depending on the number of overlaps (missed cleavages) in each peptide they will be colored according to:

0 overlaps: Gray  
1 overlap: Green  
2 overlaps: Yellow  
>2 overlaps: Red

The coloring may help you get an overview of complex mass searches.

**Print:** Prints the 'results' list, see below.

**Select font:** Change display font. This will only change the font for the current window.

### Report

The report page shows a two-panel view of the results that are checked on the 'Analyze' page.

The top panel shows the protein sequence (60 residues/line) where all enzyme cleavage points are marked in blue and all 'peptide hits' are shown as horizontal lines below, with the mass of the peptide to the right of each line. The lines are 'layered' so that overlapping peptides are displayed on different levels.

The bottom panel shows statistics on the peptide hits (e.g. residue coverage, number of hits/misses etc.), followed by listings of matching peptides,

## 6- Mass search / FastA files

matching modifications and a list of masses that does not fit anything. At the bottom of the table is listed the modifications searched for.

GLSDGEWQQLNUGKVEADIAGHGQEVLI<sup>1605.85</sup>RLFTGHPETLE<sup>1270.68</sup>KFDK<sup>1938.01</sup>KHLKTE  
L<sup>1377.83</sup>KKHGTVULTALGGIL<sup>1852.85</sup>KKKGHHEAELKPLAQSHAT<sup>1381.07</sup>KK<sup>2109.14</sup>IKYILEFISDAII  
GDFGADAQGAMT<sup>1501.66</sup>KALELFR<sup>747.43</sup>NDIAAK<sup>1359.75</sup>YKELGFQG<sup>940.47</sup>

Residue coverage: 67% [104 of 153]

Peptide hits: 12 Modified: 0 Not identified: 7

### - Identified peptides, no modifications:

input	found	dev.	mc	sequence
735.483 /	734.480	0.005	1	HKIPIK
748.444 /	747.428	-0.008	0	ALELFR
941.468 /	940.465	0.005	0	YKELGFQG

The green divider between the two panels can be shifted with the mouse cursor to change the ratio of the two panels.

To the right of the sequence is a small panel with four buttons.



The top button enables the tryptic PMS rules. If enabled, the peptide list will be colored according to number of missed cleavages (mc). Green if no missed cleavages, red if missed cleavages are present. However, if a basic residue is terminal or next to an acidic residue, it is not counted as a missed cleavage. Additionally Met-containing residues are checked for addition of oxygen (+16) and peptides with N-terminal Gln are checked for pyroGlu (-15). The modifications are colored yellow if not found.



The 'Save report' button activates a small panel giving you the option to name the coverage map (default is the title of the window), save coverage with multiple levels (if the coverage contains overlapping segments – the number of levels is shown in parenthesis) and to select a color for the coverage map. A default color is shown, which can be changed by clicking on the colored panel. The default color is chosen randomly among 16 colors.

If you select 'OK' you can save the report to a file on disk, if you click on 'Clipboard' it will be copied to the clipboard in text format. If you want to copy the coverage as a graphic, you have to click the 'Copy to clipboard' button in the right-hand panel.



**Copy report to clipboard.** This button opens a menu for copying the report to clipboard in various formats:

### Coverage map name:

Mass search results

☒ Save as multiple levels (2)

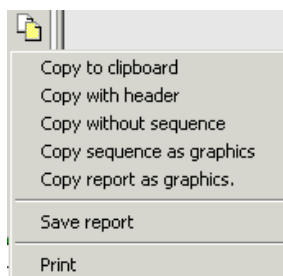
Map color:



☒ OK (to file)

Clipboard

☒ Cancel





## 6- Mass search / FastA files

*Copy to clipboard:* This puts a copy of the entire report on the clipboard in text format.

*Copy with header:* Similar to above, but adds a header with information on sequence name, file, mass, mass file, and mass type.

*Copy without sequence:* Similar to the first choice, but the sequence alignment is not included. This can be useful if you want to combine the table with a graphics picture of the alignment.

*Copy sequence as graphics:* Copies the sequence alignment as a graphics picture. The file is in Windows metafile format (vector), which means it can be rescaled without loss of resolution.

*Copy report as graphics:* Similar to the first choice, but the report is copied as graphics file and cannot be edited.



**Note:** If you want to copy the graphics to Word, Excel or PowerPoint you should start the target application before you copy to clipboard from GPMAW. If you have problems pasting, select **Edit|Paste special...** followed by **Picture (Enhanced metafile)**.



The last button closes the window.

### Print

Printing the result list gives you the choice of selecting 1- or 3-letter residue printing with the default as displayed; all other options will be as displayed. Additionally, you can set the print as 'Normal' (10 point character size) or 'Compact' (8 point).

Sequences that are too long to be printed completely will be truncated in the middle (the truncated part will be shown as three dots '...').



**Hint:** If you want to print long sequences turn the paper orientation to landscape mode (select **File|Printer setup**).

### Search all windows

You should use this option when you have a number of closely similar proteins that you want to search with a given mass list. A typical example is when you have multiple alleles. In this case you open all the different alleles on the desktop, before you perform the mass search (and remember to check this option in the 'Search for mass' dialog).

When you accept the input dialog, GPMAW will search all sequence windows on the desktop (the screen will flicker as the results of multiple search results are extracted), and the collected results will be displayed in a separate dialog box.



**Note:** As the post-processing steps in the 'Multi mass search' results window are much more limited than those in the standard 'Mass search results' you should make sure that all your parameters are accurate before searching all windows.

The results are shown in a multi-page dialog box with the following pages:

## 6- Mass search / FastA files

**Total list:** All the results are collected into a single page listed with the hits for each protein after the name protein.

**Unique:** The hits that are unique to a single protein are listed on this page.

**Common:** Hits that are common to at least two proteins are listed here. Proteins are listed by a single letter. The assignment of protein names to letters can be found on the 'Info' page.

**Individual:** This list works in conjunction with the drop-down protein name list in the toolbar, as only the protein selected in the drop-down list is shown on this page. This is mainly useful if the input search list is rather large.

**Info:** Here is listed the input mass values and the assignment of ID letters to protein names.

**Chart:** A chart showing the mass vs. deviation (see below).

Search ms	Found ms	Delta	Modif	From-To	Sequence
<b>&gt; Myoglobin - Equus caballus (Horse), and</b>					
748.380	747.428	74		134-139	ALELFRR
1271.630	1270.656	26		32- 42	LFTCHPETLEK
1351.570	1350.634	53		51- 62	TEAEHKASEDLK
1378.800	1377.834	30		64- 77	HCTVVLTALGCIK
1502.620	1501.662	33		119-133	HPGDFGADAQGAMTK
1606.860	1605.847	-3		17- 31	VEADIAHGCGEVLIR
1815.850	1814.895	29		1- 16	GLSDGEWQQVNLNVWCK
1853.940	1852.954	12		80- 96	GHHEAELEKPLAQSHATK
1884.990	1884.015	17		103-118	YLEFISDAIIHVLHSEK
1982.040	1981.049	8		79- 96	RGHHEAELEKPLAQSHATK
<b>&gt; dolphin MYOGLOBIN - Tursiops truncatus (Atlantic bottle-nosed dolphin)</b>					
748.380	747.428	74		134-139	ALELFRR
<b>&gt; sheep MYOGLOBIN - Ovis aries (Sheep)</b>					
748.380	747.428	74		134-139	ALELFRR
1271.630	1270.653	24		88- 98	HLAESEHAKHK
1271.630	1270.656	26		32- 42	LFTCHPETLEK

The Total, Unique, Common, and Individual lists shows

**Search ms:** Mass value from the input list

**Found ms:** Mass value found in the protein.

**Delta:** Difference between search and found mass value. The value can be presented either as ppm (part pr million) or as Dalton as determined by the button in the toolbar.

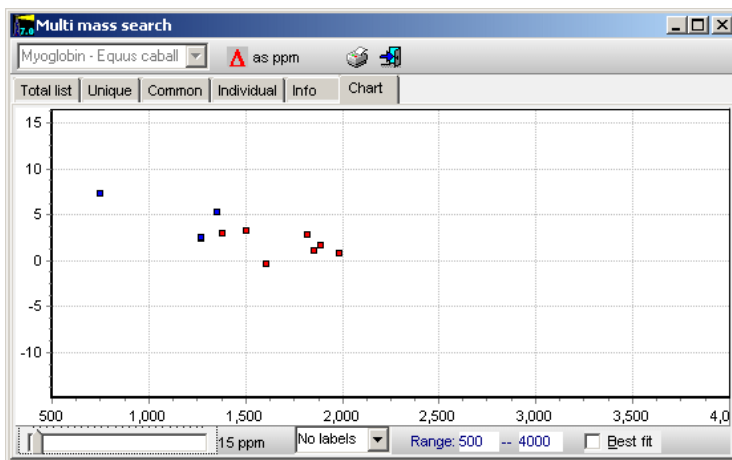
**Modif:** Name of potential modification to match search and found mass values if specified in the mass search input dialog.

**From-To:** Peptide location of found mass value in the target protein.

**Sequence:** Peptide sequence corresponding to the 'From-To' location. On the Unique and Common lists, the sequence is preceded by the protein ID characters (listed on the 'Info' page).

The chart page shows all the hits in a mass vs. precision graph where each protein shows its 'hit' in a different color:

## 6- Mass search / FastA files



Below the graph a slider enables you to zoom the precision (y-axis), a drop-down box gives you the choice of showing labels (all, mass or deviation), the mass range can be edited and the 'Best fit' check box will show you a best line through the listed points (e.g. a calibration line).

### Mass difference

6.2

The mass difference command covers three slightly different mass difference searches that are accessed through a multipage dialog box. As these functions are not linked to a sequence window, they are always available except when a dialog box has the focus.

### Mass difference

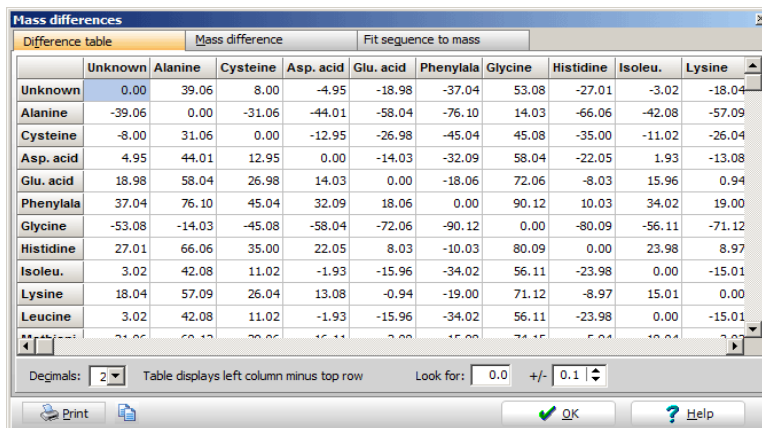
The 'Mass differences' dialog box has three tabs: 'Difference table', 'Mass difference' (selected), and 'Fit sequence to mass'. The 'Mass difference' tab contains a search interface with a text input field labeled 'Search for mass difference:' containing '0.00', a 'Search' button, and an 'Invert' button. To the right is a table with two columns: 'From AA to AA' and 'Difference'. The table is currently empty. At the bottom of the dialog are 'Print', 'OK', and 'Help' buttons.

The '**Mass difference**' enables you to search for mass differences between residues in the currently defined mass table (Chapter 4.2).

## 6- Mass search / FastA files

You enter the difference to search for in the edit box and press **'Search'** button. The differences will be shown in the central table with the closest fit first. By pressing the **'Invert'** button you can invert the search (e.g. search for 33 Da instead of -33 Da).

### Difference table



	Unknown	Alanine	Cysteine	Asp. acid	Glu. acid	Phenylala	Glycine	Histidine	Isoleu.	Lysine
Unknown	0.00	39.06	8.00	-4.95	-18.98	-37.04	53.08	-27.01	-3.02	-18.04
Alanine	-39.06	0.00	-31.06	-44.01	-58.04	-76.10	14.03	-66.06	-42.08	-57.09
Cysteine	-8.00	31.06	0.00	-12.95	-26.98	-45.04	45.08	-35.00	-11.02	-26.04
Asp. acid	4.95	44.01	12.95	0.00	-14.03	-32.09	58.04	-22.05	1.93	-13.08
Glu. acid	18.98	58.04	26.98	14.03	0.00	-18.06	72.06	-8.03	15.96	0.94
Phenylala	37.04	76.10	45.04	32.09	18.06	0.00	90.12	10.03	34.02	19.00
Glycine	-53.08	-14.03	-45.08	-58.04	-72.06	-90.12	0.00	-80.09	-56.11	-71.12
Histidine	27.01	66.06	35.00	22.05	8.03	-10.03	80.09	0.00	23.98	8.97
Isoleu.	3.02	42.08	11.02	-1.93	-15.96	-34.02	56.11	-23.98	0.00	-15.01
Lysine	18.04	57.09	26.04	13.08	-0.94	-19.00	71.12	-8.97	15.01	0.00
Leucine	3.02	42.08	11.02	-1.93	-15.96	-34.02	56.11	-23.98	0.00	-15.01

The difference table is similar to the mass difference function above, except that amino acid residue differences are pre-computed and placed in a table.

The differences are shown with two decimals as default, but by checking the **'4 decimals'** check-box, the table is redrawn with four decimals.

Printing the above table gives you a handy difference table.

### Fit mass to sequence

Occasionally you can end up with the mass of a peptide and only a partial sequence. If you need to know how many sequences that possibly exist for the remaining mass you start the **'Fit mass to sequence'** option.

Start by determining whether you are looking for a peptide or fragment (residues only, i.e. an internal fragment where you have subtracted water for the terminals).

Then establish if the mass is average or monoisotopic.

You enter the search mass in the **'Fit to mass'** box. Check that the precision is suitable. Enter the maximum number of residues to check.

If you know that certain residues are present in the peptide, you can enter them in the 'Known residues' box (1-letter code). These residues will then be omitted from the mass search.

The 'Compositional residues' contains all the residues that may make up the peptide/fragment. If you know that a few residues (e.g. C or W) are not part of the peptide, you should remove them as the number of possibilities decreases drastically even with only a few residues removed.

## 6- Mass search / FastA files

	Deviation	Sequence
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

When the search options are as stringent as possible, press the **‘Do search’** button.

A counter above this button will count down, and upon reaching zero, the table will show all possible sequence combinations that fit within the given mass window, closest fit first.

You can perform a new search by entering new values and press the **‘Do search’** button.

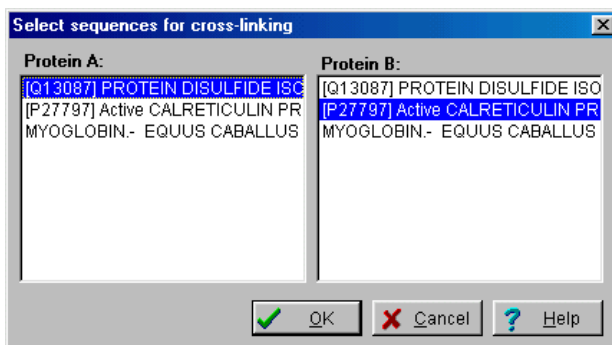


**Note:** The results of the search are compositions, not sequences. The actual sequence may be any permutation of the residues in the result.

### Search for cross-linked peptides

The basis for this function is the cross-linking of proteins using chemical cross-linkers that result in cross-links with a specific molecular mass. The two proteins to be cross-linked can either be different or identical. In the case of identical proteins, you can either search for interchain links (linking two molecules together) or intralinks (links internal in a protein). If you search for intralinks, you need to make certain that you are not analyzing multimers, i.e. you have to perform a size separation (gel filtration, gel electrophoresis) prior to analysis.

## 6- Mass search / FastA files



To analyze for cross-links you start by defining the proteins involved:

- 1) Open the relevant proteins on the GPMW desktop.
- 2) Select the **Search | MS X-link** menu option.
- 3) The 'select proteins' dialog box displays the names of all proteins on the desktop in both list boxes.
- 4) If you want to analyze intra cross-links or cross-links of homo-dimers, you select the same protein on both sides.
- 5) If you want to analyze inter cross-links between different proteins, you select the relevant proteins, one from each list box.

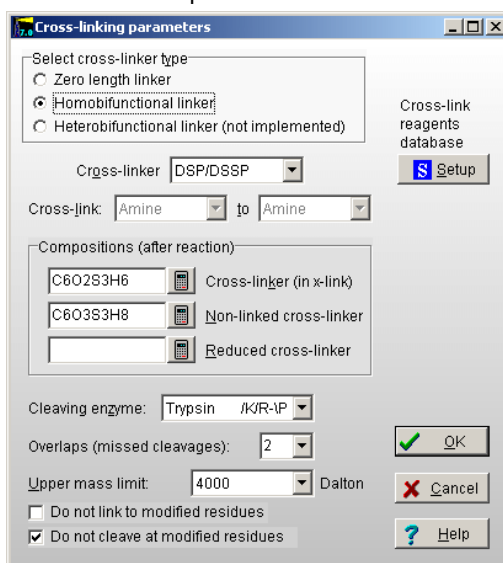
Notice that the protein in the left-hand box will be denoted 'Protein A' and the one in the right-hand box 'Protein B'.

The cross-linking reagent will be selected from the next dialog box which opens when you click 'OK'.

The selection of cross-links is done in three steps:

- 1) Select cross-linker type.
- 2) Select specific cross-linking reagent.
- 3) Select cleavage enzyme and upper mass limit of reporting peptides.

1) Start by selecting the cross-linker type. A zero length cross-linker does not introduce new atoms into the molecules, but typically removes water. Homobifunctional linkers introduce a linker molecule, but links to the same chemical group at both ends (e.g. amine, SH



## 6- Mass search / FastA files

or carboxylic acid). Heterobifunctional linkers (not implemented at present, ver. 4.22) links two different chemical groups (e.g. SH and amine).

Selecting a new cross-linker type will fill the cross-linker drop-down box below with the cross-linkers known to the system.

**2)** Select a suitable cross-linker in the drop-down box. Whenever a new item is selected, the three composition boxes are filled with the appropriate compositions for a) cross-linker; b) non-linked cross-linker (e.g. a cross-linker that is linked at one end but is 'free' at the other end, usually hydrolyzed); 3) reduced cross-linker (if applicable). The analysis of reduced cross-linkers is not supported at present (ver. 4.22).

The cross-link boxes below the compositions show what kind of chemical groups react with the selected cross-linker. If the box shows 'Amine', it means that lysine and the N-terminus can react, if the box shows 'Carb. Acid', it means that glutamic acid, aspartic acid or the C-terminus may react.

**3)** Select the enzyme to use for cleavage analysis (the list is identical to the 'Automatic digest list', see Chapter 9.1) and number of overlaps (missed cleavages). The number of overlaps can be between 2 and 8. Remember that if your cross-linking reagent modifies lysine residues and you use trypsin or Lys-C for cleavage you will need at least an overlap of 1 in all cross-linked peptides (except for the N-terminal peptide). Finally choose the upper mass limit to report. As the list generated is usually quite long, you should choose the upper mass limit that corresponds to your mass spectrometer.

Two options enable you to modify the linkage list with regard to modifications:

**Do not link to modified residues.** If checked, the list of cross-linked residues will not contain peptides where the modified residues were the only potential linkage. E.g. if you cross-link to amine groups, you will need at least one unmodified lysine residue or a free N-terminus.

**Do not cleave at modified residues.** Cleavage will not take place if any of the residues in a cleavage definition has been defined as modified.



**Note:** The first time you run this function, you may get an error message telling you that the file containing the cross-linking reagents has not been found. However, default values will be inserted. Upon completion of the dialog, the file will be saved to disk (as 'xlinker.rea' in the gpmaw\system directory).

### Edit cross-linking reagents

From the cross-linking parameters dialog box above, you can press the **'Setup'** button to go to the 'Edit cross-linking reagents' dialog. In this dialog box you can edit the various parameters for the cross-linkers in GPMW.

Each reagent needs to have a unique name and the **'Link type'** has to be specified to either 'Zero length', 'Homobifunctional' or 'Heterobifunctional' from the drop-down selection box. If either of these requirements is not met, the entry will be deleted upon exit.

The **'Link from'** and **'Link to'** fields are the residues that participate in the cross-linking. The choices are 'Carb. ac.' (carboxylic acid), 'Amine', 'Cysteine', 'Tyrosine' and 'Amide'. For carboxylic acids, the C-terminus of the

## 6- Mass search / FastA files

protein is expected to be able to participate in cross-linking, and likewise for the N-terminus when amine has been selected.

**Edit cross-linking reagents**

#	Name	Link type	Link from	Link to	X-linker	Non-linker	Red-Linker
1	EDC/EDAC	Zero length	Carb. ac.	Amine	-H2O1		
2	DSP/DSSP	Homobifunctional	Amine	Amine	C6O2S3H6	C6O3S3H8	
3	DSS/B53	Homobifunctional	Amine	Amine	C8O2H10	C8O3H12	
4	DST	Homobifunctional	Amine	Amine	C4O4H2	C4O5H4	
5	DSP/DTSSP	Homobifunctional	Amine	Amine	C6O2S2H6	C6O3S2H8	C3O1S1H4
6		Cross-link not set	Carb. ac.	Carb. ac.			
7		Cross-link not set	Carb. ac.	Carb. ac.			
8		Cross-link not set	Carb. ac.	Carb. ac.			
9		Cross-link not set	Carb. ac.	Carb. ac.			

Composition ☒ OK ☐ Cancel ☐ Help

The 'X-linker' field contains the chemical composition that the cross-linker adds to the mass of the cross-linked peptides, i.e. this is not the mass of the cross-linking reagent!

The 'Non-linker' field is the chemical composition of a cross-linker linked at one end, but not at the other end. This will usually be a hydrolyzed reagent (i.e. X-linker + H<sub>2</sub>O).

The 'Red-linker' is a reducible reagent linked to a residue, after reduction. This will typically be half of 'X-linker' + H.

Whenever the focus is on a composition field (the last three columns) the **'Composition'** button will be enabled. Pressing this will open the 'Edit composition' dialog box enabling you to safely edit chemical compositions in the GPMW format (see also Ch. 4.4).

When you exit the dialog, the new settings will be saved on disk and entries entered into the 'Cross-linking parameters' dialog.

## Results

When you in the Cross-linking parameters window select **'OK'**, the window will close and the results will be displayed in a new window.

The results are presented in a list box showing all the peptide masses that may be generated.

The list box show four columns that from left to right (depending on the state of the check-boxes below the list):

- 1) Mass of peptide in Da. The mass type is determined by the Average/Monoiso. button at the bottom of the dialog.
- 2) Residues from protein A that participate in the peptide.
- 3) Residues from protein B that participate in the peptide.
- 4) Type of cross-linked peptide(s):  
**Peptide:** Single non cross-linked peptide from either protein A or protein B.



## 6- Mass search / FastA files

**X-link:** Two cross-linked peptides with one peptide originating from protein A, the other from protein B.

**X-link + linker [1]:** Same as above but with an additional non-linked (hydrolyzed) cross-linker. The angular parenthesis indicate the number of non-linked cross-linkers attached.

**Peptide + linker [1]:** A single peptide from protein A or B with an attached non-linked (hydrolyzed) cross-linker.

The screenshot shows the X-links software interface. At the top, there are tabs for 'X-links', 'ms/ms', and 'Compare'. Below these is a table with columns: 'MH+', 'A seq.', 'B seq.', and 'Type'. The table lists several cross-linked peptides, with the entry '4027.978 400-420 337-346 X-link + linker[1]' highlighted in blue. Below the table, there are checkboxes for 'Peptides' (checked), 'Omit homopeptides' (unchecked), '+ Cross-link' (checked), '+ One free linker' (checked), and '+ Cross-link + free linker' (checked). To the right of these checkboxes, it says '19533 entries'. There are also radio buttons for 'Mgnoiso.' (selected), 'Average', 'MH+', and 'MH2+'. Below these, there is a 'Mass search' button, a 'Precision (ppm):' field set to '50', an 'Isotope pattern' dropdown set to '0', and an 'MS/MS' button. At the bottom, there is a 'Print' button, a 'Copy text' button, a 'Help' button, and a 'Done' button. The bottom of the window shows a sequence: 'LHLVDEPQNLIQNCDFEK' and 'DVCKIYQEAQ'.

MH+	A seq.	B seq.	Type
4026.981	490-498	337-359	X-link + linker[1]
4026.998	524-544	223-235	X-link
4027.003	490-498	337-360	X-link
4027.024	242-256	20- 34	X-link + linker[1]
4027.027	300-318	221-235	X-link
4027.065	562-580	219-232	X-link
4027.286	434-451	233-248	X-link
4027.301	545-561	89-105	X-link
4027.343	548-568	233-245	X-link
4027.978	400-420	337-346	X-link + linker[1]
4028.030	581-607	257-266	X-link
4028.035	319-340	229-241	X-link
4028.106	434-468		Int. X-link

What peptides are calculated?

If the cross-linker links to amine residues, both the N-terminus and lysine residues are marked as potential linked residues. If the N-terminus is blocked you have to disregard N-terminal peptides.

If the cross-linker links to carboxylic acids, the C-terminus, glutamic acid and aspartic acid are potential linker residues.

If trypsin is specified as the cleaving enzyme together with a linker that links to amine residues, a peptide that is reported as having either a X-link or a (hydrolyzed) linker has an 'internal' lysine in addition to a terminal lysine/arginine. An alternative to the 'internal' lysine is the N-terminus.



**Note:** Due to the fact that in many cases you need 'internal' cross-linking residues in the peptides, all peptides are calculated with an overlap level (missed cleavages) of 0, 1 and 2 etc up the level specified in the parameters dialog.

If you cross-link identical proteins (i.e. name of protein A is equal to name of protein B) 'reverse' links are removed from the list. I.e. if peptide 25-48 is linked to peptide 78-101 the reverse link peptide 78-101 to peptide 25-48 is not listed.

## 6- Mass search / FastA files

### Options

In order to limit the number of reported peptides you can turn off the display of **Peptide** (non-linked), **Peptide + Cross-links**, **Peptide + Free linker** and/or **Peptide + Cross-link + free linker** by 'un-checking' the corresponding check boxes below the peptide list.

The '**Omit homopeptides**' will remove links between peptides that overlap in their sequences from the list. This is both identical peptides, but also peptides that share part of their sequence, e.g. 56-68 and 60-72. If you are certain that you only have monomers in your sample, checking this option will thus simplify the list of cross-links.

The total number of entries in the peptide list is displayed after this line.

On the right side, you can choose between displaying **average** and **monoisotopic** masses, and **molecular** and **singly charged** ion species.

If the **MH2+** box is checked, the peptide list is displayed with doubly charged mass values in addition to the singly charged mass values.

**Print:** Prints the peptide list in a two-column layout to conserve space. Check the '**Print comparison list only**' to limit the print to only the mass search list (see the 'Compare to mass list' below).

**Copy to clipboard:** Copies the complete peptides list to the clipboard unless part of the list has been selected. In this case only the selected part of the list is copied to the clipboard (in the pop-up menu you can choose to copy the complete list even if part of it has been highlighted).

### Compare to mass list

As the peptide list quickly gets to be very large it is very convenient to be able to compare the list to a mass table (peak table).

Start by selecting the appropriate mass precision for the comparison. This value has to be in ppm (part per million – 1000 ppm = 0.1%). The press the '**Compare to ms list**' button and enter the mass list in the input table dialog box (see chapter 12.1 for details on file and clipboard lists).

If a mass list has been entered the peptide list will be split with the original list of potential cross-links at the top and the mass 'hits' listed below along with the deviation in Dalton and ppm (part per million) listed in the right-hand part of the list along with the search mass.

383.24	264-266	Peptide				
388.26	545-547	Peptide				
MH+	A seq.	B seq.	Type	Deviation	ppm	Mass
1518.67	246-256		Peptide + linker	0.055	36	1518.613
1518.74	139-151		Peptide	0.126	83	1518.613

When the peptide list is printed or copied to clipboard, the search results are included at the end of the list.

**Isotope pattern:** You can select isotope patterns from 1 to 12 this will limit the search to only include mass values that are separated by the indicated number of Daltons. This is efficient if your cross-linker contains isotopes, e.g. BS3 is available as a +4 Da isotope (Pierce).

### Various searches

6.4-6.8

This menu item in the **File** menu contains a number of searches that do not naturally find their place in the rest of the menu system. Most work on FastA formatted databases and were previously grouped under the FastA menu item. Most of the functions were developed for very specific purposes, but attempts have been made to make the functions as general as possible in order for others to use them.

**Peptide mass search:** Search with a list of peptide masses in a (small) FastA formatted database, enzyme, modification and sequence pattern can be specified.

**Search for motif / for sequence / with peptide list:** This is a suite of three functions that share a very similar algorithm and use the same dialog box. When they are called, each function redefines the dialog for that specific purpose.

**Mass list matching:** Use two different mass files and select the mass differences between ions that is of interest. This will typically be a chemical modification or a glycan, but can be any value.

**Peptide list mass search:** If you have a list of peptides and want to search it against a list of mass values, you generate a file containing the peptides, one per line, and you can then load it and search with a mass list.

**.mgf file filtering:** As the name implies you can perform various operations on mfg formatted peak lists. Among the operations are 'Filter mass values', 'Compare mfg files', 'Graph' and 'Find cleavage'.

### Find mass in FastA database

6.4

When you want to search for peptides in a FastA formatted database, you will in most circumstances use the ms/ms search (see 8.9) or peptide mass fingerprinting (see 8.1). However, in some cases where you get insufficient fragmentation you may want to search directly for specific mass values.

This function will generally only work efficiently if you are searching with high precision (i.e. < 5 ppm) and/or small database. You may also want to know exactly how many peptides are present with a given mass (again high precision is essential).

## 6- Mass search / FastA files

**Find in FastA database**

Input mass values:

#	Mass	Intensity
1	1003.51	
2	1014.57	
3	1031.51	
4	1045.56	
5	1057.58	
6	1136.54	
7	1159.62	
8	1163.48	

Parameters:

Precision (ppm): 5

Trypsin /K/R-IP

Sequence pattern: N;X;ST

Input charge: +1

Missed cleavages: 0

Fixed modification: 0.00000

Variable modifications:

- Oxygen O1
- Methylation C1H2
- Phospho H1O3P1
- di-Methylation H4C2
- Acetyl H2C2O1

Peptides: 25698

Run

Exit

Load database C:\Database\Campylobacter.fasta

#	Hit Da	Dev. ppm	Mod. Da	Peptide	Name
7	1159.620	3.24	16	K.EFNIKPSDVK	1 >g 157414323 ref YP_001481579.1  chromosomal replicati
4	1045.560	-4.03	16	R.GFAVVADEVR	18 >g 157414340 ref YP_001481596.1  MCP domain-contain
5	1057.580	4.56	0	R.LFLFDLYK.K	23 >g 157414345 ref YP_001481601.1  ribonucleotide-dipho
9	1187.610	3.90	16	R.EGQSSDIIVT	32 >g 157414354 ref YP_001481610.1  hypothetical protein
5	1057.580	-3.05	0	K.LINEQVSQK.L	39 >g 157414361 ref YP_001481617.1  hypothetical protein
4	1045.560	-4.03	16	R.GFAVVADEVR	136 >g 157414458 ref YP_001481714.1  methyl-accepting d
13	1267.650	-3.94	16	R.EEFFSLINSAK	181 >g 157414503 ref YP_001481759.1  GTP cyclohydrolase

In order to run the search, you have to enter the following information:

**Mass values:** can be entered manually or pasted as a list from the clipboard. The 'intensity' column is not currently used.

**Database:** Press the 'Load database' button and select the FastA formatted database.

**Input charge:** This is the charge of the input mass values. Select between -1 and +6. All input values are calculated with the same charge.

**Missed cleavage:** Select maximum number of missed cleavages (0-4).

**Enzyme:** The list of enzymes to choose between is the same as for automatic digest (see 9.1).

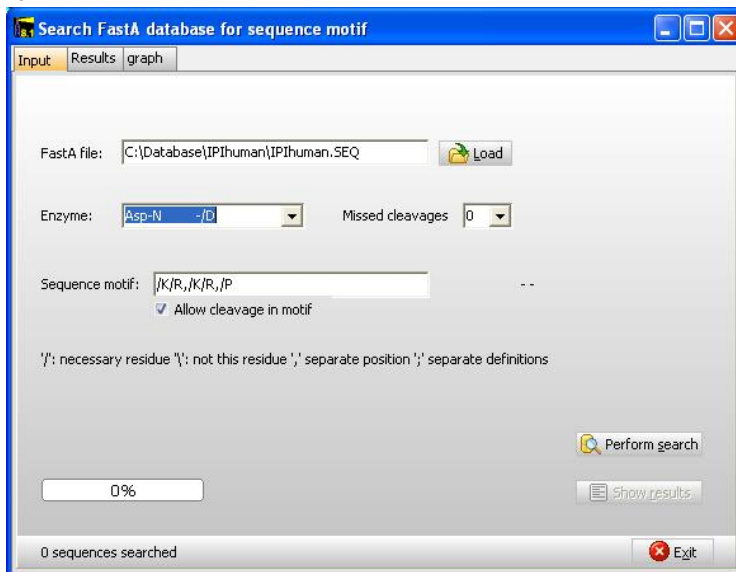
**Sequence pattern:** If you want to specify a sequence pattern, check the box and enter a pattern in the field. Positions in the pattern are separated by ';' and any residue is denoted by 'X'. Up to eight positions can be specified.

**Fixed modification:** A single fixed modification can be entered or selected from the currently loaded modification file by pressing the search button to the right of the entry field. Notice that this opens a drop-down box on top of the 'Run' button, and you have to select an entry to continue.

**Variable modifications:** You can select multiple entries from this list, which is the currently loaded modification file.

## Search FastA for sequence motif

Here (File | Various searches) you can search a FastA formatted file for a simple sequence motif and generate the peptides corresponding to a given digest.



You start by selecting the relevant FastA file through the '**Load**' button.

Then you select the relevant enzyme to cleave the proteins. The enzymes listed are the ones used in Automatic digest (chapter 9.1).

You specify **Missed cleavages** (up to four).

Enter the sequence motif. You use the same nomenclature as used for specifying enzyme specificity in automatic digest:

/ : following residue is necessary

\ : following residue is prohibited

, : separates positions in the definition

; : separates definitions if multiple definitions have to be defined in the same peptide.

Pressing the **Perform search** searches the FastA file, the progress of which can be followed in the progress bar.

When done you can switch to the results page using the **Show results** button or the tab at the top of the window.

## 6- Mass search / FastA files

### Results

Search FastA database for sequence motif

Input Results graph

#	Sequence	Mass	From	To	Motif	Name
2726	DKGSESRIRPMNAFMVWAK	2378.2100	38	57	/K/R,/K/R,	>sp IP10002777 IP10002777 TRANSCRIPTION
1665	DKGQSHPSLQLKKEKLMKAAQKESA	7384.6967	125	190	/K/R,/K/R,	>sp IP10001674 IP10001674 IL-1F7B - Homo
116	DKGPLVPTLPFPLRKPRAHKYLRLSR	6328.5719	127	180	/K/R,/K/R,	>sp IP10000138 IP10000138 CHROMOBOX PROTEIN
3462	DKGSQEKQKGSEGEKPGQEGKPA5	4314.1701	62	103	/K/R,/K/R,	>sp IP10005631 IP10005631 ISOFORM 1 OF
2133	DKGYLVGQAKL5CSYSHWSAPAPQC	4569.3321	455	496	/K/R,/K/R,	>sp IP10002172 IP10002172 C4B-BINDING PROTEIN
2483	DKGVLHNEVKYSILWRGLPNWVTSAIS	3749.1468	1329	1362	/K/R,/K/R,	>sp IP10002482 IP10002482 ISOFORM A OF
2157	DKGTILPRGPLMLSPSSLPSAFHREVIE	4553.4416	753	792	/K/R,/K/R,	>sp IP10002195 IP10002195 LIPIN-2 - Homo sapiens

Table: Copy Save as file Retrieve sequence No acc. number Exit

The results are presented in a table where the columns show:

**#:** Protein location (number) in the FastA file

**Sequence:** Peptide sequence containing the motif and cleaved by the specified enzyme

**Mass:** Monoisotopic mass of the peptide

**From:** Start of peptide in sequence

**To:** End of peptide in sequence

**Motif:** The specified motif

**Name:** Protein name

The table may be sorted by clicking on the relevant header.

The buttons at the bottom of the window allows you to:

**Copy** the table.

**Save as file.** The file will be saved in a format relevant for loading into Excel.

**Retrieve sequence.** Open the selected peptide as a sequence window.

**No acc. number.** Remove the accession number from the list of protein names – only works for some databases.



**Graph.** Shows a graph displaying the number of peptides found with a given mass.

### Search FastA for sequence

This dialog is essentially the same as above, except that you search for a given sequence:

Peptide sequence:

☒ Only full sequence

If the **Only full sequence** box is checked, only peptides corresponding to the exact match will be reported. If not checked, the specified sequence only has to be part of the peptide.

## 6- Mass search / FastA files

### Search FastA for peptide list

This function is the same as above, except that you search with a list of peptides that you load from a text file saved to disk.

Peptide list file:   Load --

Min. # peptides   

## Mass list matching

6.6


The mass list matching compares two mass lists against each other for differences that match either a defined mass difference list (i.e. a modification file) or N-glycans.


Two mass lists have to be available, both as plain text files:



**Source:** A plain list with one mass value pr line. When opened, the mass values will be shown in the second column 'Mass'.

**Target:** A text file that may contain multiple columns (ie an Excel sheet saved as a text file). If the first column is not the correct mass value, you can select it in the **Target mass column** box. As you change the column value, you will see the content of the third column in the table change to reflect the chosen target column.

**Mass list matching**

 Load source mass list Source mass file [1263]: C:\Delphi2009\Projects\GPMaw2009\DATA\sara source list.txt

 Load target mass list Target mass file [26]: C:\Delphi2009\Projects\GPMaw2009\DATA\sara\_deglyco.txt

Target mass column:   


Precision in ppm:

Fudge factor:

The fudge factor value is subtracted from the source mass.

All selected variable modifications will be searched for all comparisons with no regard for the target sequence

Precision in ppm is calculated as the sum of source and target mass

**Variable modifications** 


Oxygen [M] 15.99  
Methylation [DE] 14.02  
Phospho [STY] 79.97  
thr\_ala [T] -30.01  
Me-ester [DEST] 14.02  
D-Succ [D] -18.01  
Sodiated [DE] 21.98  
Deamidation [DE] -1.01  
Acetyl [K] 42.01  
di-Methylation [K] 28.03  
Methyl [K] 14.02

**Search type:**





☒ Search N-Glycans

☐ Search var. modifications

Note: When choosing variable modifications, only the mass (not the valid residues) will be used for searching

 Run search

	Mass	Target	Difference	Compos	Delta ppm	Modif	Sequence	Protein
1	4166.659	940.556	3226.132	CNNNNNNNHHS	3.955		Heparin cofactor 2	dFVnASSK
2	3986.562	976.457	3010.094	CNNNNNNNHNN	1.499		cDNA FLJ57622, highly	eDALnETR
3	4166.676	940.556	3226.132	CNNNNNNNHHS	1.689		Heparin cofactor 2	dFVnASSK
4	3986.558	976.457	3010.094	CNNNNNNNHNN	0.879		cDNA FLJ57622, highly	eDALnETR
5	4101.586	940.556	3161.057	CHHHHHHHHHH	3.715		Heparin cofactor 2	dFVnASSK
6	4101.587	940.556	3161.057	CHHHHHHHHHH	3.500		Heparin cofactor 2	dFVnASSK
7	4164.573	647.368	3517.228	CNNNNNNNHHS	3.603		similar to hCG2042489	nLLNGLNnNnk
8	4164.573	647.368	3517.228	CNNNNNNNHHS	3.603		similar to hCG2042489	nLLNGLNnNnk
9	4164.591	647.368	3517.228	CNNNNNNNHHS	0.779		similar to hCG2042489	nLLNGLNnNnk
10	4085.578	729.357	3356.195	CNNNNNNHSSSS	4.100		Isoform 1 of Vitamin	tYFnR
11	4137.515	976.457	3161.057	CHHHHHHHHHH	0.145		cDNA FLJ57622, highly	eDALnETR

---  Copy  Print  Help  Done

**Precision:** Select the relevant precision. Note that it is calculated as the sum of the source and target mass.

## 6- Mass search / FastA files

**Fudge factor:** This value is added to the mass difference in order to quickly make adjustment for mass changes that were not expected.

**Search type:** You may select either **N-Glycans** or **Variable modifications**. If you search for N-Glycans, the program will search for all mass differences corresponding to a bare core unit up to fully populated complex 5 arm structures with sialic acids and up to three fucose units in addition to high mannose structures up to 20 mannose units. Searching for variable modifications, the current modification file is shown in the central box, where relevant mass values can be selected.

Press the 'S' button to edit or select a different modification file (chapter 4.3).

Press the **Run search** to perform the search, and view the results in the table which can copied to clipboard or printed using the buttons in the bottom toolbar.

In addition to the source and target mass values, the table shows the **difference** in Da., **Compos:** composition of difference (for sugars: C – core unit; H – hexose; N – N-acetylhexosamine; S – sialic acid; F – fucose). **Delta** is difference in ppm, **Modif:** variable modification if specified; **Sequence:** The first column of the target file; **Protein:** the second column in the target file.

### Peptide list mass search

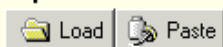
6.7

In some cases it can be advantageous to search a list of peptides for matching mass values instead of searching a digest. This can be the case if you have a list of peptides that does not fit a specific cleavage pattern and can be faster when combining several sequences, perhaps after removal of low/high mass values.

The peptide list mass search is called through the main menu option **File|Various searches|Search peptide list**.

You start out with a list of peptides - sequence in 1-letter code and one sequence per line. The maximum number of residues per line is 255 characters.

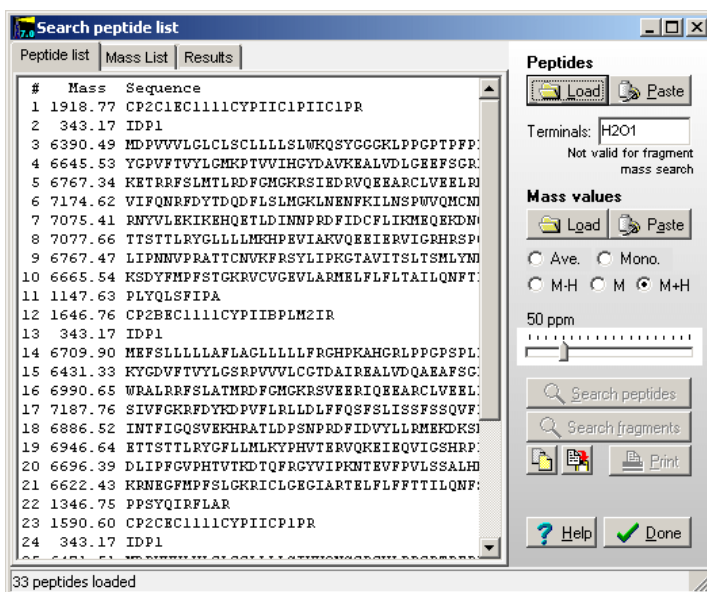
#### Peptides



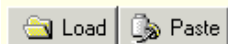
The peptide list can be either read from disk or pasted from the clipboard. The list of peptides has to be in 1-letter code and either a single sequence pr line or in FastA format (i.e. each sequence has a name line starting with '>' followed by (in this case a single) line with the sequence in 1-letter code.



## 6- Mass search / FastA files



### Mass values



After the peptide list is loaded, you can load or paste the search masses. When loaded from disk, the standard file formats used for mass searches (Ch. 6, Ch. 8 and Ch. 12.1) can be read, when pasted from the clipboard, one mass value per line is expected, each line holding one real number.

You can set the following parameters for the search:

Terminals:

**Terminals:** This is the elemental composition of the terminals, default is H<sub>2</sub>O, but any composition following the standard GPMW rules (Ch. 12.2) can be entered.

☐ Ave. ☒ Mono.

**Mass type:** Average or monoisotopic mass values.

☐ M-H ☐ M ☒ M+H

**Charge state:** Singly negative or positive or neutral charge state can be selected.

50 ppm



**Precision:** The precision of the search can be set using the slider bar in the bottom of the dialog.



Input mass search values, either by reading a mass file from disk or pasting from the clipboard. This switches to the 'Mass list' page and enables the 'Search peptides' and 'Search fragments' buttons.

Press the 'Search peptides' button to perform the search, and the dialog box switches to the last tab, showing the results.

## 6- Mass search / FastA files

The results will be listed with mass, precision (in ppm) and 'hit' peptide.

The two buttons below the search

buttons   will copy the results to the clipboard or save to a text file on disk.

517.300	3 ppm	DIAAK
662.340	6 ppm	ASEDLK
666.310	0 ppm	ELGYQG
708.320	-5 ppm	TEAEMK
748.440	6 ppm	ALELFR



**Note:** You can also import a peptide list as a sequence window -> peptide window, however, not in FastA format. Please see Chapter 9.5.

### .mgf file filtering

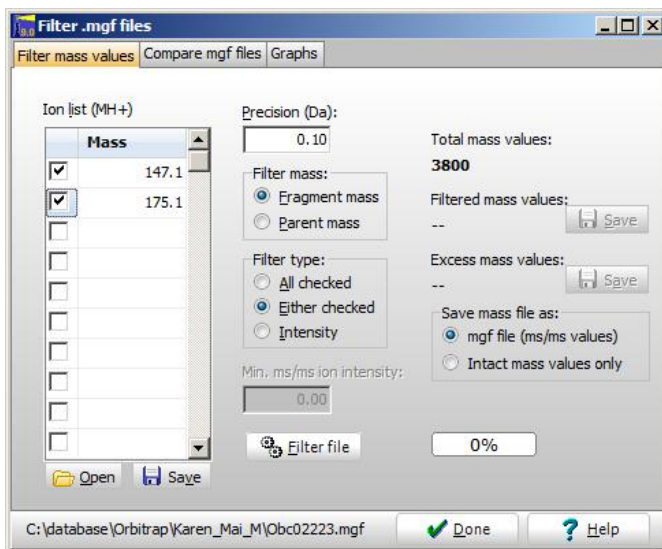
6.8

The .mgf file filtering is an option available from the main menu (File | Various searches | .mgf file filtering) and from the ms/ms search (chapter 8.3) when you have loaded an mgf file.

The functions here are either developed for a specific purpose and then just included as others may have a use for it, or I was just curious for how the content of a given mgf file looked.

Based on the mgf file, you can either filter it for specific fragment ions (i.e. in-/exclude ms/ms spectra containing signature ions) or you can compare it to another mgf file and search for similar parent ions.

When the dialog box opens, it will read the entire mgf file (may take a little time due to the large size of these files), and present you with a tabbed interface:



#### Filter mass values

The ions to search for are entered in the left-hand table. These values may be saved to and loaded from disk using the **Save** and **Open** buttons under the list box. The file format is a straight text file with one mass value / line.

## 6- Mass search / FastA files

You may also paste the list from the clipboard using the same format (right-click in the table and select from the pop-up menu).

You have to enter a search precision in the **Precision** box (in Dalton).

Then you select **Filter mass** as either **Fragment mass** or **Parent mass**.

'Fragment mass' values are the results of the ms/ms analysis, 'Parent mass' are the unfragmented mass values.

If you select 'Fragment mass', you have to select whether to search for **All checked** or **Either checked** values. In the first case only ms/ms spectra where all the values are found will be accepted into the 'filtered' list. In the second case all ms/ms spectra where any of the values is found will be accepted. When searching 'Parent mass' you are searching for all mass values.

For 'Fragment mass' you may alternatively select **Intensity**. In this case the value in the **Min. ms/ms ion intensity**, will be used for selection. This value constitute the combined ion intensity of all fragment ions in a given ms/ms spectrum.

Press the **Filter file** button to perform the search. The progress will be shown in the progress bar.

The result of the filtering will be listed to the right, with two **Save** buttons, one for saving the filtered mass values and one for the excess values (those that do not fulfill the filtering criteria). Depending on the **Save mass file as** the resulting file will either be an .mgf file (ready for applying to a search engine), or it will be the intact parent mass values (text file, one value / line).

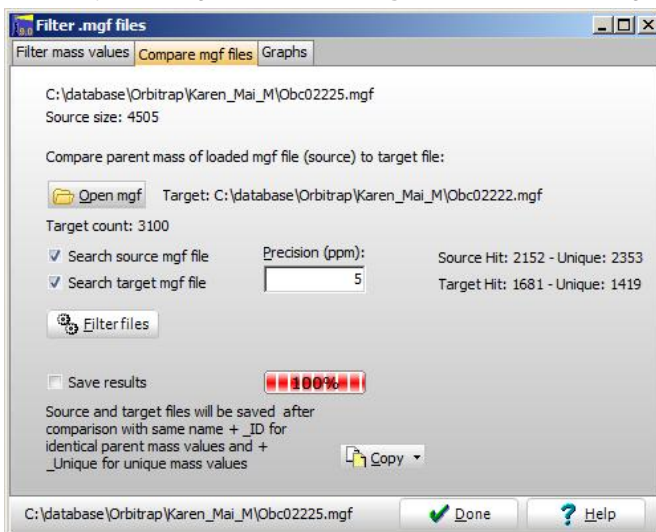
In order to process multiple files, you may drag them from the file explorer onto the dialog box. This will extend the box to reveal a list with the selected mgf files (only files with the extension .mgf will be accepted). You can now process all files by pressing **Filter file**.

The processed files can be viewed in the graph, see below.

## 6- Mass search / FastA files

### Compare mgf files

You may compare the parent mass values of the loaded mgf file (the *Source file*) to another by selecting the **Compare mgf files** tab of the dialog box.



Start by selecting the **Open mgf** button to select the file to compare to, this file is called the *Target* file.

You then select **Search source mgf file** (compare to target) or **Search target mgf file** (compare to source). Remember to set the **Precision** (in ppm).

Pressing the **Filter files** will search all parent mass values and those that are found in the other files, will be added to the 'Hit' list, while the rest will end up in the 'Unique' list.

Progress of the search will be shown in the progress bar.

Note that the result will only be saved if the **Save results** box is checked. In this case the results will be saved in mgf files with the same name as the original file with the addition of '\_ID' and '\_unique' to their names as appropriate.

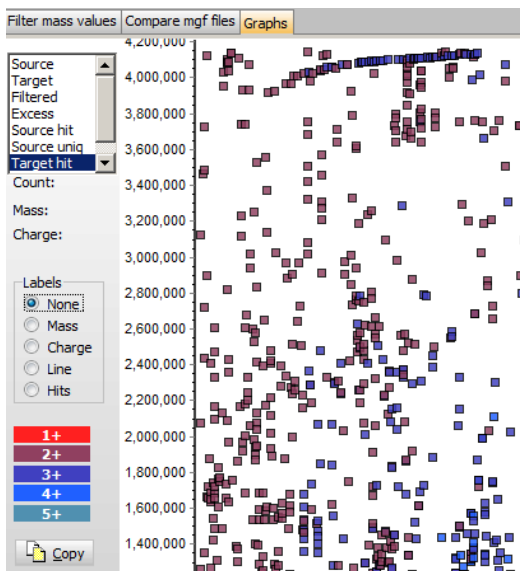
You may also copy the files to the clipboard by selecting the drop-down arrow of the **Copy** button and selecting the appropriate file.

The results of the comparison can be viewed in the graph, see below.

## 6- Mass search / FastA files

### Graph

Selecting the **Graph** page will initially show the entire Source file.



The graph will show the parent mass value along the x-axis and the line in the mgf file in the y-axis (i.e. this will be the elution order).

The appropriate mgf list can then be selected in the top left box. If you click on any of the graph boxes, the corresponding mass and charge will be shown below this box.

You may put **Labels** on the graph points by selecting the appropriate option.

Each point in the graph will be colored according to the charge of the parent ion by the color code to shown to the left.

Finally, you may copy the entire graph to the clipboard through the **Copy** button. The graph will be in metafile format.

### Find cleavage

This function sprang from an interest to determine the specificity of an unknown protease.

You start by selecting the protein which you want to search from the drop-down list. Only the sequence windows available on the GPMW desktop can be selected.

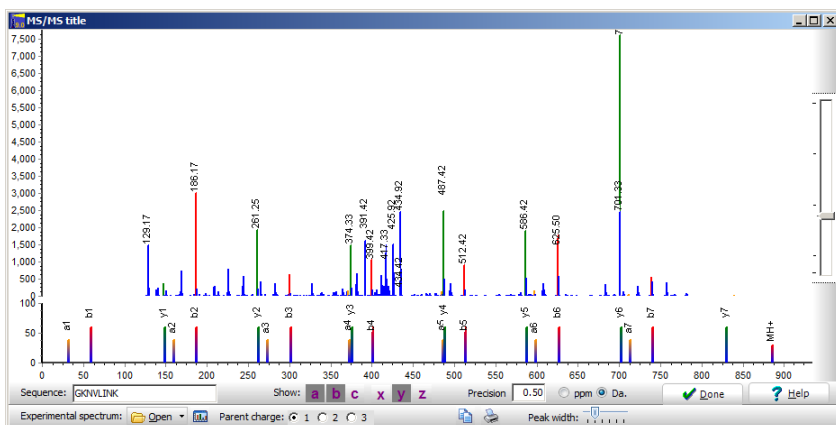
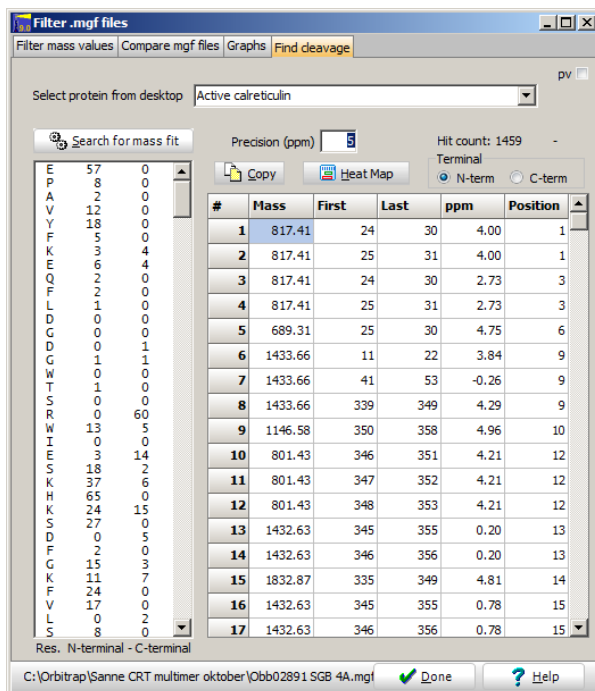
You then enter the required precision – this usually needs to be quite good (< 10 ppm) in order to get meaningful results.

Press the '**Search for mass fit**' button, and the sequence will be searched for all potential peptides that fit with the parent mass of all peptides. The results will be presented in the right-hand table as peptide mass, first and last residue in the peptide hit, precision of the hit, and the position in the .mgf file.

## 6- Mass search / FastA files

The left-hand table can be copied to the clipboard through the 'Copy' button.

When you select an entry in the result grid, the corresponding ms/ms spectrum will show in the separate ms/ms window. This shows the theoretical ms/ms spectrum at the bottom (b- and y-ions) and the experimental spectrum at top.



Click on the 'Heatmap' button to get a coverage map (see chapter 9.6) in the 'heat map' option turned on (there will be no display of a 'normal' coverage. Note that you have to select either the N- or the C-terminal column to transfer to the coverage map.

**FastA file handling****6.9**

The menu option `File | FastA database` gives access to a few commands that helps in handling and extracting sequences from FastA formatted databases.

The first command `'Open for search'` enables you to search directly in an indexed FastA formatted sequence file. For more information see chapter 2.6.

The next two commands `Process FastA` and `Reformat FastA` gives access to the same dialog box, but on different pages.

**Process FastA**

Before opening the dialog, you have to specify a FastA formatted file, and when this is done, the following dialog opens:

You start by specifying the output file, where you then have the following options:

**Save as multiple files** if checked, each sequence will be saved in a different file, with each file have an increasing number appended to the specified output file name. If this option is not checked, the results will be saved in a single file.

**Include header.** If this is checked you will have to check at least one of the following options to include them in the file: **Accession number**, **Protein name**, **Average mass**, **Species** and/or **Sequence**. Species can only be included if it is present in the FastA header, and in a format recognized by

## 6- Mass search / FastA files

GPMaw. The mass is calculated from the sequence based on the currently loaded mass file.

**Ms/ms of entire protein** This will list the mass of each b/y fragment that can be generated from the sequence – one mass/line.

**Include digest.** This option includes the sequences of peptides generated by the enzyme specified and the given missed cleavages. The list of peptides will be listed by number unless the **Sort peptides by mass** is checked, in which case they are sorted with the lightest peptide first.

As the database is processed, the results will be saved to a file and added to the display window.

### Remove signal sequence

If you specify a Swiss-Prot formatted database (e.g. UniProt, IPI, TrEMBL), you cannot get it processed as normal, however, you get the option of converting the file into FastA format with the removal of the signal peptide from the sequence, if it is specified in the sequence annotation.

The result is a standard FastA formatted file.

### Reformat sequences

An alternative option is to reformat the sequence into a format more easily read by e.g. Excel.

The screenshot shows the 'Reformat sequences' dialog box in GPMaw. At the top, the 'Input file' is 'C:\Delphi2010\Projects\GPMaw\DATA\testFastA.txt'. The 'Output file' is 'procFastA', with a note that the file will be located in the input file directory. Below this, there are two tabs: 'Extract sequences' and 'Reformat sequences', with the latter being selected. The 'Line format' section has two radio buttons: 'Reformat as single line' (selected) and 'Keep multiple lines'. The 'Single line delimiter' section has two radio buttons: 'Space delimited' and 'Tab delimited' (selected). The 'Include options' section has four checkboxes: 'Average mass' (checked), 'Monoisotopic mass' (unchecked), 'Length' (unchecked), and 'pI' (unchecked). The 'Amino acid count' checkbox is also unchecked. To the right of these options is a 'Residues to include:' dropdown menu, which is currently set to 'C'.

GPMaw loads the file, and parses it into header (accession number and name) and sequence. You can now choose to save this as a single line or as multiple (two) lines. In addition you can choose to add average mass, monoisotopic mass, sequence length, pI and/or amino acid count. The amino acid count can further be specified as either all residues or a single selected residue.

When saving as a single line, you have to specify whether each field is to be separated by a space character or a tab character. If import into Excel is wanted, using the space character is not good, as the name will be separated into different cells.

If you save as multiple lines, each field will go to a separate line.



## Searches

Search for a given composition in a sequence or local BLAST  
homology searching of a database

Like mass search it is possible to locate a peptide in a protein if you know its amino acid composition. However, as the precision in the determination of amino acid compositions is at least an order of magnitude worse than the average mass spectrum, the search precision can never be as good as 'Mass search' (Chapter 6.1). The low precision is to some degree compensated by the fact that you search for a combination of (at most) 18 amino acids (Asn -> Asp and Gln -> Glu due to hydrolysis) not just a single value.

### Search for composition

7.1

#### Data entry

Res.	Num. of res.
D Asx	0.0
T Thr	0.0
S Ser	0.0
E Glx	0.0
P Pro	0.0
G Gly	0.0
A Ala	0.0
V Val	0.0
C Cys	0.0
M Met	0.0
I Ile	0.0
L Leu	0.0
Y Tyr	0.0
F Phe	0.0
K Lys	0.0
H His	0.0
W Trp	0.0
R Arg	0.0

Number of extra residues to add: 0

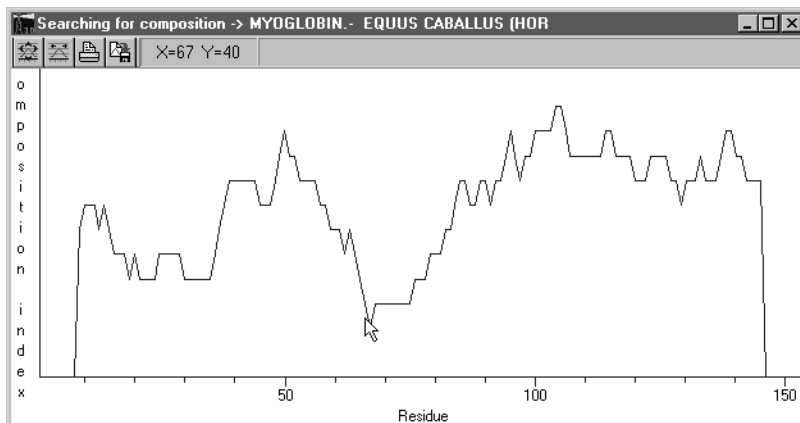
Buttons: OK, Cancel, Help, Zero

You fill in the composition search table with the expected number of residues of the search peptide (not the % composition). You can use decimal numbers as the calculations are carried out with 1 decimal.

The table is persistent between searches (remembers the table values), enabling you to carry out several searches with slight modifications. The **'Zero'** button clears the table, and the **'Number of extra residues to add:'** field adds the required number of 'unknown' residues to the search, resulting in a smoothing effect

## 6- Mass search / FastA files

### Results



The results of the search are shown as a graph of the deviation index (DI) of a sliding search window along the entire sequence

Low points in the graph show areas of the sequence, which have a composition similar to that given in the input dialog box. If the given composition reflects an actual peptide, this will usually be quite evident from the sequence (around residue 66 in the above graph). Please note that the graph starts and stops with a Y-value of zero. If the composition fits a terminal peptide, the graph has to be horizontal or make a sharp downward bend at the terminal.

The position of the cursor is shown in the first panel of the status bar.

For details about general handling of the graph, please see Chapter 11.1.

### References.

- H. Metzger, M.B. Shapiro, J.E. Mosimann & J.E. Vinton, *Nature* 219, 1166-1168 (1968)
- R.J.T. Corbett & R.S. Roche, *Anal. Biochem.* 162, 546-552 (1987).

### Local BLAST homology search

7.2

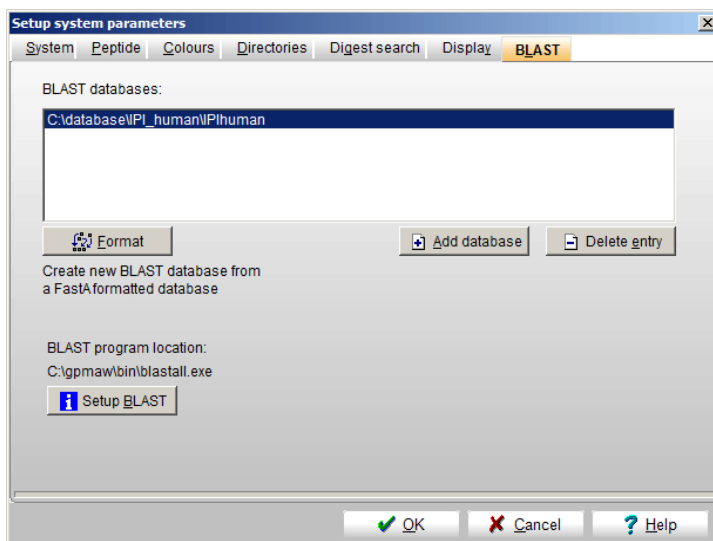
The GPMW BLAST is a local implementation of the BLAST sequence homology search available on the NCBI server (<http://www.ncbi.nlm.nih.gov/BLAST>) and uses the same code compiled for Windows. The search runs as a separate program, but all communication and interface elements are implemented in GPMW, so from a users point of use, it works like an integrated part of GPMW.

The main reasons for using a local implementation of BLAST could be:

- 1) A slow Internet connection (or none at all). When searching a large database, the NCBI server is faster than a local implementation. However, at regular intervals the NCBI BLAST server slows to a crawl.
- 2) A specialized or proprietary database. If you are searching a small genome database (e.g. *E. coli* or *A. thaliana*) a local implementation can be very fast (2-3 seconds).
- 3) Security concerns – communications across the Internet may be compromised.
- 4) Convenience – the local homology search of a protein is just a click away.

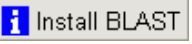
#### Preparations:

The first thing is to make sure that the BLAST homology search program is installed and recognized by GPMW. Open System setup (**Setup | System setup**) and click on the BLAST page.



## 7- Search for composition / BLAST / N-linked glycans

In the bottom left corner is the legend 'BLAST program location'. The line below should show the location of the blastall.exe program. If not, you have


to press the 'Install BLAST' button  and navigate to the blastall.exe program in order to make GPMaw aware of the location. By default the 'blastall.exe' program will be installed in the \gpmaw\bin\ directory along with the other executable files.



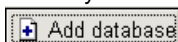
**Tech:** In addition to the 'blastall.exe' you need the file 'formatdb.exe' and a subdirectory called 'DATA' with the following files: 'seqcode.val', 'blosum42', 'blosum62', 'blosum80', 'pam30' and 'pam70'. All these files will normally be installed by GPMaw, but may also be downloaded from the NCBI web server.

When the BLAST program is installed, you need a protein database in FastA format. This can be the same database used for sequence retrieval (chapter 2.6) and/or used for digest database searching (chapter 8). The databases on the GPMaw installation CD-ROM can be used and chapter 12.4 and appendix B contains information on how to retrieve sequences from the Internet.

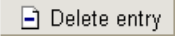
When you have the database you need to reformat it for BLAST:

Click on the '**Format**' button  and in the 'Open file' dialog you navigate to and select the FastA database (e.g. swiss.seq from the GPMaw CD-ROM – it needs to be installed on the hard drive). When selected, the database is quickly converted (note that the converted database takes up approximately the same amount of space as the original database) and you are asked to add it to the list of databases available for local BLAST.

You may add several databases to the list. The '**Add database**' button



may be used to add already converted databases to the list (e.g. if they are shared across a network) and with the '**Delete entry**'

button  you can remove databases from the list.

### Running local BLAST

The local BLAST can be called from all sequence windows, either through the main menu **Search | Local BLAST** or the same command from the pop-up menu.

This will open the BLAST dialog box with the name of the sequence in the top edit box and the sequence in the 'Input sequence' multiline edit box below. If you have installed one or more databases you can select them in the drop-down selection box 'Sequence database'. If no BLAST databases are installed you can go to the setup page by pressing the '**Setup BLAST**' button. The program will remember the most recently used database.

Both the name and the sequence can be edited before performing the BLAST search. Clear both input fields by pressing the '**Clear**' button.

Several parameters can be set to fine-tune the search:

## 7- Search for composition / BLAST / N-linked glycans

**Substitution matrix:** The amino acid substitution matrix used to calculate the score in the homology search. The matrix to use depends on what you are looking for: BLOSUM matrices are based on the comparison of blocks of homologous sequences while the PAM matrices are based on the total alignment of protein sequences. If you are looking for highly divergent proteins you should use low BLOSUM or high PAM values. Looking for closely similar proteins use high BLOSUM or low PAM values. BLOSUM62 is usually a good compromise for general searches.

**Perform gapped search:** This option is usually checked, but when searching for short closely similar sequences you should un-check this option in order to search only for contiguous sequences.

**Filter sequence for low complexity:** If checked, parts of the input sequence that have a low complexity (e.g. skewed composition or simple repetitive regions) will be masked during the search. This box should not be checked when using short sequences for the homology search.

**Expect value:** The expected number of hits from a given database with the given input sequence. E-values up to this value may be reported. When searching with protein sequences values of 1-10 are common, when searching with peptide you can increase the value to 1000-10000.

**Hits to report:** Determines the maximum number of hits to show in the results window.

**Word size:** The word size determines how many residues have to be identical in order to initiate a search of the protein. Choice is 2 (higher sensitivity, slower search) and 3 (lower sensitivity, faster search).

BLAST search input

Sequence name: SERUM ALBUMIN PRECURSOR - Bos taurus (Bovine)

Input sequence:

MKQVTFISLLLLFSSAYSRGVFRDTHKSEIAHRFKDLGEEHFKGLVLIASFQYLQCCPFDEHVKLVNELTEFAKTCVADESHAGCEK  
SLHTLFGDELCKVASLRETYGDMADCKEQEPERNECFLSHKDDSPDLPKLPDPNTLCDEFKADEKKFQGYLYEIAARRHPYFYAPE  
LLYYANKYNGVFOCCQAEKDGACILLPKIETMREKVLASSARQLRCASIQKFGERALKAUSVARLSQKFPKAEFVEVTKLVLDTKTV  
HKECCHGDILLECADDRADLAKYICDNQDTISSKLKECCDKP LLEKSHCIAEVEKDAIPENLP LPTADFAEDKDVCKNYQEAQDAFLGS  
FLYEYSRRHPYAVSVLRLRAKEYEATLEECCKADDPHACYSTYFDKLLKHLVDEPQNLIKQNCDFEKLGEYGFQNALIVRYTRKVPQ  
VSTPTLVEVSRSLGKVGTRCCTKPESERMPCTEDYLSLILNRLCVLHEKTPVSEKVTCKCSTESLVNRRPFCFSALTPDETIVYPKAFDEK

Sequence database: D:\Database\SwissPro\SWISS Expect value: 10000

Substitution matrix: BLOSUM62 Hits to report: 50

☒ Perform gapped search ☐ Filter sequence for low complexity Word size: 3

Setup BLAST

BLAST program location: C:\Delphi7\@projects\gpmaw\BLAST\blastall.exe

Clear OK Cancel Help

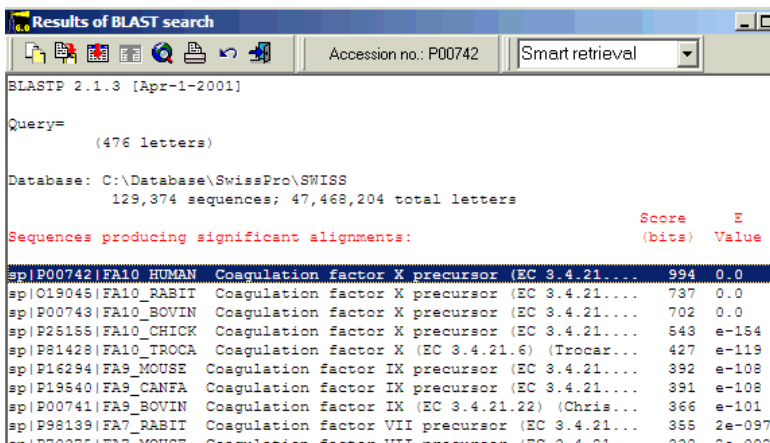
Selecting '**OK**' calls the external BLAST search program and opens the BLAST result window. This window gives the message '>Searching database<', '>Please wait<' and displays a counter that shows the elapsed search time.

```
> Searching database! <
> Please wait <

Elapsed 8 sec.
```

## 7- Search for composition / BLAST / N-linked glycans

Search times depend largely on the size of the database, but the size of the input sequence also has a minor influence.

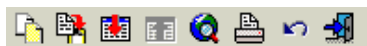


The result of the search is presented in the same window as the search timer. At the top is the date, the name of the input sequence, the name of the database followed by the list of the highest scoring comparisons. Each hit is accompanied by a score and an E-value. The E-value is the likelihood of finding a comparison with this score in a database of this size. Significant similarities are usually taken as E-values below  $10^{-4}$ , but the lower the better. Homology is a theory that can be difficult to prove.

Below the summary list, the highest scoring segments of each hit is presented. At the bottom of the display is search statistics and a reference to the article presenting the algorithm and search program (Altshul et al., 1997).

For an in depth treatment of homologies please consult the articles referenced at the end of this chapter or some of the many books on bioinformatics.

### Toolbar



The toolbar at the top of the window contains the following commands from left to right:

- 1) Copy result table to clipboard.
- 2) Save result list to a file on disk (in text format).
- 3) Move the high scoring segment comparison to the top of the window. This option is only highlighted when a line in the summary list is selected. You can accomplish the same thing by double-clicking on the line.
- 4) Scroll the top of the summary list to the top of the window. Only available when no line in the summary list is selected.

## 7- Search for composition / BLAST / N-linked glycans

- 5) Retrieve sequence into GPMW. This button works in conjunction with retrieval method drop-down list discussed below. This button is only active when a name line is selected. As the sequence retrieval option works through the accession number, it is only active when the BLAST database used for searching is in a format where the accession number is listed in the name line. The accession number used for retrieval is listed in the panel to the right of the toolbar. If there is no accession listed, the function is unable to retrieve a sequence.
- 6) Print the list.
- 7) Redo search. The input data is remembered so the search can be redone using other parameters.
- 8) Close BLAST result window.



The retrieval method drop-down list enables you to specify where the protein should be retrieved from:

**Smart retrieval:** If the accession number starts with O, P or Q the search starts with the Swiss-Prot database (Expasy). If no result, the Entrez database is searched (NCBI). Finally the sequence is retrieved from the local FastA formatted database.

**Entrez->Swiss Prot:** The Entrez database is always searched before the Swiss-Prot database.

**Swiss-Prot->Entrez:** The Swiss-Prot database is always searched before the Entrez database.

**Local FastA:** Only the local database is searched. You should choose this option only if you are not connected to the Internet or have restricted access (firewall).

The advantage of retrieving the sequence from the Internet is that in addition to the sequence you will retrieve the complete database record (see Chapter 3.9 and Appendix B).

### References (BLAST)

- S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Ahang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* 25, 3389-3402 (1997).
- S.F. Altschul, R. J. Carrol, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215, 403 (1990)
- S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89, 10915 (1992).
- S.E. Brenner, C. Chothia, and T.J.P Hubbard. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA*, 95, 6073 (1998).

**ClustalW multiple alignment****7.3****Before you start**

Although ClustalW is in the public domain (e.g. freeware) it is not included in the standard GPMaw installation package due to licensing reasons. However, you can easily download and install the latest ClustalW software by following this procedure:

You can download the ClustalW software package from a number of places on the Internet (just make a search on Google) but the following is one of the 'official' sites:

<ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW/>

**Installation of ClustalW**

If you are using Microsoft Internet Explorer you may proceed as follows (most other browser function similarly):

1. Enter the address above into the Address line of the browser.
2. A list of files on the FTP server is displayed. Right-click on the file named **clustalw1.83.DOS.zip** and select the 'Copy to folder...' option from the menu.
3. Select a temporary (empty) directory to download the file to.
4. The file is compressed (zipped) and you need to decompress the file using an unzipper like winzip ([www.winzip.com](http://www.winzip.com)).
5. After decompression the zipped file has been expanded to a large number of files.
6. Create a folder below the \gpmaw\bin\ folder called 'clustalw' (e.g. the default will be c:\gpmaw\bin\clustalw\ alternatively c:\program files\gpmaw\bin\clustalw\).
7. Copy all the decompressed files into this directory. The files with extension .h, and .c are source files and can be deleted.

**Note:** If you are using Windows XP you need a modified version of the ClustalW executable. From the ftp site above you download the file called **clustalw1.83.XP.zip**. Unzip the file and it will only contain one file called 'clustalw.exe'. Copy this file to the \gpmaw\bin\clustalw\ directory replacing the file of the same name.

You are now ready to use ClustalW from GPMaw.



**Note:** If you have any problems with downloading and/or installation, please check the <http://www.gpmaw.com> site for updated information.

**Performing a multiple alignment.**

Open all the sequences that you want to align on the GPMaw desktop. Select **Search|ClustalW** menu option.

You will now be able to select which proteins you want to align and the alignment options for ClustalW.



## 7- Search for composition / BLAST / N-linked glycans

ClustalW multiple alignment data input

Data input | Alignment

Select proteins for alignment:

OK	Align name	Full protein name
<input checked="" type="checkbox"/>	CALRETICULIN	CALRETICULIN PRECURSOR (CRP55) (CALREGULIN) (HACBP) (ERP60) (52 KDA -
<input type="checkbox"/>	SERUM_ALBUMI	SERUM ALBUMIN PRECURSOR - Bos taurus (Bovine)
<input checked="" type="checkbox"/>	Active_calre	Active calreticulin
<input checked="" type="checkbox"/>	CALNEXIN_PRE	CALNEXIN PRECURSOR (MAJOR HISTOCOMPATIBILITY COMPLEX CLASS I - Homo

Multiple alignment parameters:

Gap opening penalty (10)

Frequent  Rare gaps

Gap extension penalty (0.2)

Long  Short gaps

Temporary file location:

☒ Working directory

☐ c:\temp\

Print Copy Color ☒ Alignment ☐ Residue Bold Align sequences Done Help

All proteins are listed with the full name to the right of the table and an alignment name to the left. By default GMAW selects the first word of the full name, but as ClustalW needs **unique names** for each protein to align, GMAW will add the name line number to the alignment name if the same name occurs multiple times (see figure above). The alignment name can be edited and it can be advantageous to use short names as they are used many times in the final alignment.

The left-hand check boxes determine which proteins are actually used in the alignment – this makes it easy to perform several alignments using different sets of proteins.

When aligning proteins, a score is calculated for the alignment by comparing the amino acid residues according to a given substitution matrix. However, as homologous sequences during evolution has had amino acid residues inserted and deleted from the sequences, it is usually necessary to insert gaps (blanks) into the sequences to make them 'fit' together, in order to create the 'best' alignment.

Inserting unlimited number of gaps into the sequences will lead to alignments without any biological meaning as you will be able to make anything fit. This necessitates the inclusion of penalties for insertion of gaps. Two parameters determine this and are set by the sliders below the list of proteins:

**Gap opening penalty:** This is the penalty for creating a gap (value 5-20).

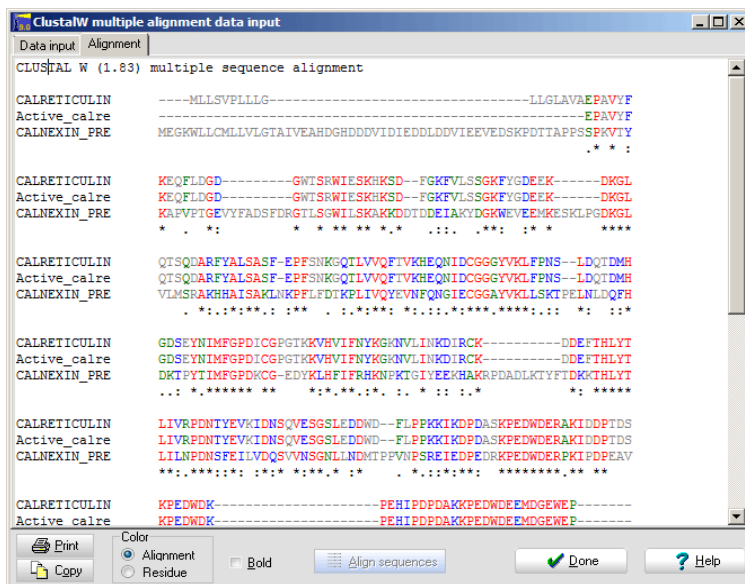
**Gap extension penalty:** Although the main event is the creation of a gap, the length of a gap also has a negative, although smaller, influence. The setting is for each residue in the gap (value 0 – 1).

Press the 'Align now' button to perform the alignment. The view will change to the alignment window, and GMAW will call ClustalW which will run as a

## 7- Search for composition / BLAST / N-linked glycans

DOS program in the background, usually finishing in 5-10 seconds depending on the length and number of proteins to align.

When finished the resulting alignment will be displayed:



The alignment will be shown with 60 residues pr line, and the residues will initially be colored according to the quality of the alignment as well as marked in the consensus line below the alignment:

Red: Fully conserved residues, marked as '\*'.

Blue: Conservative substitutions, marked as '.'.

Green: Similar substitutions, marked as ':'.

Through the pop-up menu (right-click in the alignment) you have additional options:

**Residue color:** Charged residues are blue, hydrophobic are red and Cys is colored green.

**Bold:** Residues are written in bold to make coloring easier to view.

**Copy (Ctrl-C):** Copy the alignment to the clipboard.

**Print:** Print the alignment.

If you want to **redo** the alignment, just click on the **'Data input'** button on the top of the page to return to the data entry screen.

### References (ClustalW)

Higgins D., Thompson J., Gibson T. Thompson J. D., Higgins D. G., Gibson T. J.(1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

## 7- Search for composition / BLAST / N-linked glycans

A full implementation of ClustalW with a good help section can be found on the web at <http://www.ebi.ac.uk/clustalw/#>.

### Analyzing N-linked carbohydrates

7.4

N-linked carbohydrates come in a bewildering array of various forms. As they additionally often are very heterogeneous and furthermore are branched structures, their elucidation by mass spectrometry can be very frustrating. Fortunately all N-linked glycosylations share a common core made up of 5 sugar residues (2 N-acetylgalactosamine and 3 mannose residues). To the first N-acetylgalactosamine residue core a fucose residue can be attached. To the two terminal mannose residues a number of 'arms' (typically 1-4) can be attached.

In the complex type these will be made of N-acetylgalactosamine – mannose units often terminated by a sialic acid residue. The N-acetylgalactosamine residue can be further modified by a fucose residue.

For the high mannose type the arms are made up of mannose residues.

A hybrid type exists where one mannose is extended by mannose residues and the other by N-acetylgalactosamine – mannose units.

Additional variants of the above main type exist, e.g. you can have a bisecting N-acetylgalactosamine on the central mannose residue, that can be further derivatized.

As the mass spectrometer usually cannot distinguish between the different isomers GPMW will label them by their basic structure, e.g. glucose, mannose and galactose will all be labeled hexose or hex. N-acetylgalactosamine will be HexNAc etc. please see appendix C.6.

From the Sequence window (Chapter 3) you can access the different N-glycosylation search windows through the Search | N-glycan menus:

**N-glycan predict:** Based on a known peptide sequence GPMW calculates the mass values of the most common N-linked glycan structures of the various types.

**N-glycan known base:** Based on a known peptide or reduced end modification you can search a peak list for glycans.

**N-glycan delta search:** As the glycosylations very often are heterogeneous, you can often find glycan structures that build up from fairly simple to very complex structures. This window will pull out these carbohydrate 'sequences'.

**N-glycan find peptide:** Knowing that a peak represents a glycosylated peptide is often only half the answer, as determination of the peptide part can be tricky, particularly if the fragment has been generated using a relatively unspecific protease.

The four functions are placed as individual tabs in the same window. As they share the same mass search list

The screenshot shows a window with a list of mass values (m/z) on the left and search controls on the right. The list includes values such as 987.324, 1005.321, 1102.440, 1106.403, 1119.448, 1189.562, 1207.484, 1322.498, 1414.677, 1525.560, 1572.657, 1687.751, 1755.631, 1837.604, 1957.708, and 2012.738. Below the list are buttons for 'Paste', 'Open', 'Copy', and 'M+H'. At the bottom, there is a 'Precision' field set to '50.00 ppm' and a 'Search' button.

## 7- Search for composition / BLAST / N-linked glycans

located in the left-hand side of the window, it is easy to shift between the various search modes and extract information when a single search method is insufficient.

### N-glycan predict

The N-glycan predict will calculate mass values for a number of common N-linked glycans based on a base sequence. If the window has been called from the peptide window, the drop-down box will be filled with all peptides from the peptide list which has a potential N-linked glycosylation site. Whenever one is selected, the corresponding glycosylations are calculated and displayed. If you have entered a mass list in the left-hand table, all mass values that fit within the specified precision will be highlighted.

Note that the display line of the drop-down box can be freely edited, and there is no check for relevance, e.g. does the sequence contain a potential N-linked site (NxT, NxS or NxC; x ≠ P).

Complex type glycosylation						
Core 892.82Da						
Chains	Bare	+1Sia	+2Sia	+3Sia	+4Sia	+5Sia
None	3017.20					
Mono	3382.54	3673.80				
Di	3747.88	4039.13	4330.39			
Tri	4113.21	4404.47	4695.73	4986.99		
Tetra	4478.55	4769.81	5061.07	5352.33	5643.58	
Penta	4843.89	5135.15	5426.40	5717.66	6008.92	6300.18

Core 892.82Da + fucose 146.14Da						
Chains	Bare	+1Sia	+2Sia	+3Sia	+4Sia	+5Sia
None	3163.35					
Mono	3528.68	3819.94				
Di	3894.02	4185.28	4476.54			
Tri	4259.36	4550.62	4841.87	5133.13		
Tetra	4624.69	4915.95	5207.21	5498.47	5789.73	
Penta	4990.03	5281.29	5572.55	5863.81	6155.06	6446.32


The window shows a toolbar at the top and the mass values of the potential glycosylation below.

The main display shows the mass of the peptide at edited/selected in the edit line of the toolbar.. In the main display the type of glycosylation is listed, the mass of the core unit (see appendix C), and the different chains that make up the outer arms of the glycosylation (each arm is a GlcNAc-Gal disaccharide). None means no arms, mono to penta means one to five arms. Each arm can terminate in a sialic acid residue (Sia), but as the stability of this sialic acid is not very high, you will often experience a very heterogeneous population.

The complex table is repeated below with the inclusion of a deoxyhexose unit (fucose) attached to the core.

The '**Bisecting**' check-box adds an N-acetylgalactosamine unit to the core unit, and the '**Extra fucose**' check-box adds an extra deoxyhexose unit.

## 7- Search for composition / BLAST / N-linked glycans

Clicking the **'Glyco type'** button  **Glyco type** toggles to a list of high mannose (0 to 6 mannose units) and hybrid type glycosylations (mono and di- glycosylation arms +/- sialic acid and mannose).

The hybrid table is repeated including a deoxyhexose unit.

When a glycol-peptide mass in the listing on the page match one of the search mass values in the left-hand search list table, these will be highlighted with a colored background.



**Hint:** The drop-down edit box can be edited to any peptide (cut and paste is also supported). This means that you can modify residues and are not limited to the peptides present in your digest. You may even enter a sequence without the N-glycosylation motif, although this is not likely to have any biological relevance.

When the 'Predict N-glycan' function is called from the peptide window (Chapter 9.4), the drop-down peptide list will be populated with all the peptides with a potential N-linked glycan site. When selecting a new peptide, the mass lists below will change to reflect the new masses.

Please read Appendix C.6 for information on the core unit and individual monosaccharide unit masses.

### N-glycan known base

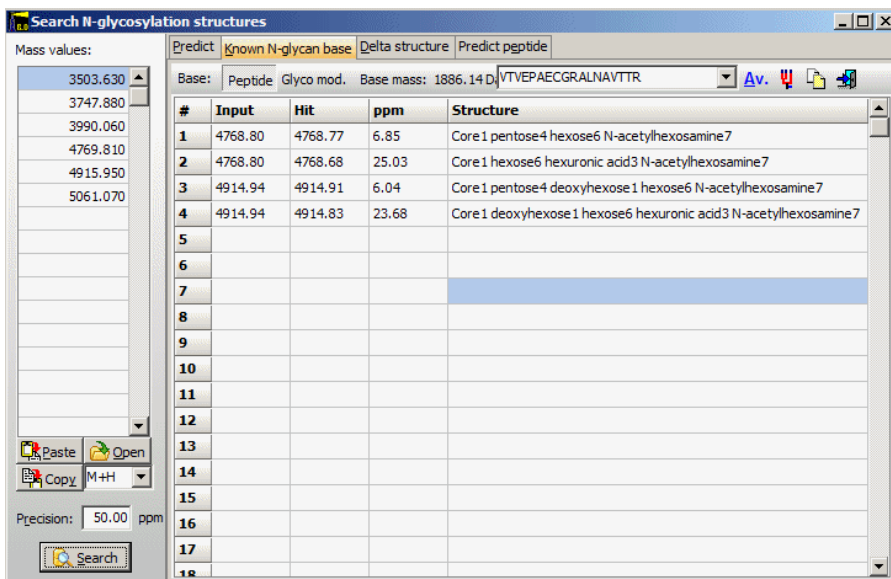
The search section on the N-glycosylation window allows you to use a mass list to search for N-linked glycosylations linked to a given peptide.

This option can also be called directly from the sequence window, in which case the peptide line will show the highlighted peptide. If no peptide is highlighted in the sequence, the peptide window will be empty.

If the search form is called from the peptide window, the edit box will show the first peptide with an N-glycosylation motif and all other peptides having the motif will be available in the drop-down list.

Based on the setting of the base mass type (Peptide/Glyco mod.) this mass is either calculated based on the peptide sequence entered in the edit box in the control bar or it is calculated based on a modification mass selected when pressing the **Glyco mod.** button. This modification will be taken as the base mass, i.e. added to the value of each sugar. The main purpose is to analyze liberated sugar structures that have been modified in the reducing end.

## 7- Search for composition / BLAST / N-linked glycans



#	Input	Hit	ppm	Structure
1	4768.80	4768.77	6.85	Core 1 pentose4 hexose6 N-acetylhexosamine7
2	4768.80	4768.68	25.03	Core 1 hexose6 hexuronic acid3 N-acetylhexosamine7
3	4914.94	4914.91	6.04	Core 1 pentose4 deoxyhexose 1 hexose6 N-acetylhexosamine7
4	4914.94	4914.83	23.68	Core 1 deoxyhexose 1 hexose6 hexuronic acid3 N-acetylhexosamine7
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				

If you have pasted a mass list into the left-hand table, you can now press the **'Search'** button to display all potential glycosylations starting from the base core and then added hexose and N-acetylhexosamine units up to five arms. For each sugar added, all sugars in the 'sugar' database are tried as an addition, as is an additional deoxyhexose unit. The **'Glyco display'** button



toggles between a display of 'core' unit + sugar composition, and a simpler display of 'core' unit and 'arms' (hex + HexNAc units).

### N-glycan delta search


The glycan delta search takes the mass list and searches for differences corresponding to sugar unit mass differences (e.g. 162 Da for hexose units). The program will try to extend any sugar unit differences to the maximum length in order to find related glycan structures.

As the precision in differences of the large fragments often encountered is not necessarily a relative ppm difference, you have to specify a maximum difference (delta) in Da.


If special precautions are not made, glycans will often ionize as sodiated species instead of protonated. To check for this, a check-box in the toolbar controls the check for adducts. In the edit box you can enter a different adduct to check for (e.g. Li).


## 7- Search for composition / BLAST / N-linked glycans


Mass values:

Predict	Known N-glycan base	Delta structure	Predict peptide
Delta:	0.20	Da	<input checked="" type="checkbox"/> Check sodium (Na+)
			21.9825 Da  Use internal table
1	987.32	1119.45	1322.50 1525.56 1687.75
	pentose	N-acetylhex	N-acetylhex hexose
2	2012.74	2174.83	2335.80 2357.78 2519.80 2681.93 2843.87
	hexose	hexosamine	Adduct (Na) hexose hexose hexose
3	2012.74	2174.83	2335.80 2497.78 2519.80 2681.93 2843.87
	hexose	hexosamine	hexose Adduct (Na) hexose hexose
4	2012.74	2174.83	2335.80 2497.78 2659.92 2681.93 2843.87
	hexose	hexosamine	hexose hexose Adduct (Na) hexose
5	2012.74	2174.83	2335.80 2497.78 2659.92 2821.89 2843.87
	hexose	hexosamine	hexose hexose hexose Adduct (Na)
6	2809.01	2971.06	3133.08
	hexose	hexose	
7	3156.12	3318.26	3479.39
	hexose	hexosamine	
8	3457.28	3479.39	
	Adduct (Na)		


Precision: 50.00 ppm



By default the search use the built-in sugar unit table, but you may change to a user-defined modification file by selecting the down arrow next to the 'sugar table' button  Use internal table. The user defined table has to reside in a modification file of the name 'sugar.mod' – see section 4.3 for modification files. Pressing the 'sugar table' button will display the current sugar mass table. The currently active table is shown to the right of the button.

Check sodium (Na+) 21.9825 Da 

Core	892.32
deoxypentose	116.05
pentose	132.04
N-deoxyhexose	146.06
hexosamine	161.07
hexose	162.05
hexuronic acid	176.03
heptose	192.06
N-acetylhexosamine	203.08
muronic acid	275.10



### N-glycan find peptide

Even when you have determined that a given mass value represents a glycosylated peptide, it can be very difficult and time consuming to determine what is the glycosylation and the base peptide. Particularly if you are using a relatively non-specific enzyme to cleave your protein, it will take time.

The 'N-glycan find peptide' function takes your mass list and tries to fit the mass values to a peptide given a number of restraints:

**Peptide length:** The maximum length of peptide, can be adjusted between 1 and 12.

**HexNAc-Hex units:** The number of 'arms' to fit to a glycosylation. Note that this is only used for calculating complex glycosylations. In the calculations, HexNAc units are added up to the specified number. For each iteration, the number of Hex units are varied between 0 and HexNAc units. Deoxyhexose units are also added to a number of 1 (core) + number of HexNAc units.

**NeuAc units:** The number of sialic acids to add to the sugar. The number of sialic acids cannot exceed the number of hex units added.

## 7- Search for composition / BLAST / N-linked glycans

High mannose is searched up to 6 Hex units. Hybrid structures are searched with 0-2 HexNAc-Hex arms and up to 6 Hex units.

Complex, high mannose and hybrid structures can be selected in the search. Each will be shown in a different color in the table – complex as yellow, high mannose as red and hybrid as blue.

**Search N-glycosylation structures**

Mass values: 987.324, 1005.321, 1102.440, 1106.403, 1119.448, 1189.562, 1207.484, 1322.498, 1414.677, 1525.560, 1572.657, 1687.751, 1755.631, 1837.604, 1957.708, 2012.738

Predict: Known N-glycan base Delta structure Predict peptide

Peptide length: 6 HexNAc-Hex units: 4 NeuAc units: 2 ☒ Complex ☒ High mannose ☐ Hybrid

#	Seq.	Site	Mass	Da.	ppm	from	to	NHx	Hex	Fuc	Sia
1	N	1	2809.01	0.01	4	232	232	3	3	3	1
2	NY	1	2194.82	0.01	3	232	233	3	0	3	0
3	FN	1	2194.82	0.01	3	231	232	2	2	0	1
4	KFNY	1	3318.26	0.00	-1	230	233	4	3	2	1
5	KFNYT	1	3318.26	0.00	-1	230	234	3	2	4	1
6	KFNYTE	1	3318.26	0.00	-1	230	235	3	3	4	0
7	N	2	2809.01	0.01	4	288	288	3	3	3	1
8	NMTR	2	3457.28	-0.02	-5	288	291	4	4	0	2
9	GNM	2	3156.12	0.00	0	287	289	3	3	2	2
10	GNMTR	2	3457.28	-0.02	-5	287	291	4	3	3	1
11	KGNM	2	2174.83	-0.01	-4	286	289	2	1	2	0
12	KGNM	2	3457.28	-0.01	-2	286	289	4	2	3	2
13	N	3	2809.01	0.01	4	352	352	3	3	3	1
14	NE	3	1957.71	-0.01	-3	352	353	2	0	3	0
15	NMTR	2	2194.82	-0.01	-4	288	291	0	4	1	0
16	NESG	3	1755.63	0.00	1	352	355	0	2	1	0
17	NY	1	3156.12	0.00	1	232	233	2	8	2	0
18	FN	1	2194.82	0.01	3	231	232	2	2	0	1
19	TNKFN	1	3318.26	0.01	3	228	232	2	6	1	1
20	NMTD	2	3318.26	0.01	2	288	292	1	8	0	1

Precision: 5.00 ppm

Search



---

## Database mass search

Identifying proteins based on mass spectrometric peptide maps, directly in a database based on mass (8.8) or by using ms/ms mass spectral data (peak lists – 8.9).

### Introduction to digest mass search

### 8.1

This is a very powerful and sensitive way of identifying proteins. The method uses the concept that the mass values of peptides generated by a specific enzyme from a given protein (i.e. the peptide mass map) is specific for each protein in the protein database. The peptide mass map is usually so redundant that even a small sub-fraction of peptides is sufficient for identification.

In practical terms you need at least 6-8 peptide masses in the mass range 1000-3000 Da in order to get a reasonable 'hit' in the database. In addition the mass precision has to be reasonably good (0.02% or better). As the number of proteins in the databases gets larger, you can expect to need more peptides and/or higher precision. Mass values below 1000 Da are often not very specific (i.e. a given mass is shared among many proteins) and above 3000 Da mass precision is not very good and usually also contain missed cleavage points (i.e. overlapping peptides).

The sensitivity of the method is entirely dependent on the mass spectrometric identification of peptides, and is usually in the sub-picomole range. Samples can be pure proteins in solution, isolated by gel electrophoresis or by other means.

A major limitation is that, generally, only proteins present in the database can be identified, e.g. you cannot count on finding homologous proteins.

For references, please see end of section.

### The GPMW search.

The search in GPMW is based on a scoring system. The system is quite flexible and the user can easily change the scores and thus optimize the search for particular systems. The scores are set in 'Setup' on the 'Digest src.' page. The scoring system is divided into three parts:

1. **Direct match.** A score is given based on the number of overlaps (missed cleavages) in the database peptide (e.g. zero overlap may give a score of 10, 1 overlap 8, 2 overlaps 6 etc.). This is to reflect fact that the more cleavage points present in a peptide the more unlikely it is.
2. **Better fit:** If the difference between the search and the database peptide is better than half/quarter the given precision and additional

## 8 – Database mass search

score is given. This enables you to specify a looser search precision than you actually have (e.g. 200 ppm instead of 100 ppm) which enables you to catch outliers in your search data while still enabling true 'hits' to get to the top of the result list.

3. **Scoring type:** The normal search type is 'Linear' where scores are listed directly. Use this when you specify a narrow mass search range (e.g. search 30-60 kDa. proteins). When you search a large mass range (e.g. 10-150 kDa.) you should use one of the alternative scoring types. Alternatives are 'Score/NumPep' (score divided by number of peptides in the database protein) and 'Score/Sqrt(NumPep)' (divided by the square root of peptides). The score calculated by these types will compensate for the fact that large proteins tend to give false positives. The 'Score/NumPep' tends to over-compensate (favor small proteins).
4. **Sequence tags:** Finally you can give a score to a sequence tag (a short sequence you have identified in the protein, e.g. by ms/ms experiments) and to an amino acid composition (in some cases you can identify certain residues to be present in a given peptide).

In order to speed up the search, the protein database is 'pre-digested' with the cleavage agent used in the search (e.g. trypsin). This is done in order to speed up the search dramatically.

### Setup digest mass databases

8.2

Before making a digest search you have to set up a number of parameters in the **Setup | Setup system** dialog box on the 'Peptide src.' page (Chapter 5.5). Furthermore, you have to generate digest databases based on a protein sequence databases in FastA format.

#### Directories

The digest database directory (Setup, directories page) specifies the directory where GPMaw looks for digest databases. By default this directory is C:\GPMaw\DATABASE, but can be located anywhere on a local hard drive, CD-ROM or network. It is strongly recommended that you place the digest database on your local hard drive as the actual search is heavily I/O dependent. The protein database itself can be placed on a slow media (i.e. network or CD-ROM) as the speed penalty in retrieving sequences is much less.

#### Peptide search parameters

The peptide search tab specifies search parameters and scoring parameters. The search parameters can be changed in the 'Digest mass search parameters' page of the Setup system (Chapter 5.5).

#### Make digest database

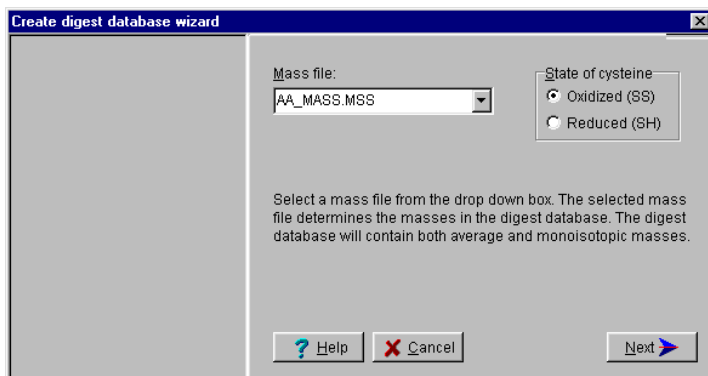
Digest databases can be generated from most protein databases in FastA or PIR/NBRF format (see Appendix B). Swiss-Prot is in a different format and is not accepted directly as input., but has to be converted to FastA format. The EMBL and NCBI non-redundant protein databases need to be modified slightly as the sequences can have extremely long sequence names.

## 8 – Database mass search

The conversion of databases and reduction of complexity can be carried out by the 'Dbindex' utility (Appendix B). More detailed information on the databases and how to obtain them can be found in Appendix B.

The creation of digest databases is carried out by a series of questions in a multipaged dialog box (a 'wizard'). Before you start the wizard you should make certain that you have a proper FastA formatted database ready (see Appendix B) and that you have sufficient space on your harddisk. The final databases will typically take up space corresponding to approximately one quarter to one third of the original database.

The wizard is started from the main menu option **Setup|Make digest database**.



**Mass file:** The initial choice in the wizard is to select the mass file pertinent to the digest search. The drop-down selection box is similar to the one in the main menu. The choice is usually between different modifications of cysteine. If you choose the default file AA\_MASS.MSS (i.e. Cys is defined as mass 102/103) you should also choose whether Cys is in the oxidized or reduced state. Press the **'Next'** button to go to the next choice.

The selections made on each page of the wizard are shown in the left-hand list. You can at any point use the **'Previous'** button to go back and make changes.

## 8 – Database mass search

**Create digest database wizard**

Mass file:  
AA\_MASS.MSS  
Cys is reduced (SH)

Database (input file)  
D:\Database\EMBLnr\EMBLNR.SEQ

Output directory  
D:\Database\EMBLnr\'

☒ Synchronize database and output directory

(1) Select database (in FastA format) to use for creating digest file.  
(2) Select output directory.  
(3) Uncheck the 'Synchronize' checkbox to make output directory different from database directory

? Help X Cancel < Previous Next >

**Database:** In the top edit box you enter the position of the FastA formatted protein database to convert. You can either enter the file path and name manually or you can use the **'Open file'** button to the right of the edit line. The **'Output directory'** is where you want to place the digest database. If the **'Synchronize..'** check-box is checked the output directory will match the database directory. If you want the output to be placed in a different directory than the database you have to un-check this box before entering/selecting the output directory.



**Note:** You will not be able to proceed from this page before you have selected a valid FastA formatted database.

**Create digest database wizard**

Mass file:  
AA\_MASS.MSS  
Cys is reduced (SH)

Database directory:  
D:\Database\EMBLnr

Database file:  
EMBLNR.SEQ

Digest mass file directory:  
D:\Database\EMBLnr\'

Select Enzyme: Trypsin

Cleavage params: IK/R-I-P

Filename: TrypAA\_M

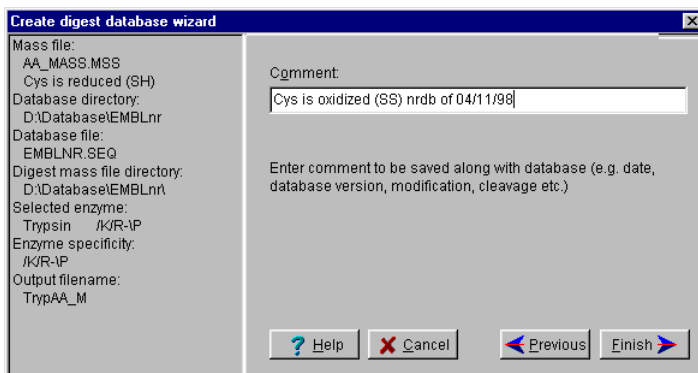
(1) Select enzyme from the drop-down box. If you choose user defined remember to fill out 'Cleavage parameters'.  
(2) Check/edit for correct cleavage parameters.  
(3) Edit filename to reflect enzyme and mass file.

? Help X Cancel < Previous Next >

Next you have to select the **cleavage agent (enzyme)**. The drop-down selection box is similar to the one in automatic digest (Chapter 9.1). This is also where you go to make necessary changes. The **Cleavage parameters** are changed automatically when you make a selection of an enzyme (cleavage agent). However, this box can be edited if you need to make changes.

The **'Filename'** determines the name of the final digest database file. This is created automatically from the first four characters of the enzyme (cleavage agent) and the first four characters of the mass file. You can change the name to a more appropriate one before going to the next page.

## 8 – Database mass search



In the final page of the wizard you can add a comment to the digest database. By default information regarding the state of Cys is included, but you can enter any information up to a maximum of 80 characters. When you press the **'Finish'** button, the digest database files will be created. This is typically a process that takes 1-5 minutes. A dialog with a progress meter will show the development of the database. As the protein database is an ASCII (text) file, the actual state of the progress meter will only be an approximation.

When the creation of the digest database is finished you will see a temporary dialog stating that the mass file has been reinstated. This is because during creation of the digest database, the mass file of your choice (wizard page 1) has been temporarily loaded.

### Digest mass search - data input

8.3

#### Database selection

When you start a search the program will ask you to select a digest database if no database has been previously selected or the previously selected database is no longer present. If you have previously run digest mass search and the previously selected database is still present, it will be selected automatically.

GPMAW can only search FastA formatted databases that have been indexed using the correct mass file (particularly taking modifications of cysteine into consideration). The safest way is to use the **Make digest database** command (see Chapter 8.2) to generate the correct files, as the wizard will take you through all the necessary steps.

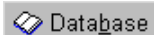
The **'Open database'** dialog box is a standard Windows open file dialog from which you can either

- 1) Select an already prepared database by selecting the .DA2 file (e.g. swiss.da2).
- 2) Select a FastA database (it will have the extension .seq). This will create the correct 'digested' data files (e.g. \*.da2 etc.). This will take considerably more time than using the pre-digested files, however, on subsequent uses of the database you can use the files generated on

## 8 – Database mass search

the first try. **Remember** to set the correct mass file before selecting the database.

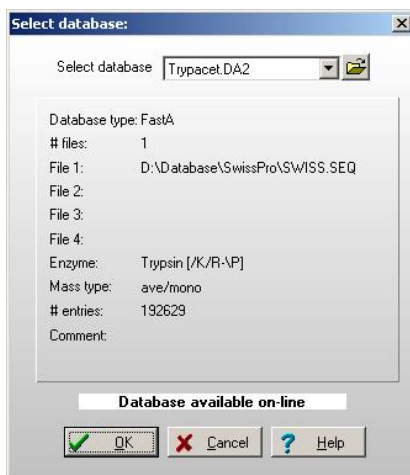
You can change database at any time by selecting the **'Database'** button



### Database information



Pressing the **'Info'** button when a database is opened shows the database information dialog.



**Select database:** By clicking on the drop-down list box at the top, you can select between the digest databases present in the currently selected digest database directory.

**Database information:** The panel below will show the characteristics and data entered when the currently opened database was created.

**Database on-line:** If the protein database is present in the current directory or the location specified in line three, the database is available for information during searching, and the message **'Database available on-line'** will be displayed below the panel. If not, the message will be **'Database not available!'**. In this case you will not be able to retrieve a sequence, obtain pl information, or view the extended report.

### Input of search data:

In the left column you enter the peptide masses to search for. Whenever a mass is entered manually, it will be selected in the next column (indicated with a check-mark, 'v'). Mass values can be selected and deselected by checking and un-checking this box. In the next column you can enter a sequence or composition to search for. The sequence has to be in the standard 1-letter residue code. Depending on the radio buttons below the input field, this column will be interpreted either as a sequence (i.e. sequence as

## 8 – Database mass search

entered), a composition (sequence of residues is ignored) or the N-terminus of the peptide.

**Peptide mass search - data input**

Search data [17]

Mass	OK	Residues
732.4720	<input checked="" type="checkbox"/>	<input type="checkbox"/>
748.3840	<input checked="" type="checkbox"/>	<input type="checkbox"/>
993.4030	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1262.5730	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1271.6340	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1351.5680	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1378.7980	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1486.0880	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1502.6190	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1518.6130	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1606.8630	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1815.8490	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1853.9450	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1884.9890	<input checked="" type="checkbox"/>	<input type="checkbox"/>
1982.0440	<input checked="" type="checkbox"/>	<input type="checkbox"/>
2010.0630	<input checked="" type="checkbox"/>	<input type="checkbox"/>
2394.1270	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Treat 'Residues' as partial:  
☒ sequence ☐ composition ☐ N-terminus

Database  
D:\Database\SwissProt\Trypsin.DA2

Search limits:

Low mass 5 kDa.  
High mass 150 kDa.  
Precision 100 ppm  
Minimum 0.1000 Da.  
Mono mass 3000 Da.  
Overlaps 2  
Min. hits 5  
Mass type M+H

☒ Save settings

Peptide mass list

Load Save  
Paste Copy  
Edit & Pre-screen

Data file info:  
abrf\_horsemyo.PEP

OK Cancel  
Help

### Peptide mass list

**Load and Save:** Enable you to load and save peptide mass lists. GPMW can read peak lists from PerSeptive (GRAMS), Bruker-Daltronik and Hewlett Packard laser-tof mass spectrometers (other file formats will be supported if the demand is present, please contact Lighthouse data). Saving is only supported for GPMW's own peak file format (.PKS, see Appendix A).

**Paste and Copy:** Paste a mass list into the search data table (alternatively use Ctrl+V or the pop-up menu). Copy the mass list to the clipboard.

**Edit and pre-screen:** Enables you to quickly select, de-select and remove masses from the search list. You can also pre-screen the list against a pre-compiled list of masses (e.g. a list of autodigest and/or common background peaks). See also mass search, Chapter 6.1.

## 8 – Database mass search

### Search limits:

The values in the 'Search limits:' box can be edited by selecting the appropriate box and start typing. Alternatively you can click twice or press <F2>. The 'Overlaps' and 'Min. hits' have an up/down arrow when in active edit mode while the 'Mass type' field is a drop-down selection box.

**Mass range:** The minimum and maximum mass of the protein to search for. Used to select only part of the database. You should always give a large allowance for variations in mass assignments (pre- and pro-proteins, fragments etc.). The low limit can usually be left at 10-20 kDa. This will leave out a large number of fragments and very small proteins that in almost all cases will be irrelevant for the mass search. The high limit can be set at 100-200 kDa. This is to remove the influence of a small number of very large proteins in the database that tend to give false positives due to the large number of random hits. This is particularly important if you run the current setup with the 'Score type' as '**Linear**' while not so important if you have selected '**Score divided by the square root of the number of peptides**' (**Score/Sqrt(NumPep)**).

**Precision:** Precision of the mass data obtained. This will either be in % or ppm as defined in Setup (Chapter 5.1).

**Min. Prec.:** Minimum precision of the mass data. If you have difficulties in assigning mass data with absolute precision, you can set this to the best attainable precision, otherwise set it to 0.0.

**Monoiso<:** This field determines the crossover point for monoisotopic masses. If you have a high-resolution mass spectrometer your low mass ions will usually be isotopically resolved enabling you to read the more precise monoisotopic mass. However, above a certain m/z you can no longer resolve the isotopes and you have to revert to average masses. If you only use average masses, you set the 'Monoiso<' value to 0.



**Note:** If you enter only monoisotopic mass values, you have to enter a value in the 'Monoiso<' field higher than the largest monoisotopic mass in your list!

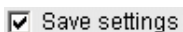
**Max. overlap:** Specifies the maximum number of overlapping peptides that can be allowed in a search mass (e.g. the tryptic peptide GFESRNITK contains an internal tryptic cleavage site and is thus an overlapping peptide with a value of 1). Searches are much faster using a value of 0, but a value of 1 or 2 will usually give a more realistic search pattern (see also 'Optimize' under results below).

**Min. hits:** This value sets the minimum number of peptides that have to match the input masses before being added to the score list. As the score list is sorted during the search, the highest scores are always kept in the list even when there is an 'overflow' of hits. A low value will slow down the search while important hits may be lost with a high value.

**Mass type:** M-H, M or M+H can be selected depending on the input mass type.



## 8 – Database mass search



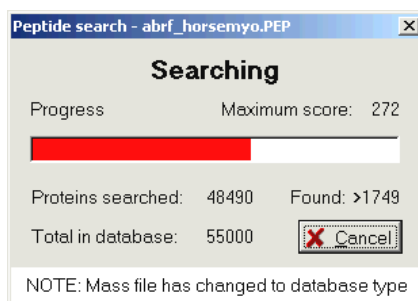
**Save settings:** When this check-box is set, the values entered in the various fields of 'Search limits' will be saved when selecting 'OK'. Default values can be entered in the 'Setup dialog' on the 'Digest src.' page (Chapter 5.5).

### Digest mass search - status window

8.4

When you start a search, GPMW will check whether your currently loaded mass file is identical to the mass file used to generate the digest database. If there is a difference you will be asked if you want to change to the database mass file. The 'correct' mass file is only important when you look at the 'Detailed report' where the calculated mass values are calculated dynamically. In all other instances, the values are based on the ones saved in the digest database.

**During a search,** the status window will be displayed. The red horizontal bar shows the progress of the search. When the search is finished the dialog closes and the result list is displayed.



**Maximum score:** The maximum score encountered in the search.

**Proteins searched:** The number of proteins in the database that fall within the specified mass window (min. and max. mass).

**Total in database:** Total number of proteins encountered. When the search has finished, this value equals the total number of proteins in the database.

**Found:** The number of proteins in the database that have at least the 'min. hits' number of peptides with a mass that fits the search specifications. A '>' in front of the number means that the maximum number of proteins that can be reported has been exceeded (at present 500).

If the mass file used to compile the digest database is different from the currently loaded mass file, and the '**Autoload**' option has been set in 'System setup' Ch. 5.5, you will be notified by a message in the bottom of the dialog box that the program has changed to the correct mass file for the digest analysis.

### Digest mass search - results

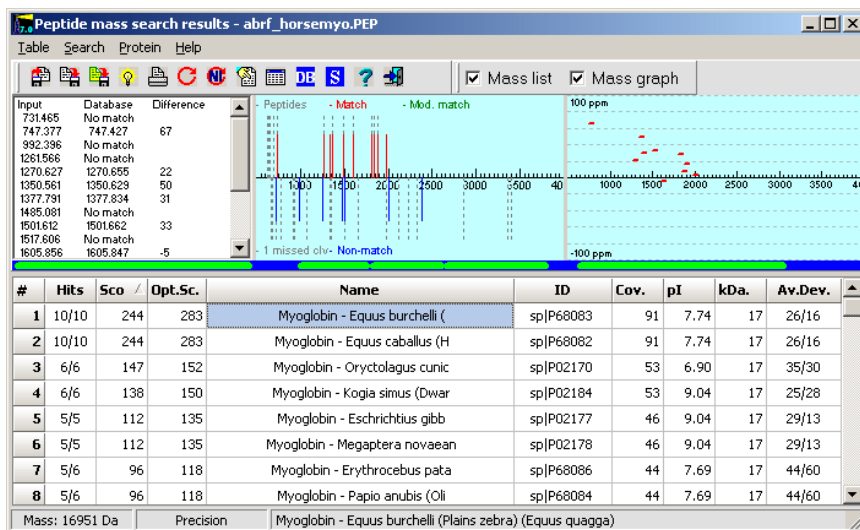
8.5

The results dialog box lists all the proteins found that matches the search criteria up to a maximum limit of 500. If this limit is exceeded, only the highest scoring 500 protein hits will be listed.

## 8 – Database mass search

The dialog is divided into three parts:

1. Top left shows **search information** on the protein selected in the score table. Only the first occurrence of peptide 'hits' will be shown.
2. Top right shows graphically the **precision** of the hits listed in the left box
3. Bottom shows the actual **protein score table** including number of hits, score, short name, ID, mass coverage, pI and mass.



### Score table

The score table is initially sorted by the score (column 3). After performing optimization, the list will be sorted by the optimized score (column 4). By right-clicking on the table you can select either sorting order from the pop-up menu.

#### Table content:

**#:** Line number of table.

**Hit:** Number of peptides that fit the input masses without/with optimization. If several peptides fit the same input mass, only the first one will be reported.

**Score:** The score calculated for the given protein based on the scoring system specified under Setup (Chapter 5.5, Digest mass search). If a given search peptide results in more than one 'hit', only the first is displayed and counted as part of the score even if later 'hits' have a higher precision. The extended report (see below) displays all possible 'hits' in the target protein.

**Opt. sc.:** The optimized score after optimization (see below – Toolbar | Optimize).

**Name:** Name of the protein truncated to 32 characters (except for the note below). The full name of the database entry is shown in the detailed report (see below). Notice that even though you search a non-redundant database

## 8 – Database mass search

you will often experience multiple hits of the same protein. This is because non-redundant databases are seldom really non-redundant, but contains a multitude of proteins with only one or a few amino acid differences. This is particularly noticeable when you hit a protein that has been analyzed by X-ray or NMR analysis (e.g. like the proteins with the 'pdb' in the ID of the figure of the score table).

**ID:** ID or accession number of the protein. These numbers are unique and enable a positive identification in the relevant databases. If the database used is a combined (non-redundant) database (like Owl, NCBI-nr, EMBL-nr) the ID field will often show the origin database. Likely abbreviations are: sp – Swiss-Prot; spn – Swiss-Prot New; spt – Swiss-Prot TREMBL; tr – TREMBL; trn – TREMBL new; gp – GenPept; PIR – Protein Identification Resource; pdb – Protein Data Bank (Brookhaven 3D structure database);

**Cov.:** Coverage of the identified peptide in the given protein (calculated as mass percentage).

**pl:** The calculated pl of the protein. This value is only available if the feature has been turned on in Setup (Chapter 5.5, Digest mass search) and the database is available on-line. The algorithm used is unable to calculate the pl for some proteins; in these cases the pl reported would be 0.0. Three different pl tables are available for calculation; please see Chapter 5.6 Setup Advanced.

**kDa:** The mass of the intact protein.

**Av.Dev.:** The average deviation of the hits. First number is the average deviation in ppm, the second number is the average deviation in mDa.



**Note:** If the complete FastA formatted database is available on-line, the first 25 proteins will be loaded and the pl and full name entered into the list irrespectively of the setting of the pl calculations.

### Hit evaluation help

In order quickly to evaluate whether a 'hit' is significant or not, three panels along the top of the 'hit list' displays relevant information on the currently selected 'hit'. Whenever a new line is selected in the hit list, all three panels are updated. Two of the panels, the left-most 'Mass list' and the central 'Mass graph' can be turned on and off by checking the corresponding boxes in the command bar. The status of these check boxes is remembered between sessions.

The three evaluation windows are divided by resizable splitters, which you may grab and move with the mouse. Likewise the division between the hit list and the evaluation windows is a resizable splitter.

If the entire window is expanded horizontally, only the 'Search precision' window will expand. The other windows have to be expanded manually.

### Search information

When a protein is selected in the score table, information about the number and precision of the peptides constituting the 'hit' for the selected entry will be

## 8 – Database mass search

shown in the top left score table. If multiple 'hits' are present in the protein, only the first 'hit' will be displayed and counted as part of the score.

Input	Database	Difference
1090,613	1090,571	0,042/ 39 ppm
1177,655	No match	
1279,691	1279,672	0,019/ 15 ppm

The list box display input peptides in the left column (if the input mass list was entered as  $M+H^+$  the mass of a proton will be subtracted). The central column will list the 'hit' peptides from the database and the right column will list the mass difference in Da and in ppm (part per million) – a precision of 0.1% is equal to 1000 ppm.

Just above the score table there is a blue line representing the selected protein. The green lines show the 'hit' peptides relative size and position in the protein.

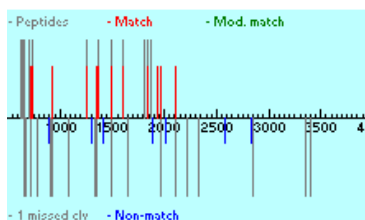
### Mass graph

The mass graph displays theoretical peptide masses for the currently selected protein as well as the input search masses. The graph is essentially identical to the corresponding graph in mass search (Chapter 6.1).

Theoretical peptide masses are shown as gray lines with 'straight' peptides without overlaps going up, while peptides containing a single overlap (or missed cleavage) will point down.

Search peptides will be red and pointing up when there is a hit, and they will be blue and point down when they do not fit with any theoretical mass.

The graph can be zoomed by clicking on it twice, once on the upper and once on the lower zoom limit. After the first click, the mass clicked on will be shown in the bottom right corner. You can reset to default mass range by double-clicking in the window.



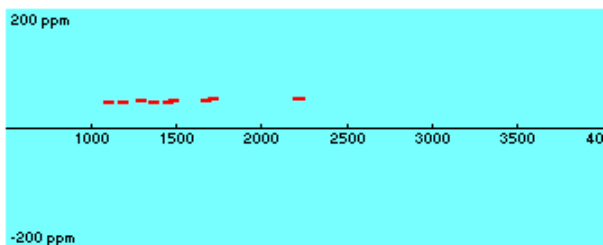
### Search precision

The precision of the currently selected 'hit' in the protein score table can be viewed graphically in the top right box.

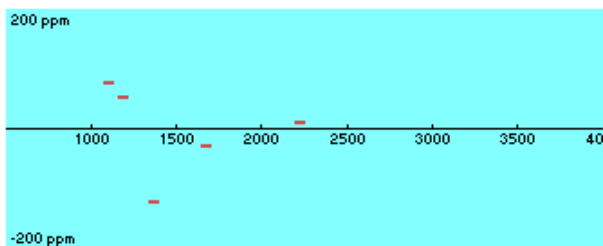
The x-axis shows the masses from 500 to 4000 Da and the y-axis the precision in ppm. The scale is the current mass search precision defined in the mass input dialog.

Each peptide mass 'hit' is displayed as a short red bar. If the 'hit' is the result of optimization it will be drawn in dark red color.

## 8 – Database mass search



The graph is a visual aid in determining the validity of the current 'hit'. In the above example there is a correct 'hit' having a calibration offset of approx. 45 ppm while below is a typical 'random' hit that shows random fluctuations around zero. Another typical calibration error is when you have a more or less constant offset. This will show up as a sloping line.



Although the graph is a convenient aid in determining false positives, you should be careful in the interpretation, as peptide masses are not randomly distributed but falls into mass ranges.

### Toolbar

The buttons in the toolbar are placed in a band that can be 'torn off' with the mouse and positioned anywhere on the screen. When the band is a free-floating window it can be resized.



The table commands are from left to right:

**Load tbl.** and **Save tbl.:** Enable you to load and save the score table. The main reason for saving the score table is to compare different digests of the same protein (see below). If you load a table from disk, you will not be able to view the information that requires the search data (optimization) or the protein database on-line (pl).

**Optimize.** The optimization works only on the proteins in the score table. In the System setup | Digest mass search (chapter 5.5) you can turn on the optimization by linear fit, number of overlaps (missed cleavages) to use and tryptic peptide mass search rules. The linear fit is carried out for each of your hits against the given protein in the database. This will increase the score for all proteins; however, correct hits are more likely to benefit than chance hits. The article by V. Egelhofer contains more details. At the same time the

## 8 – Database mass search

number of overlaps will be increased to the number specified in the setup. The tryptic PMS rules are: if the basic residue is terminal or next to an acidic residue it is not counted as an overlap (and will therefore result in a higher score). You will get an additional score if the peptide starts with Gln and a mass is found at -15 (corresponding to pyro-Glu). If the peptide contains Met and a +16 mass is present an additional score will be added for oxidized Met.

**Print:** Print the score table. You will be given the option of printing the first page of the list (default) or the whole list. In most cases the first page will be sufficient.

**Redo:** Repeat the search. You will be returned to the data input dialog box with all the search data intact, thus enabling you to redo the search using other parameters.

**Redo NI:** Similar to the above command except that only peptide masses **not identified** for the currently selected protein will be reused for the next search, i.e. the identified masses are deleted.

**Get sequence:** If the database is available on-line, this button will be enabled, and by pressing it, the currently selected protein from the score table will be retrieved from the database and displayed in GPMW as a sequence window. Peptides that have been identified during the mass search will be underlined and colored (Pre/post AA see Chapter 5.3, System colors).

**Extended report:** Displays the extended report (also called the second pass search) for the currently selected protein. See below. This option is only available if the database is on-line.

**Database information:** Displays a dialog showing the search database name, database comment, enzyme used in creating the database, enzyme cleavage specificity and number of proteins searched/present in the protein database. The information is essentially the information saved in the '.INF' file.

**Setup:** Opens the Setup system dialog box on the digest search parameter page. Any changes you make will not take effect until your next search.

**Help:** Context sensitive help.

**Exit:** Closes the digest search result table.

### Detailed report (second pass search)

Select the '**Protein|View report**', the '**Detailed report**' button in the toolbar or just double click on an item in the score table to open the 'Detailed report' dialog.

This dialog lists the available information on the currently selected protein in a separate window. If you select a different protein in the score table and requests the Extended report, a new window will not open, but the content of the sequence report window will change to reflect the newly selected protein. The report only displays non-optimized data (i.e. without mass shift and max. overlaps).

In addition to straight 'hits' the report will also show potential 'hits' that corresponds to oxidized methionines (i.e. peptides having a mass 16 Da higher and containing at least one methionine).

## 8 – Database mass search

### The detailed report includes:

The full **protein name** as it appears in the FastA database.

**ID** is accession number. Other accession numbers may appear in the name.

The mass is average mass calculated base on the currently selected mass file.

The **pl** is a calculated value and should be regarded as indicative only. In 'Setup - Advanced' you can choose between different tables for calculating the pl. Appendix C lists the tables for pl calculation.

The **sequence** is shown with identified peptides in red upper case characters and non-identified residues as blue lower case characters. When printing the report, the sequence will be in black and white, but you will be able to differentiate identified residues due to upper/lower case.

The **coverage** is percentage of residues in identified peptides (unlike the coverage in the hit list which is mass percentage identified and may be higher due to multiple peptides covering the same residues).

The screenshot shows a software window titled "Peptide search sequence report". It contains a menu bar with options: Prev, Next, Copy, Save, Print, Help, and Close. There are also checkboxes for "List peptides" and "Graph in print". The main text area displays the following information:

Peptide mass search report (non-optimized data)

Hit no.: 1 of 85      Score: 244

Protein: Myoglobin - Equus burchelli (Plains zebra) (Equus quagga)

ID: sp|P68083

Mass (av): 16951.49 Da      pI: 7.74

GLSDGEWQQV LNVWGKVEAD IAGHCQEVLI RLFTCHPETL EKfdkfkhlk TEAMKASED      60

LKRGHTVVLTL ALGGILKKG HHEAEKPLA QSHATKkip ikYLEFISDA IIVHLSEKHP      120

GDPCADAQCA MTKALELFRn diaaakykelg fgg      153

Coverage: 80.4%

Peptide mass hits:

Measured	Computed	AM/ov	Diff	ppm	Res	Seq
731.465	No match					
747.377	747.427	M/0	0.05	67	134-139	K ALELFR N
992.396	No match					
1261.566	No match					
1270.627	1270.655	M/0	0.03	22	32- 42	R LFTCHPETLEK F
1350.561	1350.629	M/1	0.07	50	51- 62	K TEAMKASEDLK K
1377.791	1377.834	M/0	0.04	31	64- 77	K HCTVVLTLALGGILK K
1485.081	No match					
1501.612	1501.662	M/0	0.05	33	119-133	K HPCGDAQQAQGMATK A
1517.606	No match					

### The peptide mass table:

Unlike the mass table presented in the overall hit list above, which only list the first occurrence of multiple peptides that fit the search profile, the peptide mass table in the detailed report includes all peptides that fit the search mass profile.

- **Measured:** Measured mass (data input corrected for protonation)
- **Calculated:** Calculated masses based on database entry ('No match' means that the given input peptide was not identified in the protein displayed).
- **AM/ov:** A - average mass; M - monoisotopic mass; ov - number of overlapping cleavage sites.
- **Diff:** Mass difference (in Da.) between measured and calculated mass.

## 8 – Database mass search

- **0/00 or ppm:** Mass difference as parts in 1000 or ppm (parts per million) as selected in Setup (Chapter 5.1).
- **Res:** Position of identified peptide.
- **Seq:** Identified peptide sequence with one preceding and following residue.

Below the peptide mass table is: Average differences, number of matches and mismatches. Then follows a list of potential 'hits' if the methionines are oxidized.

At last is given some reference data on the database and input parameters. If the '**List peptides**' check-box in the toolbar is checked, a list of all theoretical peptide mass data for the given protein as found in the digest mass database, no overlapping peptide masses.

If the **Graph in print** is checked, the precision vs. mass graph from the results page will be printed in top of the first page of the printed output.

### The toolbar.

**Prev. / Next:** Display the previous/next protein in the 'hit' list. If you double click in the score table of the parent window, the report will be updated to reflect the 'hit' clicked upon.

**Copy:** Copies the content of the report window to the clipboard. The copy on the clipboard will not contain any formatting characters.

**Save:** Saves the content of the report to disk. This is an ASCII (text) file and can easily be incorporated in a report. It is not particularly amenable for spreadsheet analysis.

**Print:** Prints the report.

**Close:** Close the report window.

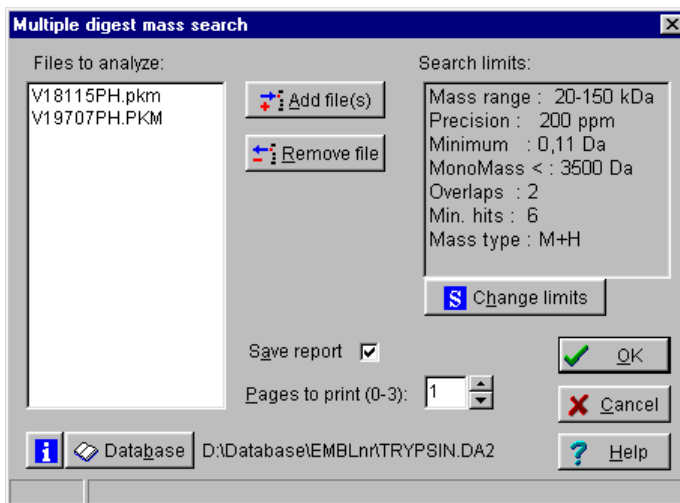
**List peptides:** If checked a list of all potential peptides in the target protein will be listed at the bottom of the report.

**Help:** Open the context sensitive help.



**Multiple digest mass search****8.6**

The multiple digest mass search option enables you to run digest search on multiple peak lists without operator intervention.



You start by selecting the database. Like in the normal database search described above, the most recently used database is automatically selected if present.

The peak files to analyze are either dragged into the list box in the left part of the dialog from File Explorer, or they are selected using the '**Open file**' dialog box that can be activated by pressing the '**Add file(s)**' button. Multiple files can be selected in one operation.

Files can be removed from the list by highlighting the file name and pressing the '**Remove file**' button.

Each search shows a dialog with a progress bar like the single search above. After the last search you are back in GPMW.



**Note:** The multiple digests mass search only works on disk files (peak lists). Furthermore, all searches have to be performed using identical parameters.

**Options.**

**Save report:** At the end of each search, the result list is saved as a .PMS file which enables you to retrieve the results, perform optimization, view the detailed report and rerun the search using different parameters.

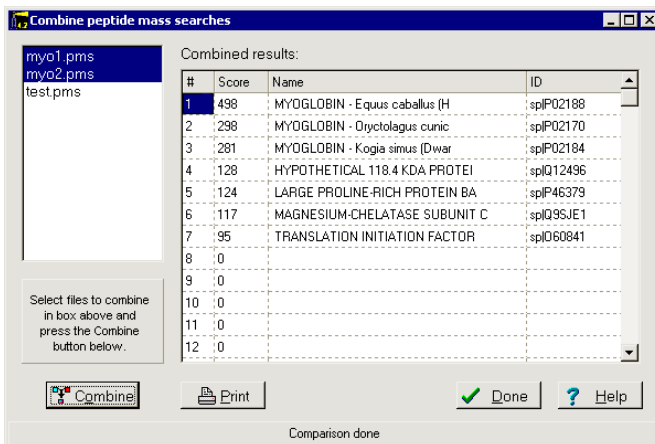
**Pages to print:** Determines how many pages of the result list are to be printed. The default is 1 page (if you save the report, you can go back and print more pages).

**Change limits:** These are the search limits for the current digest search. The limits are identical to the search limits for the single digest search command (see above).

**Combine digest mass search****8.7**

If the specificity of your digest mass search is too low, you can perform multiple mass searches, either using different input parameters or, preferably, different digests, and combine the search results afterwards.

After you have performed a search, you can save the search results in a PMS file (see above and Appendix A). You then select **Search|Combine digest search**.



In the left-hand dialog box, you select the digest results that you want to combine (2-5), press the button and all the selected digest search files will be compared and entries having the same ID will have their scores combined. The final list will be sorted and displayed in the **'Combined results'** table.

**Note:** The list of PMS files is taken from the currently defined 'User directory' (see chapter 5.4).

Notice that the table does not have any links to the original database. The different digest searches have to be carried out on the same database as the comparison based on the ID field and different databases will most likely have different ID's for the same protein.

**Protein mass search****8.8**

Instead of using the peptide masses from a digest, you may also use the mass of the intact protein.

However, this approach is fraught with dangers. Unlike the digest mass search where the information is usually redundant, the search for a protein only contains a single piece of data. Furthermore, the likelihood of the protein being modified is very high. For example, the presence or absence of an initiating methionine (i.e. is it present/absent in the protein and/or in the database). Residues may be chemically modified (e.g. oxidation of methionine), the possibility of adduct ions is greater, and, finally, the protein may be post-translationally modified.

## 8 – Database mass search

If you take the appropriate considerations, you may be able to use the database to search for proteins.

### Database

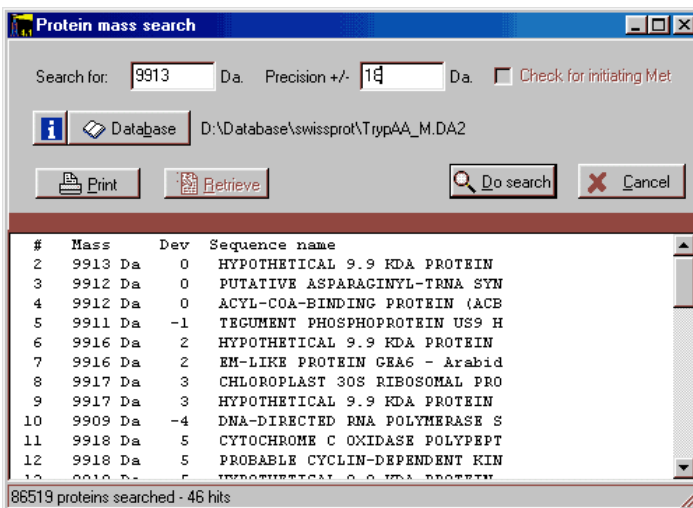
You can use the same kind of databases as used for digest mass search (Chapter 8.2). However, you have to make certain that the mass file used for constructing the derivatised database use the correct mass file. For peptide mass searches you will often derivatise cysteine (i.e. with vinylpyridine) while proteins will often be either oxidized or reduced.

### Data input and searching

Data input is quite simple, as the only parameters needed are:

- Protein mass
- Precision
- Database

The mass of the protein is taken as the average mass in Dalton. The precision is also in Dalton and is +/- . The database option works just like in peptide database searching, Chapter 8.3. If a search database has been used, the previously opened database is automatically opened. You can select a new database by pressing the **'Database'** button and selecting a new database (.DA2 file). Pressing the **'i'** button will open a dialog with information on the protein database.



Only integer data can be entered in the **'Search for'** and **'Precision'** data fields.

When search data has been entered you press the **'Do search'** button. A status bar moves across the dialog box just below the buttons, indicating the progress of the search.

## 8 – Database mass search

The time for a search is in the order of 10-15 seconds for a non-redundant database (~500 000 proteins).

The results are sorted with the best hit at the top. If the 'hit' mass is ~131 Da higher than the search mass, it is because the 'Check for initiating Met' is checked, and the corresponding protein sequence both fits with an additional Met and the sequence actually starts with Met.

### Options

**Check for initiating Met:** When checked, proteins in the database will be checked for the presence of a methionine in position 1. If the methionine is present in the database sequence and the search mass + mass of Met, it will be added to the search results.

**Print:** Prints the search results including the search data.

**Retrieve:** Select a hit result and press the '**Retrieve**' button to load the corresponding protein into GPMaw as a new sequence window. Alternatively you may double click on the relevant entry to open it as a sequence window in GPMaw.

### References (digest mass search)

W.J. Henzel, T.M. Billeci, J.T. Stults & S.C. Wong, Proc. Natl. Acad. Sci. (USA) 90, 5011 (1993).

M. Mann, P. Højrup & P. Roepstorff, Biol. Mass Spectrom. 22, 338 (1993).

D.J.C. Pappin, P. Højrup and A.J. Bleasby, Current Biology 3, 327 (1993).

P. James, M. Quadroni, E. Carafoli & G. Gonnet, Biophys. Biochem. Res. Comm. 195, 58 (1993).

J.R. Yates, S. Speicher, P.R. Griffin & T. Hunkapiller, Anal. Biochem. 214, 397 (1993).

V. Egelhofer, K. Büsow, C. Luebbert, H. Lehrach and E. Nordhoff: Improvements in Protein Identification by MALDI-TOF-MS Peptide Mapping, Anal. Chem., 72, 2741-2750 (2000).

### MS/MS search

8.9

The ms/ms search of GPMaw is based on the public domain search engine 'X! Tandem'. The X! Tandem program is a professional class search engine; Although it is able to perform proteome wide searches, the GPMaw implementation is targeted towards characterizing individual proteins and small collections of proteins. It should be possible to analyze large databases, but the handling and analysis of these data has not been thoroughly tested using the GPMaw implementation, so other search engines (e.g. Mascot, Sequest etc.) can be recommended.

The focus of the GPMaw implementation is ease of use and fast analysis time. Currently all the functionality of X! Tandem is not implemented, but this is likely to improve in future versions of GPMaw.

## 8 – Database mass search



**Note:** For more details on the X! Tandem search engine, particularly all the parameters that can be adjusted, see the web site of The Global Proteome Machine Organization, <http://www.thegpm.org>.

The MS/MS search is implemented as an external search program called 'tandem.exe'. GPMW prepares all data and saves them in files on disk. Then Tandem is called with in-line parameters that specifies which files to load. X! Tandem then runs in a separate window. **Note:** while X! Tandem runs, you should not make changes to GPMW. Upon completion of the search, X! Tandem closes and a message is sent to GPMW which opens the result file created by X! Tandem.

The ms/ms search is accessed through the main menu Search | MS/MS search or by pressing the F5 key.

An example file has been included with the GPMW installation; please see section 8.14 for details.

### MS/MS data input

### 8.10

When you select the MS/MS search option, you are greeted with a window containing three tabs in the left-hand side:

**Input:** This is where you define your search: specify input mass file, search mass or database, enzyme used, output file and parameters.

**Output:** This is the primary output: List of all high-scoring proteins and all peptide hits.

**Hit:** On this tab you can analyze each individual hit on the 'Output' page.

The screenshot shows the X! Tandem search interface with the 'Input' tab selected. The window title is 'Search sequence - file1\_0b01782\_040209\_0924.xml'. The interface is divided into several sections:

- Data input file name:** (.mgf, .dta, .pkl) with an 'Open file' button.
- Select database:** No sequence/database found - cannot search.
- Enzyme:** Trypsin [KR][P] (dropdown menu).
- Output file name:** Output (text field).
- Search crap list:** (checkbox).
- Parameters:**
  - Parent, mono. error: -10.0/+10.0 ppm
  - Fragment, mono. error: 0.8 Daltons
  - Max. valid expect value: 0.010
  - Minimum ion count: 6
  - Scoring ions: Y A B
  - Spectrum
    - Minimum number peaks: 13
    - Maximum number peaks: 50
    - 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
- File:** Edit params (button).
- Fragment mass type:** Monoisotopic (radio button), Average (radio button).
- Variable modifications:**
  - Oxygen [M] 15.99
  - Methylation [DE] 14.02
  - Phospho [STY] 79.97
  - Thr\_ala [T] -30.01
  - Me-ester [DEST] 14.02
  - D-Succ [D] -18.01
  - Sodiated [DE] 21.98
  - Deamidation [DE] -1.01
  - Acetyl [K] 42.01
  - di-Methylation [K] 28.03
  - Methyl [K] 14.02
- Fixed modifications:**
  - Carbamidomethyl [C] 57.02
  - Carboxymethyl [C] 58.01
  - PyridylCys [C] 105.06
- Enable refinement:** (checkbox).
- Variable modifications:**
  - Oxygen [M] 15.99
  - Methylation [DE] 14.02
  - Phospho [STY] 79.97
  - Thr\_ala [T] -30.01
  - Me-ester [DEST] 14.02
  - D-Succ [D] -18.01
  - Sodiated [DE] 21.98
  - Deamidation [DE] -1.01
  - Acetyl [K] 42.01
  - di-Methylation [K] 28.03
  - Methyl [K] 14.02
- Max. e-value:** 0.100
- Semi cleavage:** (checkbox).
- Spectrum synthesis:** (checked checkbox).
- N-term acetylation:** (checkbox).
- Run Search:** (button).



The bottom of the window features a toolbar with icons for 'Open result file', 'Copy', 'Print', 'Display', 'Done', and 'Help'.

## 8 – Database mass search

### Performing a search:

- 1) Open your ms/ms peak list file (can be in mgf, dta or pkl format).
- 2) Specify sequence or database to search
- 3) Specify cleavage enzyme and output file name (a default name is created by GPMaw).
- 4) Check and set search parameters
- 5) Set variable and fixed modifications
- 6) Click the 'Run' button
- 7) X! Tandem opens in a command line window where you can follow the progress of the search.
- 8) When the search is done, GPMaw loads the result file and displays it on the 'Output' tab, see below.

### Open file

Pressing the Open file button, opens a standard dialog where you select your ms/ms data file. The file has to be in .mgf, .dta or .pkl format (appendix A.3). Once you have selected an mgf file, two buttons become visible to the right of the input field   Filter. Pressing the first button opens a window showing a graphical view of the input file (unsupported function) and the second button opens a filter window, enabling you to filter the input file (only works for .mgf files), please see below.

### Search sequence

By default GPMaw wants to analyze the topmost sequence window on the desktop. This is selected in the 'Search sequence' line below the data file input. If this option is used for analysis, the sequence will be saved in FastA format in the X!Tandem directory with the name 'searchSeq.fasta'.

If you want to search a database instead, you select this by pressing the 'Select database' button. **Note:** this database has to be in FastA format; a standard GPMaw file does not work.



**Hint:** If you want to make a FastA formatted file consisting of a limited number of sequences, an easy way is to load the relevant sequences in GPMaw and save them as a single FastA formatted file using the File | Export sequence command (Chapter 2.8).

### Enzyme:

In this box you select the enzyme used for the digestion of the protein(s). Currently only the most common enzymes are implemented.

If the enzyme of your choice is not on the list, you can select the last option '-user defined -'. This opens an input box where you can enter any most kinds of enzyme specificity.

The enzyme specificity is a formatted string that contains three elements: two specifiers and a separating vertical bar.

The specifiers are either accepting residues in square brackets [ ] or inhibiting residues in curly brackets { }.

## 8 – Database mass search

The character representing any residue is X.

This means that the cleavage of the enzyme trypsin is specified as [RK]{}{P} - cleavage takes place after Arg or Lys, but not if the following residue is Pro. Endoproteinase Asp-N is defined as [X]{}[D] - cleavage takes place after any residue but before Asp.

Multiple rules can be strung after each other separated by a comma e.g. [RK]{}{P},[X]{}[D] - cleavage following either a tryptic cleavage site or Asp-N cleavage site.

[X]{}[X] - cleavage after any residue (remember to increase 'missed cleavages').

### Output file name:

An output file name is suggested by GPMW based on the input file name combined with the current date and time (hours and minutes). The file name can be freely edited. Note that if you run multiple searches with the same input file but varying the parameters, you have to change this file name as GPMW does not check for an existing file of the same name.

This file is saved in the xtandem directory.

### Search crap list:

The cRAP list is a file, 'crap.fasta.pro', containing a number of advantageous proteins, primarily keratin, bsa and trypsin, maintained by the Global Proteome Machine Organization ([www.thegpm.org](http://www.thegpm.org)). These proteins are typical contaminants in proteolytic experiments and may in some cases obscure your expected data. When this option is checked, your data file will additionally be searched against these proteins. If the file 'crap.fasta.pro' is not present in the xtandem directory, the option will be greyed.



**Note:** The .pro format is a FastA file compiled with the fasta\_pro.exe program in order to make a more compact file that is faster to search by X!Tandem. The fasta\_pro.exe program is included on the disk and can be used to compact your own FastA files. The current cRAP file can be downloaded from <ftp://ftp.thegpm.org/fasta/cRAP> in FastA format. A list of proteins in the file can be found at <http://www.thegpm.org/crap/index.html>. Here you will also be able to download updated versions of the list.

### Parameters:

If you need to change the search parameters, you select the 'Edit params' button, which opens a dialog box where all parameters can be edited (see section 8.11 below). Alternatively, sets of parameters can be saved and re-loaded from the drop-down menu activated by the down-arrow on the button.

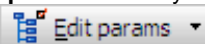
Select variable and fixed modifications from the two list boxes by clicking once on each modification to include in the search. You deselect a choice by clicking once again.

The **variable modifications** are loaded from the currently selected modification file (standard GPMW modification file), see Chapter 4.3 for details on how to construct and edit this file.

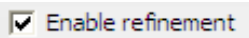
## 8 – Database mass search

The **Fixed modifications** is also a standard GPMW modification file, but with the name 'fixedmod.mod'. You can load and edit this file in the 'Edit modification dialog', Chapter 4.3. The first time you run the MS/MS search function, this file will be created automatically.

**Saving parameters:** If you click on the down-arrow on the 'Edit params'

button , you will get a menu that enables you save and load up to five sets of parameters. This is handy when using different instruments or experimental settings.

### Refinement

The refinement option  makes X! Tandem perform a second pass search through the identified protein hits. This makes it possible to define a second round of modifications, particularly useful if you search for multiple modifications on the same residue (e.g. methylation and acetylation of lysine residues).

In addition you can search for

**'Semi cleavage':** This option searches for peptides that fulfill cleavage requirements only at one end of the peptide. Trimming at the N- and C-terminally end of a protein can typically be found using this option.

**'Point mutations':** Searches for point mutations in the proteins already identified.

**'Spectrum synthesis':** If checked, the following bond cleavages are favored relative to other cleavages (e.g. results in a higher score): N-terminal to Pro and C-terminal to Asp, Glu, Asn, Gln, Val, Leu and Ile.

**'N-terminal acetylation':** searches for N-terminal acetylation.

Please note that the run time for 'refinements' is often much longer than the initial search. Please also note that the search for 'semi cleavage' and 'point mutations' can lead to spurious identifications, so you should be critical towards these identifications.

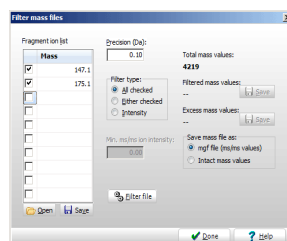


**Important:** The mass values used for searching is downloaded from GPMW, using the currently loaded mass file. This makes it imperative that you have the correct mass file loaded prior to running the search and that the SS button (oxidized/reduced Cys) is in the correct position. This means that if you have loaded a modified Cys table (i.e. acetamide) you should not also specify Cys as modified under 'fixed modifications'.

### Filter mass file

When you have specified an mgf file in the input field, the **'Filter'** button becomes visible. This option enables you to filter an mgf file according to specific fragment ions (reporter ions) or total fragment ion intensities.

You may enter mass values to search for in the table, and enable/disable individual items through the check-boxes. In the **'Filter type'** box you can select either **All checked**, all checked mass values have to be present to





## 8 – Database mass search

select a given ms/ms mass list, or **Either checked**, only one of the checked mass values have to be present to select the ms/ms mass list. Alternatively, you can select **Intensity** to filter the list based on total ion ms/ms intensity. In this case you have to enter a value in the **Min. ms/ms ion intensity** box.

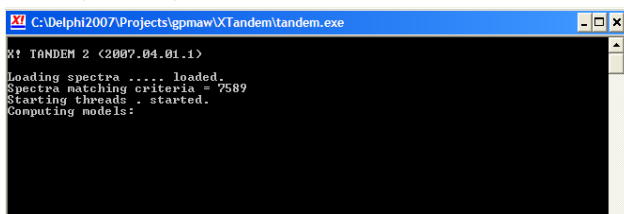
Press the **Filter file** button to perform the actual filtering.

After filtering, the number of filtered and excess mass values will be displayed, and you can save either as a file on disk in either mgf file format or just save the intact parent ion mass values depending on the setting of the **Save mass file as** box.

The mass values in the table can be saved and loaded from disk as simple text files (extension .txt) with a single mass value pr line.

### Run

When you press the '**Run**' button, the parameters are saved in a file called 'input.xml'. This file works in conjunction with the file called 'default\_input.xml' which contains all the available parameters for the X! Tandem search engine. The parameters in 'input.xml' override the corresponding values found in 'default\_input.xml'. A file called 'taxonomy.xml' is also saved, which allows for the use of databases with taxonomy information, but this function is not currently used by GPMaw.



```
C:\Delphi2007\Projects\gpmaw\XTandem\tandem.exe
X! TANDEM 2 <2007.04.01.1>
Loading spectra .... loaded.
Spectra matching criteria = 7589
Starting threads . started.
Computing models:
```

GPMaw then calls the XTandem program with the given parameters. The XTandem program opens in a separate command window, where you can follow the progress of the search. **Do not close this window unless the XTandem search encounters problems.**

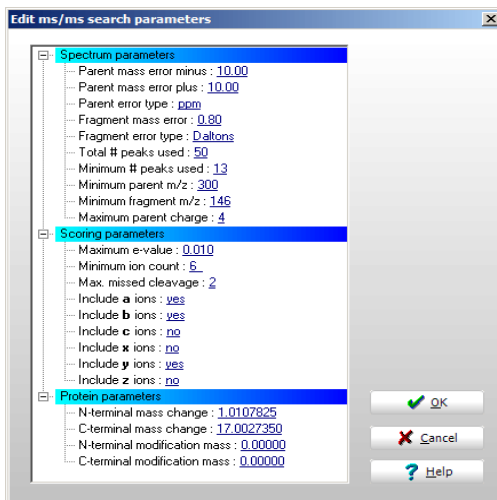
Upon completion of the search, the window will close automatically and the results are saved to a file on disk with the given 'output file name' and the extension '.xml'.

This file is then read by GPMaw, parsed and displayed on the 'Results' tab (see section 8.12 below). The saved file can at any time later be opened in GPMaw by selecting the 'Result file: Open' button at the bottom of the ms/ms search window.



**Hint:** The .xml result file can also be viewed on the Proteome Machine home page. Go to [www.thegpm.org](http://www.thegpm.org), select one of the search sites in the left-hand margin and on the search page, you can select 'view saved xml data' in order to see another representation of your data. This can also be used to view data that you ship to colleagues.

When you edit the parameters, the following dialog box is shown:



#### Parameters:

**Parent mass error:** This is the precision of the parent ion, and is set as a plus and minus error to accommodate the (few) mass spectrometers that have an asymmetric mass profile. The precision can be set as either ppm (parts per million) or Da. (Dalton).

**Fragment mass error:** The precision of the fragment ions. Can be set as ppm or Da.

**Total # peaks used:** Maximum number of peaks used for identification of the target peptide.

**Minimum # of peaks used:** Minimum number of peaks required for a spectrum to be considered.

**Minimum parent mz:** Minimum  $M + H^+$  required for a spectrum to be considered.

**Minimum fragment mz:** Minimum fragment  $m/z$  to be considered.

**Maximum parent charge:** Highest charge of parent ion to be considered.

**Max. e-value:** Highest e-value for peptides to be recorded in the output list.

**Minimum ion count:** Sets the minimum number of ions required for a peptide to be scored.

**Max. missed cleavages:** Largest number of missed cleavages considered for a peptide.

**Include ions:** Allow the checked ions to be used in scoring.

## 8 – Database mass search

**Refine search:** Controls whether the refinement module of XTandem is used. Setting of these parameters is not currently implemented in GPMW, but the user can set them manually in the 'default\_input.xml' file.

**N-terminal mass change:** Moiety added to the peptide N-terminus upon cleavage. Default is 1.0107875.

**C-terminal mass change:** Moiety added to the peptide C-terminus upon cleavage. Default is 17.0027350.

**N-term modif. mass:** Moiety added to the N-terminus of the protein. Default is 0.

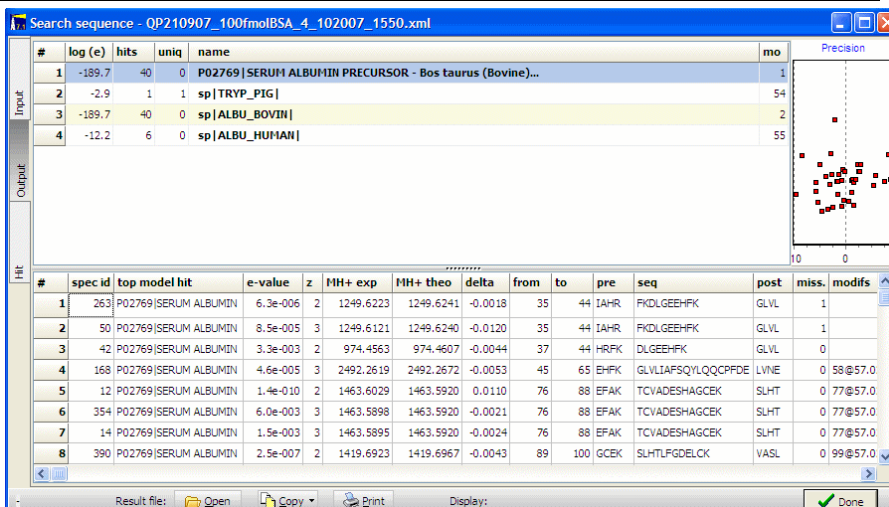
**C-term modif. mass:** Moiety added to the C-terminus of the protein. Default is 0.



**Note:** For more details on the X! Tandem parameters see the web site of *The Global Proteome Machine Organization*, <http://www.thegpm.org>, look in the API section.

## MS/MS search results

8.12



The output tab is divided into three panels:

Sequence hits are listed at the top. This is a list of all the models in the database that fit to the current search. Each hit is listed with log of e-value, number of peptide hits, unique hits and name of protein. In the above example, a test run on bovine albumin, the protein has been determined with a score of log e at -189.7, 40 hits, none of which are unique, as the same protein is part of the crap file, which of course have the same hits. In addition porcine trypsin has been found with a single hit, and human albumin with 6 hits, all of which are shared with bovine albumin.

To the right of the protein names are listed a model number, this can be correlated with the model numbers in the peptide list below.

## 8 – Database mass search

**Double-click** on a model name to view the peptide hit results on the 'Hit' tab.

A graph at the right-hand side of the window shows the precision of all the peptide hits of the currently selected protein model. The graph has the x-axis in ppm (autoscaled) and mass values along the y-axis (500 – 4000).

The bottom table shows all the peptide hits as they have scored during the search:

**Spec id:** Spectrum id, location in the input mass list.

**Top model hit:** Name of protein hit.

**e-value:** The score of the peptide hit.

**Z:** Charge of parent ion.

**MH+ exp:** Experimental mass.

**MH+ theo:** Theoretical mass value calculated based on the database hit.

**Delta:** Deviation between experimental and theoretical mass.

**From – To:** Peptide location in the model sequence.

**Pre:** Residues prior to the peptide hit.

**Seq:** Sequence of peptide hit.

**Post:** Residues (up to 4) following the peptide hit.

**Miss:** Number of missed cleavages in the peptide hit.

**Modifs:** Modifications found in the peptide hit. These are labeled as position in the sequence followed by '@' and finally the mass of the modification.

Multiple modifications can be listed after each other, all of which have been found in the same peptide hit.

**Models:** Number of the model(s) that have been allocated this peptide hit. In this case all of the first hits have model number 1 and 2 (bovine serum albumin, first and third in the model protein list above).

**Check box:** When copying the list to the clipboard, the menu option 'checked peptides only' will only copy those lines where this check box is checked. On the 'Hit' tab only the top hit for each peptide in the protein sequence will be selected when the 'unique' option has been selected.



**Note:** The results of ms/ms searches are always saved in .xml files in the search directory. You can at any point load a previous result file by clicking on the 'Result file: Open' button and select an older file.

Result file:  Open

If you perform multiple searches on the same input data file and you want to keep the intermediate results, please remember to change the output file name as GPMaw only sets the name when you load the input file.

### Analyzing individual hits

8.13

When you have double-clicked on a protein name on the output tab, GPMaw switches to the 'Hit' tab and shows the details of the peptides making up the hit.

## 8 – Database mass search

Input	Sequence	MKWVTFISLL	LLFSSAYSARG	VFRRDTHKSE	IAHRFKDLGE	EHFKGLULIA	FSQYLQQCPF	60
		DEHVKLUNEL	TEFAKTCUAD	ESHAGCEKSL	HTLFGDELCK	VASLRETYGD	MADCCCKQEP	120
		ERNECFLSHK	DDSPDLPKLK	PDPNTLCDEF	KADEKKFWGK	YLVEIARRHP	YFYAPELLYY	180
		ANKYNGUFQE	CCQAEDKGAC	LLPKIETMRE	KULASSARQR	LRCASIQKFG	ERALKAWSUA	240
		RLSQKFPKAE	FUEUTKLUTD	LTRVHKECCH	GDLLCADDPR	ADLAKYICDN	QDTISSKLKE	300
		CCDKPLLEKS	HCIAEVEKDA	IPENLPPLTA	DFAEDKDUCK	NYQEAKDAFL	GSFLVEYSRR	360
		HPEYAVSULL	RLAKEYEATL	EECCAKDDPH	ACYSTUFDKL	KHLUDEPQNL	IKQNCQDFEK	420
		LGEYGFQNAL	IURYTRKUPQ	USTPTLVEUS	RSLGKUGTRC	CTKPESERMP	CTEDVLSLIL	480
		NRLCULHEKT	PUSEKUTKCC	TESLUNRRPC	FSALTPDETY	UPKAFDEKLF	TFHADICTLP	540
		DTEKQIKQT	ALUELLKHKP	KATEEQLKTU	MENFUAFUDK	CCAADDKEAC	FAVEGPKLUU	600
Output	Grid	STQTALA						
		Coverage: 46.29% P02769   SERUM ALBUMIN PRECURSOR - Bos taurus (Bovine)						
		IAHR FKDLGEEHFK GLVL 35-44 1249.6223 Da -1 ppm						
Hit								

M/Z 1249.62 [2] 35-44 FKDLGEEHFK

M/Z 974.46 [2] 37-44 DLGEEHFK

M/Z 2492.26 [3] 45-65 GLVLIAFSQYLQQCF

120

100

80

yg

The Hit tab shows a three window layout:

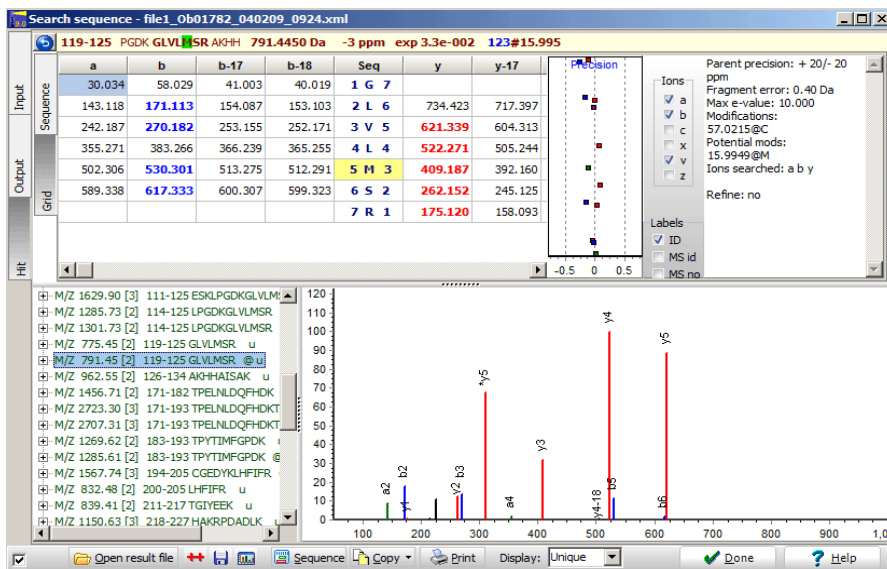
At the very top is a peptide information line which shows details on the currently selected peptide.

Below the peptide line is a two-tabbed window, which initially shows the protein sequence with the total coverage and each identified peptide highlighted in red.

At the bottom left is a list of all peptides belonging to this protein, and to the right is a window showing the corresponding ms/ms spectrum.

When you select a peptide in the table, the topmost tabbed list changes to show the corresponding fragment ions:

## 8 – Database mass search



The tabbed window at the top now changes to a grid, displaying the theoretical fragments that can be generated. Fragments that are found in the mass spectrum are highlighted and colored. Modified residues are listed with a colored background. Each fragment type has its own color. The same colors are used to display the precision and the ms/ms spectrum, bottom right.

To the right of the grid, is displayed the precision with which the mass spectrum fits the theoretical mass value.

Then a band is displayed with control options for fragment ions to display in the table, default are the ions used during the search. Below this control is another for the ms/ms labels: **ID** displays the ion type for the corresponding identified ion (e.g. y10, b7); **MS id** displays the mass values for identified ions; **MS no** displays the mass values for non-identified ions.

Finally is a table for the most important search parameters.

At the bottom left is a list of all peptides identified during the search. The actual peptides displayed depends on the setting of the 'Display' option in the bottom toolbar: **Display: Unique**. Three options are available: **Unique** only the highest scoring peptide of each kind is displayed. Different modifications count as different peptides, even same modification in different locations. **Modified** only modified peptides are displayed. **All** all peptides that fit the search criteria are shown.

In the peptide list each line shows the experimental mass of the parent ion, number of charges, location of the

M/Z 1463.60 [2] 76-88 TCVADESHAGCEK @  
 M/Z 1419.69 [2] 89-100 SLHTLFGDELCK @  
 ... ID 8 Theo 1419.70 Delta -0.0043 Da,  
 ... e-value 2.5e-007 mc=0  
 ... 99@57.022  
 M/Z 1494.53 [2] 106-117 ETYGDMADCCEK @

## 8 – Database mass search

peptide, the peptide sequence and whether the peptide is modified (@).

If you click on the '+' sign, the list expands with additional information for the given peptide: ID is location in the peptide table (output tab). Theoretical mass, difference between theoretical and experimental mass, e-value, missed cleavages (mc) and lastly the modifications found in the peptide (listed as location@mass change).

Whenever a peptide is selected, the middle line of the display is updated with peptide information:

GCEK SLHTLFGDELK VASL 89-100 1419.6923 Da -3 ppm 99#57.022

The information line shows: peptide position (from-to), sequence with four residues before and after (modified residues are highlighted), experimental mass, deviation from theoretical mass, e-value and modified residues in the format: location@mass of modification.


If you click on the left-most blue button, the display changes to show the location of the hit in the result table (#7) and the location of the spectrum in the input file list (ID 12).

76-88 EFAK TGADESHAGK SLHT #7 ID 12 77#57.022 86#57.022

The ms/ms graph at the bottom right shows the mass data of the peptide as listed in the input file. Identified peptide fragments are colored along the same scheme as the fragment table. The label of each peak is controlled in the gray band above. The graph can be zoomed by click-and-drag right and down, you un-zoom by click-and-drag left and up. You can scroll the display by right-click and drag left/right.

### Sequence and coverage map

When on the **Hit** page, you can open the sequence analyzed in a standard

sequence window by pressing the **sequence** button  **Sequence**. This will open a sequence window with all identified peptides as underlined residues, information on each peptide can be found in the left-hand information panel by mouse-over. In addition a coverage map displaying the hits will open (see chapter 9.6).

### Running the BSA example

8.14

The BSA example included with GPMaw is a short calibration run performed on an Orbitrap machine. It is present in the \gpmaw\bin\xtandem\ directory. Two files are included: The data file "BSA\_100fmol.pkl" and an example of a search "BSA\_100fmol\_120107\_1729.xml".

The ms/ms search has a lot of parameters to set, but once you have 'trimmed' the search parameters to your instruments/experiments, there will usually only be the input file, perhaps the output file, and the modifications to specify for each search.

## 8 – Database mass search

### Before you begin

Before you call the ms/ms search function, you load bovine serum albumin on the desktop – the accession number is P02769 if you need to download it from the Internet).


Download by typing number in toolbar window and click the 'Web' button (needs Internet access).



Make sure that the BSA sequence is the topmost window before proceeding.

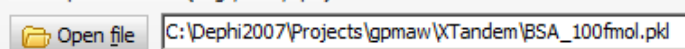
### Setting up the search

Begin by selecting 'MS/MS search' from the 'Search' menu, press the F5 key

or click the magnifier  in the toolbar, this opens the MS/MS search window with the 'Input' tab selected.

Start by selecting the 'Open file' button. In the 'Open MS file' dialog you navigate to \gpmaw\bin\xtandem\ and select the "BAS\_100fmol.pkl" file.

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



As you already have the BSA sequence opened as the topmost window on the desktop it will be selected by the ms/ms search:

Search sequence: Serum albumin precursor (Allergen Bos d 6) (BSA), - Bos taurus (Bovine).

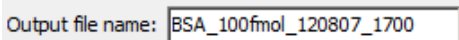
If you will rather search a file, you need to have the sequence(s) saved in FastA format. This is also the way to go if you need to search multiple sequences.

You now need to select the enzyme used to cleave the protein. Select from the drop-down box labeled 'Enzyme':



If the enzyme of your choice is not available on the list, select the last option: '- user defined -'. This will open a dialog box where you can specify your enzyme (for nomenclature see section 8.11 above).

Then enter the output (result) file name. By default gpmaw selects a name consisting of the input file name with date and time added.



The field can be freely edited. Note that if you perform multiple searches on the same input file without closing the ms/ms window in between, you have to change the name manually, as gpmaw will overwrite without asking permission.

You now have to select the parameters to use for the search. Please see section 8.11 for details on the various settings.

The most important settings for this search are +/- 10 ppm as the parent ion error and +/- 0.8 Da as the fragment error.



## 8 – Database mass search


Click on the 'Edit params' button to set the parameters, you should end with settings similar to this:

Parameters:

Parent, mono. error: -10.0/+10.0 ppm  
Fragment, mono. error: 0.8 Daltons  
Max. valid expect value: 0.010  
Minimum ion count: 4  
Scoring ions: Y A B

Spectrum  
Minimum number peaks: 13  
Maximum number peaks: 51  
Maximum parent charge: 4  
Minimum parent m+h: 300  
Minimum fragment m/z: 146

Protein  
Max. missed cleavage: 2  
Cleavage N-term mass change: 1.0107825  
Cleavage C-term mass change: 17.0027350  
N-term residue mod. mass: 0.00000  
C-term residue mod. mass: 0.00000

File:  Edit params ▼


Note that by clicking the down-arrow on the 'Edit params' button, you can save and retrieve up to five different settings to use on different instruments/experiments.

Now select the variable modifications:

Variable modifications [1]

Oxygen [FMPY]	15.99
Methylatio [DE]	14.02
Phospho [STY]	79.97

Each modification can be selected / deselected by clicking with the mouse. If you need to load a different modification file or edit one of the modifications,

click on the 'S' button  below the box. Oxygenation of Met is the usual choice, but a number of others like deamidation of Asn and Gln are also typically seen for most runs.

For the fixed modifications you select carbamidomethylation of cysteines (Cys has been alkylated with iodoacetamide).

Fixed modifications [1]

Carbamidomethyl [C]	57.02
Carboxymethyl [C]	58.01

You can select to the perform 'refinements' (second pass search) on your data, please see section 8.11 for details.

Now you can perform the search by clicking the Run button



## 8 – Database mass search

### The run

If parameters are correctly set and XTandem is correctly installed, a command line window will open, showing the progress of the search.

```
X! C:\PROGRA~1\gpmaw\bin\XTandem\tandem.exe

X! TANDEM 2 <2007.04.01.1>

Loading spectra . loaded.
Spectra matching criteria = 361
Starting threads . started.
Computing models:
      t

      sequences modelled = 0 ks
Model refinement:

Creating report:
      initial calculations ..... done.
      sorting
```

Without the refinement option set, this search should take less than two seconds.

### Results

When the search is done, XTandem saves the results to disk and closes. GPMaw is informed, loads the result file and displays the contents on the 'Output' page.

As we only searched a single protein, there will be (or should be) only one hit in the list of proteins found:

#	log (e)	hits	uniq	name
1	-172.6	39	39	P02769   Serum albumin precursor (A

Below the protein list is a list of all the peptides identified in the search which fall inside the parameters set (particularly below the e-value of 0.01).

#	spec id	top model hit	e-value	z	MH+ exp	MH+ theo	delta	from	to	pre	seq
1	50	P02769 Serum albumin	8.5e-005	3	1249.6121	1249.6240	-0.0120	35	44	IAHR	FKDLGEI
2	263	P02769 Serum albumin	6.3e-006	2	1249.6223	1249.6241	-0.0018	35	44	IAHR	FKDLGEI
3	42	P02769 Serum albumin	3.3e-003	2	974.4563	974.4607	-0.0044	37	44	HRFK	DLGEEH

For details on the peptide parameters, please see section 8.12.

Now we want to see the details of the search, so we double-click on the Serum albumin precursor name in the protein list.

### Hit (protein results)

This opens the protein details tab (called 'Hit') which is a quite crowded display:

The top left display shows the protein coverage (i.e. the parts of the protein that has been identified by peptides in the ms/ms run):

## 8 – Database mass search

MKWUTFISLL LLFSSAYSRG UFRDTHKSE II  
**DEHUKLUNEL** TEF**AKTCUAD** **ESHAGCEKSL** H'  
 ERNECFLSHK **DDSPDLPLK** **PDPNTLCDEF** KI  
 ANKYNGUFQE **CCQAEDKGAC** LLPKIETMRE KI

Below on the left is a table of all unique peptides identified:


⊞ M/Z 1249.62 [2] 35-44 FKDLGEEHFK  
 ⊞ M/Z 974.46 [2] 37-44 DLGEEHFK  
 ⊞ M/Z 2492.26 [3] 45-65 GLVLIQSQYLQQCF  
 ⊞ M/Z 1463.60 [2] 76-88 TCVADESHAGCEK

Click on one of the peptides will switch the coverage display to show the fragment ions of the identified peptides.

a	b	b-17	b-18	Seq	y	y-17
101.071	129.066	112.040	111.056	<b>1 Q 8</b>		
215.114	<b>243.109</b>	<b>226.083</b>	225.099	<b>2 N 7</b>	940.383	923.35
375.145	403.140	<b>386.113</b>	385.129	<b>3 C 6</b>	<b>826.341</b>	<b>809.31</b>
490.172	518.167	<b>501.140</b>	500.156	<b>4 D 5</b>	<b>666.310</b>	<b>649.28</b>

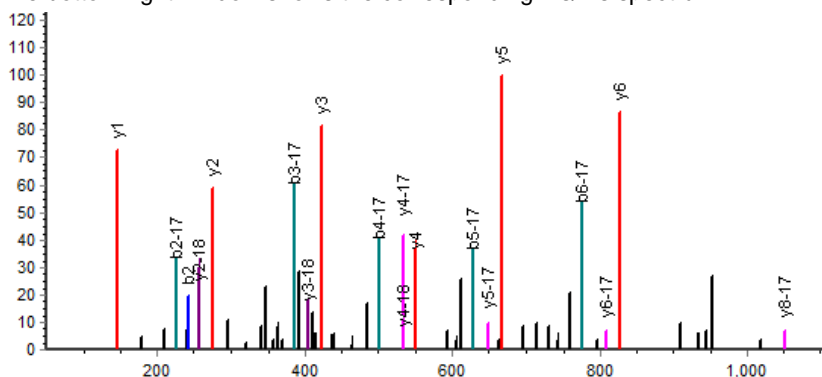
This table shows all the possible fragments, with the ones identified in the spectrum shown highlighted in a specific color for each fragment type. The identified residues (sequence) are displayed in the 'Seq' column. If a residue is identified as modified, is shown with a yellow background.

The top line in the window shows the identified peptide sequence with modified residues highlighted and the most important parameters:

 **413-420** NLIK Q**N**DQFEK LG**EY** **1051.4171 Da** **-1 ppm** **exp 3.9e-003** **415#57.022**

By clicking on the left-hand blue button, the display changes to show the peptide ID parameters (location in the table on the 'Output' tab and location of the ms/ms spectrum in the input file).


The bottom right window shows the corresponding ms/ms spectrum



The colors of the fragments correspond to the colors in the table above. The label of the peaks can be set in the control bar to the right of the table, which also controls the table columns to display. The top central graph shows the deviation of the ms/ms fragments from the theoretical mass values, and the right-hand list shows the search parameters.

## 8 – Database mass search

The results are saved on the computer as an **xml** file in the same directory as the input mass file. This file can at any point be recalled (click the bottom

open button ) and viewed again. You can also go to the Global Proteome Machine ([www.thegpm.org](http://www.thegpm.org)) and load the file for an alternative view of your data.

You can get a coverage map of your hit by pressing the **Sequence** button



at the bottom of the display. This opens a sequence window containing the selected protein with the coverage displayed as underlined residues (see chapter 3.4). From this a window containing an interactive coverage map (see chapter 9.6) opens.

On the disk is a file 'BSA\_100fmol\_120107\_1729.xml' which contains a search performed by Lighthouse data. You can compare your results with this file.

### References:

TANDEM: matching proteins with mass spectra, Robertson Craig and Ronald C. Beavis, *Bioinformatics*, 2004, 20, 1466-7.

A Method for Reducing the Time Required to Match Protein Sequences with Tandem Mass Spectra, Robertson Craig and Ronald C. Beavis; *Rapid Commun. Mass Spectrom.*, 2003, 17: 2310-2316.

## Cleavages

Cleaving a protein into peptides using specific or nonspecific cleavage methods.

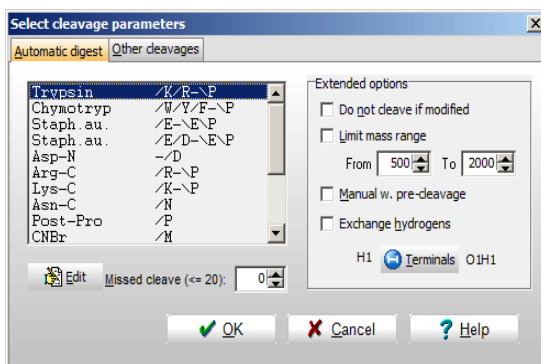
This chapter describes how to use GPMW to simulate cleavages of a protein using specific proteases or chemical cleavage methods. It also describes how the resulting peptide list can be viewed in various modes and can be used for generating simulated HPLC chromatograms and mass spectra.

If you need to mass analyze a protein digest, check out the 'Cleavage analysis' section, 9.3. The results of the cleavage (the peptide window) is described in section 9.4)

### Automatic digest

### 9.1

The **Cleavage | Automatic digest** command opens the 'Select cleavage parameters' on the Automatic digest page. The other page of the dialog 'Other cleavage' is covered in the next section.



In the automatic digest page, you can select a pre-defined enzyme to cleave your protein. By default, 10 enzymes are defined in the list of enzymes. All of the 'enzymes' can be edited, and there is room for five additional enzymes. Note that the CNBr (/M) does not yield the correct masses for cyanogen bromide cleavage (use 'Other cleavage' for this purpose, see below).



**Tip:** As the most commonly used enzyme is trypsin, this is placed at the top of the list. If you preferentially use another enzyme, you should move this to the first entry in the list.

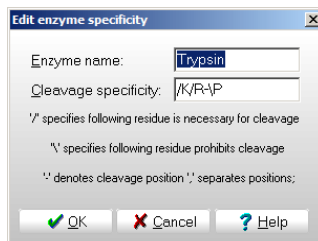
Each enzyme is listed with name and specificity (see below).

## 9 – Cleavages / Coverage map

As enzymes do not always cleave at all potential cleavage sites, it is possible to specify '**Partials**' (missed cleavages). A partials level of 1 means that the resulting peptide list, in addition to all completely cleaved peptides, will contain all peptides that contain a single potential cleavage site (e.g. in addition to HITLK and SEAQR the list will also contain HITLKSEAQR). A partials level of 2 means that all peptides containing at most two potential cleavage sites will be shown. A peptide containing potential cleavage points will be indicated in the peptide list with a '\*' after the peptide number.

The '**Edit**' button enables you to edit the currently selected enzyme line. For each line in the enzyme list, you specify a name (e.g. chymotrypsin) and cleavage specifications (see 'Specifying enzyme cleavage definitions' below).

The results of an automatic digest are shown in the peptide daughter window (below, 9.4).



### Extended options

The peptides resulting from a cleavage can be modified through the '**Extended options**'. The different options can be combined.

**Do not cleave if modified:** If checked, automatic cleavage will not take place if one of the residues that are part of the cleavage specification is modified. E.g. if you have specified a lysine residue to be hydroxylated, trypsin will not cleave.

**Limit mass range:** If checked, only peptides having a mass in the range specified by the two edit number boxes will be displayed in the peptide window. If this function is activated, the number of peptides below, in, and above the range will be shown in an extra information pane below the toolbar in the peptide window.

**Manual w. pre-cleavage:** Enables you to modify the automatic cleavage. After clicking '**OK**', all cleavage points will be shown inverted in the sequence window. Clicking on the residue with the mouse can now toggle extra as well as existing cleavage points. Right-clicking the mouse in the window terminates data input.



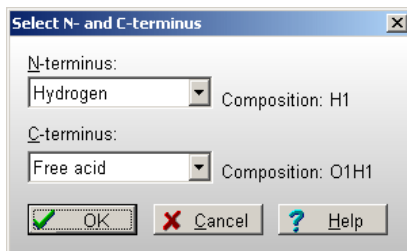
**Note:** As it is not possible to scroll the display when specifying 'manual' cleavage points you should make certain that the whole sequence is displayed in the sequence window before selecting this option.

**Exchange hydrogens:** The mass of all exchangeable hydrogens will be changed to the mass of deuterium.

## 9 – Cleavages / Coverage map

**Terminals:** The 'Terminals' button enables you to specify the N- and C- terminals of the generated peptides.

Pressing the button opens a dual drop-down dialog box. The selections are the same as for specifying the termini of the intact protein (for setting up terminals, see 'Edit' Ch. 4.1).




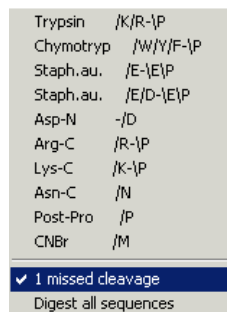
The default is hydrogen, H (mass 1 Da) for the N-terminus and free acid, OH (mass 17 Da) for the C-terminus.

If you define a new N-terminus, it will be part of all peptides in the resulting peptide table. If you perform ms/ms of such a peptide, the modification is 'carried on' to the ms/ms window.

### QuickDigest

The 'Automatic digest' can also be accessed from the sequence window by clicking on the 'Digest'

button . The down arrow next to the 'scissor' button opens a menu for the '**QuickDigest**'. This menu lists the top eight enzymes from the automatic digest option. Selecting one of these enzymes performs a quick 1-click digest. Please note that the 'QuickDigest' performs a 'straight' digest with no extended options.



Two options are listed at the bottom of the QuickDigest menu:

**1 missed cleavage:** Checking this will give you one missed cleavage in the resulting peptide window. Note that this option is persistent, i.e. the setting will be remembered between sessions.

**Digest all sequences:** If you check this menu option, all sequences opened on the desktop will be digested, i.e. generate a peptide window. Note that this option is not persistent, but will have to be selected before each operation on all sequence windows.

### Specifying enzyme cleavage definitions

Enzyme specificity is defined by the following symbols:

- / following residue is necessary for cleavage
- \ following residue prohibits cleavage
- cleavage position
- , separates multiple residues in the same position
- ; separator for individual cleavage specifications. This enables you to combine two or more enzymes within the same specification (e.g. if you digest with trypsin and Endoproteinase Asp-N the definition will be /R/K-

## 9 – Cleavages / Coverage map

\P;-/D). If you want to specify overlapping peptides for one enzyme and not another, you specify the required overlap level for the combined specification. When you have the peptide list you can remove the overlapping peptides for the 'clean' cleaving enzyme(s) through the local menu command **'Remove partials'** (see 'Pop-up menu' under 'Peptide window' below)

If no dash ('-') is present in the specifications, cleavage takes place after the last residue. Up to 6 positions can be specified. You may combine multiple cleavages into a single line (see above), the limit on the line is then 32 characters.

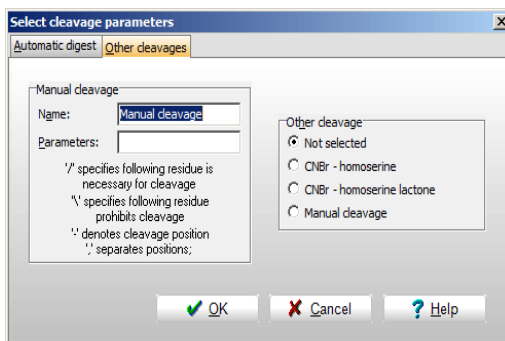
### Examples:

Trypsin:	Cleavage takes place after Arg or Lys, but not if the following residue is Pro. <b>/R,K-IP</b>
Chymotrypsin:	Cleavage takes place after Trp, Phe, or Tyr, but not if the following residue is Pro. <b>/W,F,Y-IP</b>
Endoproteinase Asp-N	Cleavage takes place before Asp. <b>-/D</b>
Acidic cleavage	Cleavage after Asp <b>or</b> before Ser or Thr. <b>/D;-/S/T</b>

## Other cleavage

9.2

### Manual digest



The manual digest option functions are identical to 'auto digest' except that you enter the enzyme name and cleavage specificity manually (see how to specify cleavages above).

The result of manual digest is, like automatic digest, displayed in a peptide window (see below).



## 9 – Cleavages / Coverage map

### Other digest

The other digest page lists a couple of cleavage options that cannot easily be defined using the standard cleavage nomenclature.

**CNBr:** The cyanogen bromide cleavage is a chemical cleavage that cleaves after Met, but modifies the C-terminal methionine residue into either homoserine or homoserine lactone depending on cleavage conditions. Apart from the homoserine/lactone formation the cleavage is similar to automatic digest and manual digest described above.



**Hint:** If you need to obtain information on overlapping peptides in a CNBr digest you have to perform an automatic digest (using /M- as cleavage parameter) and modify the C-terminus according to homoserine or homoserine lactone formation.

**Manual cleavage:** The manual cleavage enables you to specify all cleavage points yourself. After clicking '**OK**' you have to click on all residues in the sequence window **preceding** the bonds to be cleaved. The residue clicked on will be shown in inverted color to indicate that cleavage will take place.

Cleavage positions can be toggled both on and off.

You terminate cleavage definition input by clicking the right mouse button inside the sequence window. The resulting peptide window will then open.



**Hint:** You can modify an automatic digest by checking the 'Manual with pre-cleavage' option in the extended section of automatic digest (see above).

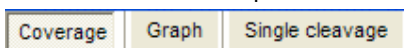
### Cleavage analysis

9.3

The **Cleavage analysis** command enables you to obtain a quick overview of the results of enzymatic or chemical cleavages.

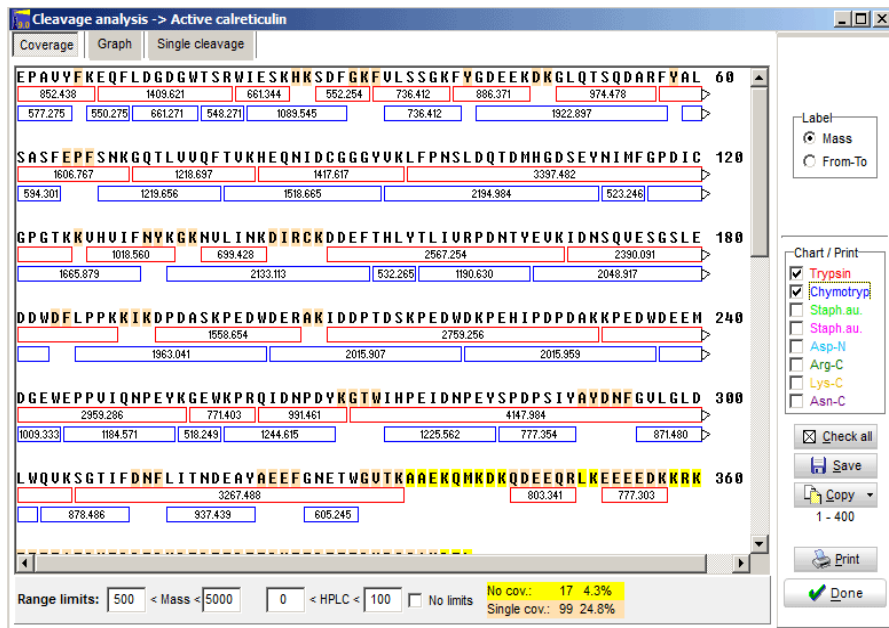
The window is divided into a large left-hand window showing the coverage analysis, and a small right-hand command bar giving you most of the available options.

There are three views/options available:



## 9 – Cleavages / Coverage map

### Coverage:



This is the default option with most option and information.

The left-hand window displays the protein sequence with 60 residues pr. line. Below each sequence line are the peptides created by the options in the right-hand panel and limitations defined in the bottom panel.

The **caption** in each peptide box can in the right-hand Label panel be set to either 'Mass' or 'From-to' (number of first and last residue in the peptide).

You select an enzyme to use for digestion in the right-hand panel by setting a check-mark with the mouse cursor. You may select up to all 8 enzymes. For each enzyme selected, the resulting peptides are displayed below the sequence lines in the same color as the enzymes are written in.



**Note:** The eight enzymes listed are the topmost eight 'enzymes' from the automatic digest list. You may edit the digest list to reflect your cleavage preferences (see section 9.1).

The colors cannot be redefined by the user.

The buttons below the enzyme selection box enables you to:

Select (**check**) all enzymes with a single click.

**Save** the cleavage analysis to disk – this makes it possible to load it into the 'Coverage analysis' windows for later analysis.

**Copy to clipboard.** The graph is copied in Windows metafile format, which enables the picture to be re-scaled without loss of quality after pasting into

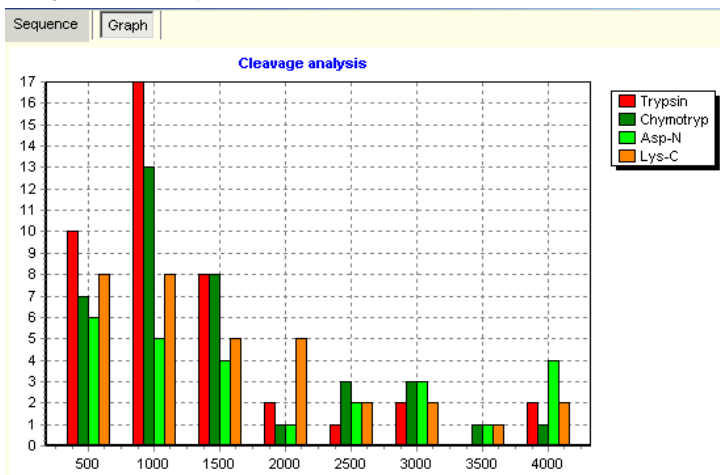
## 9 – Cleavages / Coverage map

other applications like Word or PowerPoint. By clicking on the down-arrow you can select a range to be copied to clipboard.

**Print.** Print the coverage map.

### Graph:

If you click on the 'Graph' tab at the top of the dialog box the display changes to a graphic display of the peptide distribution.



Each enzyme is presented in each own color as shown in the top right-hand box. The height of each bar presents the number of peptides in each mass segment giving a quick view of the size distribution of peptides. The last section '4000' counts all peptides larger than 3500 Da.

The enzyme cleavages to display in the chart are selected in the right-hand panel in the 'Chart/print' section, just like for 'Coverage'.

Several graph display are available and can be selected by right-click in the chart to open the pop-up menu.

Display options:

*Side by side:* Default, each mass segment shows all peptide counts.

*Side all:* Peptide counts are separated into each enzyme cleavage.

*Stacked:* The bars are stacked above each other, the height representing number of peptides.

*Stacked 100:* Similar to 'stacked' except each bar is scaled to 100%. This option shows the relative distribution in each segment.

The chart can be copied to the clipboard both in bitmap and in vector format (for best quality) through the pop-up menu or the command panel.

Before **printing** you have to select all the cleavage methods you want to analyze by checking the appropriate 'enzyme' check boxes above the '**Print**' button. The printout will show the graph at the top followed by all the peptide parameters on the left-hand side of the page with the protein sequence to the

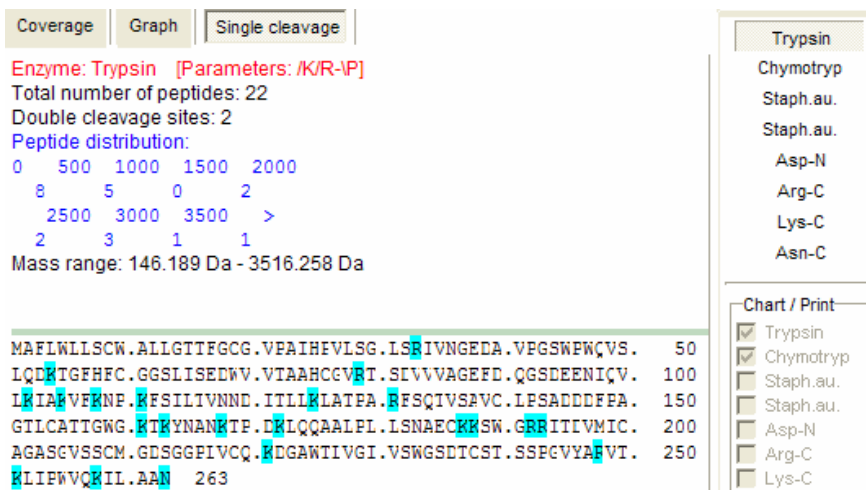
## 9 – Cleavages / Coverage map

right of the appropriate cleavage statistics. This makes for a compact presentation of all relevant cleavages.

The graph printed will be in the same format as displayed on screen.

The 'Save to disk' button is not available from the 'Graph' page.

### Single cleavage:



On the 'Single cleavage' page you analyze each enzyme cleavage individually.

In the right-hand panel, you cannot choose multiple enzymes, and the 'Chart/Print' panel has been disabled. Instead you now have eight buttons, one for each enzyme. When depressed, the corresponding cleavage will be displayed in the left-hand window.

The enzyme name and cleavage parameters will be in red, followed by the number of peptides (assuming complete cleavage) and number of double cleavage sites (e.g. two consecutive basic residues in a tryptic digest). The reason for listing the number of double cleavage sites is that many endoproteases are less active towards terminal residues leading to incomplete digests and more complex peptide mixtures. Below is listed the distribution of peptide masses divided into 500 Da ranges up to 3500 Da. The number of peptides with a mass larger than 3500 Da is listed between the '3500' and '>'. Finally the mass of the smallest and largest peptide is listed (the peptide mass range).

The bottom part displays the protein sequence with all cleavage points highlighted. Note that cleavage takes place **after** each highlighted residue. The last residue will always be highlighted as the termination of the sequence is counted as a cleavage point. It is not possible to alter the format of this display (e.g. change to a 3-letter display).

## 9 – Cleavages / Coverage map



**Note:** The cleavages performed in the Cleavage analysis are 'straight' cleavages. This means that you cannot specify overlapping sequences, modified terminals etc. For this you have to perform a 'normal' automatic cleavage as detailed in chapter 9.1.

### Printing:

When you print the graph and Single cleavage you will get a combination of the two. At the top will be a graph showing the results of all the checked enzyme digestions, and below will be the single cleavages of each enzyme including the sequence with highlighted cleavage positions.

### Peptide window

9.4

[1] Trypsin -> Myoglobin - Equus caballus (Horse), and

Mo. S 1/2 Alt i Low MS MS Trypsin [K/R-P] - p0 Seq. sort Sync. windows

Num	From-To	MH+	HPLC	Mass	pI	Add. Mass	Sequence
20	146-147	310.1761	5.58	309.1689	8.84	351.1809	YK
6	48- 50	397.2558	7.57	396.2485	9.25	438.2605	HLK
4	43- 45	409.2082	8.14	408.2009	5.69	450.2129	FDK
15	99-102	470.3337	13.97	469.3264	9.25	511.3384	IPFK
19	140-145	631.3410	5.66	630.3337	5.49	672.3457	NDIAAK
21	148-153	650.3144	15.78	649.3071	3.85	691.3191	ELGFQ
8	57- 62	662.3355	8.54	661.3283	4.13	703.3403	ASEDLK
7	51- 56	708.3233	7.01	707.3160	4.54	749.3280	TEAEMK
18	134-139	748.4352	18.00	747.4279	6.36	789.4399	ALELFR
3	32- 42	1271.6630	18.70	1270.6558	5.53	1312.6678	LFTGHFETLEK
10	64- 77	1378.8417	21.08	1377.8344	9.25	1419.8464	HGTVVLTALGGILK
17	119-133	1502.6693	15.69	1501.6620	5.06	1543.6740	HPGDFGADAQGMATK
2	17- 31	1606.8548	19.57	1605.8475	4.58	1647.8595	VEADTAGHGQEVILR
1	1- 16	1815.9024	23.46	1814.8952	4.13	1856.9072	GLSDGEWQQVLNVWGK
13	80- 96	1853.9617	16.95	1852.9544	7.46	1894.9664	GHHEAELKPLAQSHATK
16	103-118	1885.0218	26.60	1884.0145	6.19	1926.0265	YLEFTISDAITHVLHRSK

The peptide window shows the list of peptides that is the result of one of the enzymatic or chemical cleavages described above. The window is a daughter window that is linked to its parent sequence window so the peptide window will close when the parent sequence window is closed.



**Note:** Up to three peptide windows derived from the same sequence window can be open simultaneously. Subsequent peptide windows will all replace window number one. The number of the peptide window is displayed as the first item in the title bar in sharp brackets (e.g. [1]).

The initial display of the peptide list is determined by the 'Setup peptide parameters', Chapter 5.2. Several of the display parameters can be turned on and off, either through the tool bar or by right-clicking the mouse to get to the pop-up menu.



**Hint:** The amino acid residues are shown colored if they are colored in the parent sequence window. This means that if you want to color residues after you have created your peptide window, go back to the parent window and create the colored residues you want before returning to the peptide window. If the colors are not displayed immediately, force a repaint by minimizing and restoring the window.

## 9 – Cleavages / Coverage map

The peptide list shows the peptides generated in a tabular fashion, typically with a number of physiochemical properties as shown above. The **'Alt'** button switches to an alternate table that can be set up different from the primary. Typically one is set up with various physical chemical properties and the other as multiply charged ions (below) (see also chapter 5.2). Independent of the setup chosen, the first column lists the peptide number (in the linear sequence) and the last column lists the amino acid sequence (in either 1- or 3-letter code). The actual parameters shown will depend on the setup. Up to six columns (parameters) can be shown simultaneously in addition to the peptide number (always first) and the peptide sequence (always last). The currently supported peptide parameters are:

**Mass:** Molecular mass, M (e.g. non-charged).

**M-H to M5H-:** One or two negative charges (e.g. (M-H)<sup>-</sup> and (M-2H)<sup>5-</sup>).

**M+H to M8H+:** One to eight positive charges (e.g. (M+H)<sup>+</sup> to (M+8H)<sup>8+</sup>).

**From-To:** First and last residue in the displayed peptide.

**HPLC:** HPLC retention index.

**Ch:** Net charge at the pH defined in Setup (Ch. 5.2).

**pI:** Theoretical pI of the peptide calculated with the table defined in Setup (Ch. 5.1).

**B&B:** Bull and Breese index (hydrophobicity index).

**Add. mass:** Additional mass. Masses in this column have the mass defined in Setup (Ch. 5.2) added to their peptide mass M. Note that the addition is to M not to (M+H)<sup>+</sup>.

**Alt. mass:** Mass values in this column is calculated by the alternate mass file as defined in Setup (Ch. 5.2). Note that the settings here affects both the charge and the terminals.

**Av/Mo:** This column shows the opposite mass type as displayed by the (M+H)<sup>+</sup> column. E.g. if the window is set to monoisotopic display, this column will show the average (M+H)<sup>+</sup> value.

You can easily change the layout of the various columns. Right-click on the header button in question and a pop-up menu will show all the display options with a check-mark by the current selection. Select another column display type, and the display will be updated to reflect the new selection.

HPLC	Ch	pI	Av/Mo
14.46	-0.	M2H+	18.6493
19.35	0.	M3H+	94.6566
12.64	-2.	Mass	52.5046
18.70	-1.	MH+	72.4429
9.92	0.	From-To	72.4089
12.19	-0.	HPLC	87.2846
8.05	-0.	B&B	35.1471
9.01	-0.	✓ Ch	07.0934
9.01	-1.	pI	42.0367
7.99	-1.	M-H	06.9730
8.54	-0.	M4H+	90.8912
18.00	-0.	M2H-	48.8996
10.66	1.	Add.Mass	21.9209
7.01	-1.	Alt.M5	08.8100
13.97	0.	M5H+	84.8124
13.59	1.	M6H+	72.8478
8.54	-1.	M7H+	62.7172
16.32	-0.	M8H+	20.7257
		Av/Mo	



**Note:** The number of displayed columns will not be changed (maximum is 8 with the first and last column fixed at 'Number' and 'Sequence'). To change the number of columns you still have to go through Setup (Ch. 5.2).

## 9 – Cleavages / Coverage map

For details see Chapter 5.2.

Num	From-To	MH+	M2H+	Sequence
1	1- 16	1836.05	918.53	AGSYLLEELFEGHLEK
2	17- 29	1659.84	830.42	ECWEEICVYEEAR
3	30- 43	1818.85	909.93	EVFEDDETTDEFWR
4	44- 67	2490.78	1245.89	TYMGGSPCASQPCLNNGSCQDSIR
5	68- 93	2748.07	1374.54	GYACTCAPGYEGPNCFAESECPLR
6	94-114	2338.63	1169.82	LDGCQHFCYGPPESTYTCSCAR
7	115-117	341.39	171.20	GHV

The header for each column acts as button. When a button is pressed, the corresponding column will be sorted, the first time in ascending order, when pressed again, the sorting will be in descending order.

The header of the sorted column will be displayed in red.

Num	From-To
1	1- 16
39 <sup>1</sup>	1- 29
76 <sup>2</sup>	1- 43
2	17- 29
40 <sup>1</sup>	17- 43
77 <sup>2</sup>	17- 67
3	30- 43

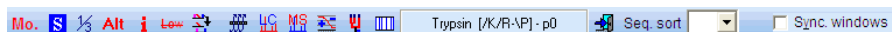
If you have specified overlapping peptides in the 'Select cleavage parameters' box (chapter 9.1) overlapping peptides will be shown with a superscript after the number indicating the number of overlaps (= number of missed cleavages) present in the peptide.

The commands available for peptide windows are accessed through the toolbar, the **Peptide list** menu, or the pop-up menu. The commands are listed below with a short description first, followed by a more detailed description of the individual functions.



**Note:** The commands have different scopes. Some commands are for the display only (e.g. 1/3 code toggle), some functions work on the peptide list as a whole (e.g. HPLC chromatogram) and some commands are for the individual peptide (e.g. ms/ms fragmentation).

### The toolbar



**Average/mono mass** toggle. Default is set in the setup dialog (chapter 5.2).



**Setup:** Opens the 'System setup' on the Setup peptide parameters page, see Chapter 5.2.



**1/3:** Toggle between 1- and 3-letter code.



**Alternate display:** Toggles between normal peptide list and alternate list. The actual columns displayed are set in the Setup peptide parameters (Chapter 5.2). The available columns are mass, singly to quadruply charged positive ion, singly and doubly charged negative ion, from-to, pI, HPLC index, Bull & Breese index and charge.



**Info:** Detailed peptide information window, see below.

## 9 – Cleavages / Coverage map



**Low mass filter:** When pressed, the low mass peptides are hidden in the display. The low mass filter limit can be set in System setup (chapter 5.2). This option is very handy when viewing mass spectrometric peptide maps where you very often do not see low mass ions.



**Partial modifications:** When the button is depressed, selection of a modified peptide will open a frame at the bottom of the window with the unmodified peptide and a list of all different combination of modifications. Please note that only mass values will be shown for the different version as GPMW does not calculate the pI, retention index etc for modifications.

**Note:** If you **print** the peptide list (and make sure the 'Print partial modifications' option is checked) you will get a list of all partial modification listed after the standard peptide list.



**Ms/ms:** Displays ms/ms fragmentation pattern of the selected peptide, see Chapter 10.1.



**HPLC chromatogram:** Simulated HPLC chromatogram, see below.



**Mass spectrum:** Simulated MS spectrum, see below.



**Charge vs pH:** Displays a graph of the charge of the peptide at all pH between 0 and 14. See below.



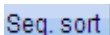
**View/Search N-glycan structures:** Use the selected peptide to search for N-linked glycan structures. See chapter 7.4 for details.



**pI strip:** Displays a window with a simulated pI strip (isoelectric focusing) of all peptides.

Staph.au. [/E/D-\E\P] - p0

The status panel shows the cleavage agent, the cleavage parameters and the partials level or number of missed cleavages (e.g. p0 = no overlapping peptides; p1 = fully cleaved peptides + peptides containing 1 potential cleavage site).



The sequence sort enables you to sort the peptide list according to highlighted sequences – select a highlight color in the drop-down box, and then click on the 'Sequence' header. Normally when you click on the 'Sequence' header in the table, the sequences will be sorted according to characters. When you have selected a color in the '**Seq. sort**' box, sorting will take place based on the selected color.



## 9 – Cleavages / Coverage map



**Close:** Close current peptide window. Also closes derived daughter windows like simulated HPLC chromatogram. Does not close the sequence window.



**Synchronize**

**Synchronize windows:** When this box is checked the peptide window will synchronize with the 'Sequence window' so selection of a peptide in the peptide window will result in the underlining of the corresponding sequence in the sequence window. Other windows like **ms/ms fragmentation** and **charge vs. pH** are also updated whenever the focus changes between peptides (in 'normal' mode you have to select the corresponding command in order to update the windows).



**Hint:** If you have a sequence, a peptide, an ms/ms fragmentation and a charge vs. pH window open in GPMW you can select the **Window|Tile** command to have all related windows tiled optimally in the main GPMW window. If you then check the '**Synchronize windows**' in the peptide window, all windows will be updated whenever the focus changes in the peptide window.

### Peptide list commands in the main menu:

The first three menu items (**1/3 letter residue**, **Multicharged**, **Info**) correspond to the toolbar buttons mentioned above.

**Predict SS cross-links:** Lists a combination of all masses of combinations of all peptides containing cysteine residues.

In order to limit the number of potentially linked peptides you are asked to limit the number of peptides to combine to 2, 3 or 4. The list can be constrained to show only combinations of peptides having an even number of Cys residues, i.e. there will be no free cysteines.

The disulfide cross-links are more fully discussed below.

**N-glycosylation:** Displays the masses of peptides with potential N-glycosylation sites with the most common combinations of glycosylations (e.g. high mannose, complex and hybrid type), see discussion below.

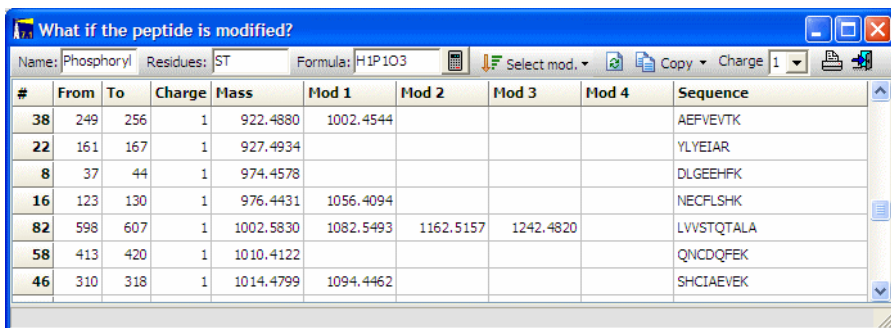
**DigestAlyzer:** Compare any number of protein digests against each other using any of two parameters: mass, hplc index, pI, hydrophobicity. For more details see Chapter 11.9 for details.

The remainder of the Peptide list menu (**MS/MS**, **HPLC chromatogram**, **Mass spectrum**, **Charge vs. pH**) is identical to the toolbar buttons listed above and discussed in Chapter 10.1 and below.

### What if?:

The 'What if the protein was modified' option enables you to view the mass values up to four modifications of a given kind on any peptide from the digest, if it contains the residues specified for the modification.

## 9 – Cleavages / Coverage map



#	From	To	Charge	Mass	Mod 1	Mod 2	Mod 3	Mod 4	Sequence
38	249	256	1	922.4880	1002.4544				AEFVEVTK
22	161	167	1	927.4934					YLVEIAR
8	37	44	1	974.4578					DLGEEHFK
16	123	130	1	976.4431	1056.4094				NECFLSHK
82	598	607	1	1002.5830	1082.5493	1162.5157	1242.4820		LVVSTQTALA
58	413	420	1	1010.4122					QNCDQFEK
46	310	318	1	1014.4799	1094.4462				SHCIAEVEK

The command opens a window listing all the peptides in the peptide window. Along the top of the window is a toolbar with the following commands available (from left to right):

*Name (of modification):* This field is not essential.

*Residues:* Single letter code of residues for which the modification is valid.

*Elemental formula:* Enter the elemental composition of your modification, you may use the 'formula calculator' button next to the edit field.

*Select mod.:* Click on the down-arrow to select a modification from the currently loaded modification file (see Chapter 4.3).

*Recalc button.* Press to recalculate the modified mass values.

*Copy to clipboard button:* If you just click the button, the table is copied to the clipboard; if you click on the drop-down arrow, a menu with options is displayed: *Copy to clipboard* standard copy; *Copy mass list* copies the complete mass list; *Copy just mass values* just mass values are copied; *Copy just modified mass* only the mass values of modified peptides are copied. Several of these options are handy if you need to create inclusion lists.

*Print.* Print the results.

*Exit.* Close the What if? window.

The results show each peptide with up to four modifications (if the peptide contains four residues that are allowed to be modified).

In the example above a peptide list is analyzed for possible phosphorylations by specifying Ser and Thr as potential phosphorylation targets and the phosphorylation formula P1O3 (mass ~80 Da). Peptide 4 (in the middle of the list) shows for example, that it can potentially be modified with up to two phosphorylations.

Use the local menu (right-click) to copy to clipboard or print the results.

## 9 – Cleavages / Coverage map

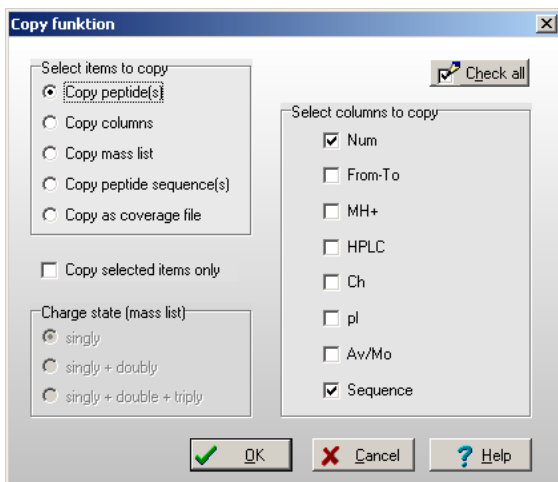
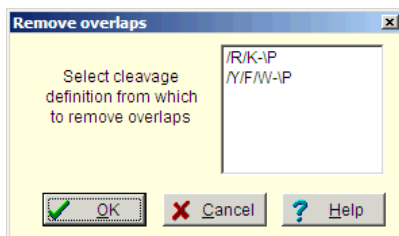
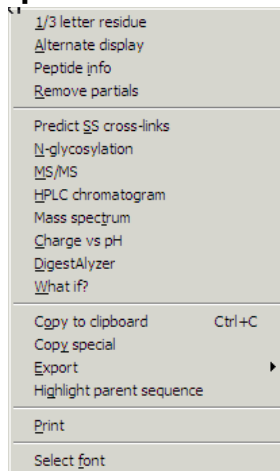
### Pop-up menu:

The pop-up menu (opened by clicking the right mouse button in the window) contains the same menu items as the **Peptide list** menu above with the addition of **Remove partials**, **Copy/Export** [see below], **Print** and **Select font**.

**Remove partials:** This is a special command for the case where you have specified multiple enzymes as cleavage parameter (e.g. trypsin combined with chymotrypsin as /K/R- $\downarrow$ P;/Y/F/W- $\downarrow$ P) and have overlapping peptides. If you then want to remove the overlaps from one of the definitions (e.g. if your experience tells you that trypsin cleaves completely while chymotrypsin generates overlapping peptides) you select '**Remove partials**' and from the list of enzymes (above) you select the cleavage definition from which to remove overlapping peptides. The peptide list will be redrawn reflecting the changes to the definitions. It is not possible to have different levels of partials (e.g. 1 for trypsin and 2 for chymotrypsin), only a partials level of 0 combined with another level (1, 2 ...). If you have combined more than two specifications and you want to remove additional overlaps, you select the options several times.

**Copy special:** The 'Copy special' command opens a dialog box that enables you to fine-tune what you want to copy. *Copy peptides* copies all the information displayed.

See below how to make a multiple selection, you can limit the number of items to copy by checking the '*Copy selected items only*' box. *Copy columns* opens yet another dialog box where you select which columns to copy. *Copy mass list* copy only the mass values, but enables the right-hand option of multiply charges. *Copy*



## 9 – Cleavages / Coverage map

*peptide sequence(s)* only copies the peptide sequences in 1-/3- letter format as displayed. 'Copy as coverage file' copies the list of peptides to the clipboard in coverage map format, this can then be pasted into the coverage map window (Chapter 12.7). If 'Copy selected items only' is checked; only the peptides selected in the list will be copied.

### Copying the peptide list to clipboard:

**Copy to clipboard:** The peptide list is copied to the clipboard. The format of the copied list is defined in the system setup (**Setup | Setup system**) .



**Note:** If you want to copy to a spreadsheet you should select 'Tab delimited', if you copy to a report select 'Copy as text'. With 'Tab delimited' each column will be in a spreadsheet column by itself. You also have the choice of copying the sequence with a limited length or full length. For more information see Chapter 5.2 'Setup peptide parameters'.

You can highlight and then copy part of the peptide list by using the usual Windows selections keyboard shortcuts

**Copy a continuous list:** Click with the mouse on the first entry, hold down the <Shift> button and click on the last entry. All entries between the two will now be selected. Select '**Copy to clipboard**'.

**Copy a discontinuous list:** Hold down the <Ctrl> button while clicking on the different entries. You can combine the two selection methods by first making a continuous list and then de-selecting individual items by holding down <Ctrl> while clicking on items to de-select. Select '**Copy to clipboard**'.

**Copy columns to clipboard:** Lets you select which columns of the peptide list to copy to the clipboard. The complete column will be copied, you will not be able to select a range.

**Select peptide as new protein:** A new sequence window will open on the GPMW desktop containing the currently selected peptide. This command is particularly effective if you want to carry out additional cleavages and experiments on an isolated peptide.

**Export: Export sel. peptide as new protein** will open a new sequence window containing the currently selected peptide. This enables you to perform all the standard sequence related functions on a peptide. The function is most useful for large peptides (see also Fragment window Ch. 3.7). **Create in-/exclusion list** starts a small 4-page wizard: 1<sup>st</sup> page in-/excludes modifications; 2<sup>nd</sup> page adds variable modifications and selects charge states; 3<sup>rd</sup> select output format and mass mode; 4<sup>th</sup> page enables you to review the mass list before writing it to disk in Micromass .pkl format.

**Export list to GRAMS** will export the peptide list to the now discontinued PerSeptive implementation of the GRAMS mass spectrum analysis software.

### Print

Printing the peptide list essentially gives an output that matches the display – a header showing protein information and a table with the same layout as the displayed list. The main difference is that in monochrome (e.g. laser printers) colors will be shown in bold. The sequence printed will be extended to the

## 9 – Cleavages / Coverage map

right margin of the page – if you select to print in landscape mode you will get more of the sequence printed.

**Print options** to be set immediately prior to printing.

**Printer.** The drop-down box enables you to select any printer installed on your system. The system default printer will always be the one initially selected.

**Print orientation:** Portrait (vertical) is usually the default. Landscape (horizontal) enables longer peptide sequences to be printed at the expense of the number of peptides.

**1-letter code/3-letter code.** The default will be what has been selected in the peptide window.

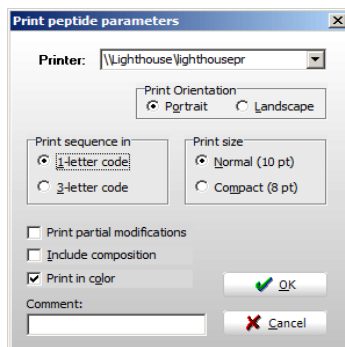
**Print size:** Normal (10 point) or Compact (8 point). This option refers to the peptide table only; the header and the (optional) composition will be printed in normal size.

**Partial modifications:** For all modified peptides, a list with partial modifications will be printed. E.g. a peptide having two phosphorylations will also be listed as having one and no phosphorylations. The partial modifications will be listed after the normal list.

**Include composition:** The amino acid composition will be printed in a table for every 10 peptides.

**Print in color:** Highlights and colored residues will be printed in color. On a monochrome printer (e.g. laser printer) the colors will in most cases be simulated in gray tones.

**Comment:** Here you can write any text that you want to include in the printout. A limit of 80 characters apply. The text is not preserved between printouts.



### Peptide info



The peptide info window can be accessed either by double-clicking on a line in the peptide list or by selecting a line followed by 'Peptide info' from the pop-up menu, the 'Peptide list' in the main menu or the '**Info**' button in the toolbar. The peptide info opens a dialog box showing physical/chemical information on the selected peptide.

The peptide information window can also be called directly from the sequence window (after highlighting a peptide – see chapter 3.2).

The peptide information window is divided into four panels with the following content:

- Top left: The top blue line represents the protein sequence and the green bar shows the relative position and coverage of the selected peptide. Below is shown the sequence position, length of peptide (with

## 9 – Cleavages / Coverage map

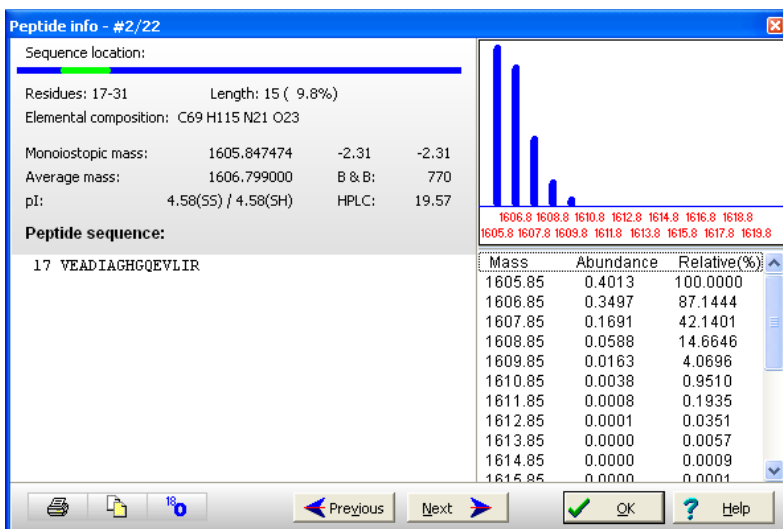
percentage of total sequence) and the elemental composition of the peptide.

Then follows various physical chemical characteristics of the peptide: Monoisotopic and average mass (6 decimals), charge (at the pH selected in System setup, chapter 5.2), Bull & Breeze index, theoretical pI and HPLC retention index. The charge and the pI labels have fly-by help showing the pH and the method used in the respective characteristic.

Bottom left: The sequence of the selected peptide.

Top right: The isotopic distribution of the peptide is shown as a stick diagram. The first 15 isotopes are displayed and the graph is always scaled to the largest isotope. The mass of each isotope is shown beneath each stick. Dragging the edge of the window will expand the isotopic distribution.

Bottom right: The isotopic distribution of the peptide in table form. First column shows the mass. Second column shows the abundance of each isotope and the last column shows the relative abundance with the most abundant isotope as 100%.



The arrows '**Previous**' and '**Next**' replaces the content of the window with the characteristics of the previous and next peptide in the peptide window. If the 'Peptide info' window is called from the sequence window (chapter 3.2) these two buttons will be grayed (non-active).

Selecting '**Print**' will make a hardcopy of all the information in the window except the blue/green peptide location line.

## 9 – Cleavages / Coverage map

Selecting **'Copy'** will copy all the text in the window to the clipboard. No graphics (peptide location line or isotope graph) will be copied to the clipboard.

The **'<sup>18</sup>O'** button enables you to quantitate peptides based on the incorporation of stable isotopes of <sup>18</sup>O during tryptic digestion. The method is based on the comparison of two samples, one digested by trypsin in normal water and the other digested by trypsin in <sup>18</sup>O water. When analyzing the resulting mass spectra, you will observe the incorporation of two oxygen molecules as isotope distributions where the +4 Da peak is much increased.

The ratio of unlabeled to labeled peptide can be calculated using the following formula:

$$\text{Ratio} = (I_4 - (M_4 * I_0) / M_0 - (M_2 / M_0) * (I_2 - (M_2 * I_0) / M_0) + (I_2 - (M_2 * I_0) / M_0)) / I_0$$

Here M<sub>0</sub>, M<sub>2</sub> and M<sub>4</sub> are the theoretical intensities of the peptide base peak (0) and the +2 and +4 Da peaks. Likewise I<sub>0</sub>, I<sub>2</sub> and I<sub>4</sub> are the observed intensities.

When you click the **'<sup>18</sup>O'** button, a dialog window opens for the calculation of the peptide ratio when incorporating heavy water.

The M values will be entered automatically by the program. You enter the ratios, measured from the experimental mass spectrum. Press the **'Calculate'** button to get the ratio of the two peptides (in this case 1.24).

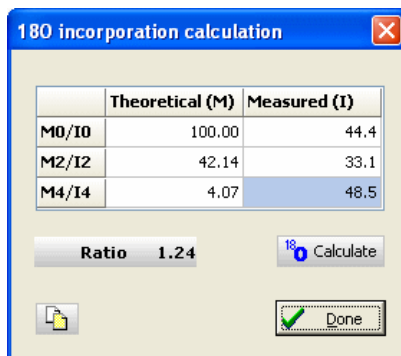
The **'Copy'** button  puts a copy of the results on the clipboard.

For references please see:

Xao, X. et al., Proteolytic <sup>18</sup>O Labeling for Comparative Proteomics: Model Studies with two Serotypes of Adenovirus. Anal. Chem., **73**, 2836-2842 (2001)

Johnson, K.L., and Muddiman, D.C., A Method for Calculating <sup>16</sup>O/<sup>18</sup>O Peptide Ratios for the Relative Quantification of Proteomes. J. Am. Soc. Mass Spectrom., **15**, 437-445 (2004)


Sigma currently have a kit <sup>18</sup>O Proteome Profiler™ for performing the quantitation experiments (Product Code P3623) along with a Technical Bulletin.



	Theoretical (M)	Measured (I)
M0/I0	100.00	44.4
M2/I2	42.14	33.1
M4/I4	4.07	48.5

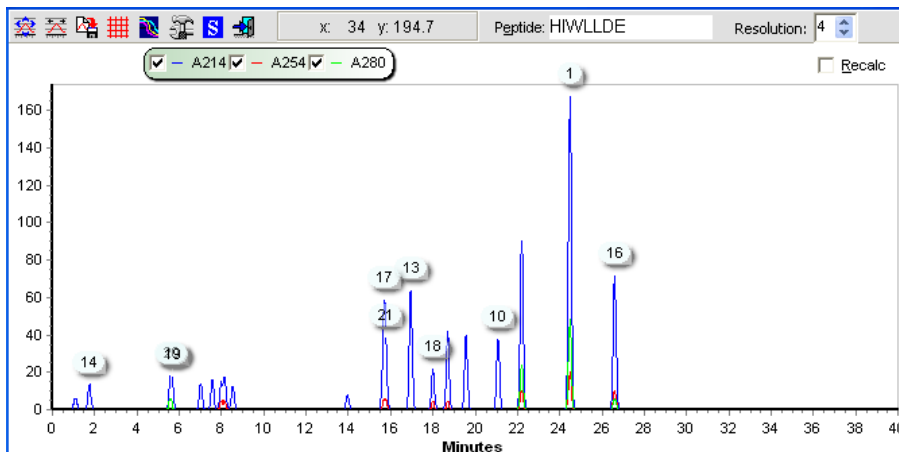
Ratio 1.24

<sup>18</sup>O Calculate

 Done

## 9 – Cleavages / Coverage map

### Simulated HPLC chromatogram



The simulated HPLC chromatogram [Y. Sakamoto, N. Kawakami & T. Sasagawa, J. Chrom. 442, 69-79 (1988)] is based on the separation taking place on a C18 column running a 0.1% TFA/water/acetonitrile gradient. The retention values are the ones displayed in the peptide list and are relative, you cannot translate them directly to minutes on your own separation system. Each peptide is labeled with the number from the peptide list (e.g. linear order of the polypeptide chain, overlapping peptides after the fully cleaved peptides).

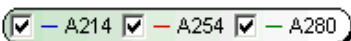
The peak heights are based on relative absorption at 214, 254 and 280 nm. The following relative values are given to the various bonds and amino acid residue side chains:

	214 nm	254 nm	280 nm
Peptide bond :	1	0	0
Cys and Met :	1	0	0
His :	5	0	0
Phe :	5	4	0
Tyr :	5	6	6
Trp :	33	10	24

The relative proportions between 214, 254 and 280 nm absorption does not reflect the absorption observed as different instruments have different bandwidth etc. and is only intended as a guide to enable you quickly to locate peptides containing aromatic residues.




## 9 – Cleavages / Coverage map



The graphs for 214, 254 and 280 nm can be disabled individually by unchecking the corresponding check-boxes in the legend. The colors of the three graphs can be set either in Setup System Colors (Ch. 5.3) or by using the graph control (see Ch. 11.1).

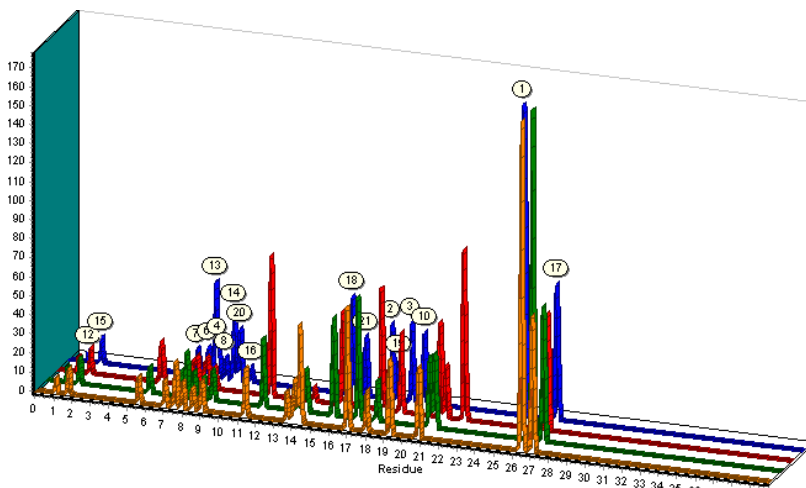
The buttons in the local toolbar are explained in Chapter 11.1 except for the **'Import peptide lists'**. This feature enables you to compare the retention time (e.g. the theoretical separation) of a number of peptide digests (these have to open on the GPMaw desktop). You may create up to 4 different

peptide lists (digests) and when clicking on the import button  they will be incorporated in the current window as new traces. The peptide lists can be the same digest of different proteins (see figure below) or different digests of the same protein (e.g. in order to determine which digest separates best on reversed phase HPLC).



**Note:** Only the 214 nm traces will be shown for each graph when multiple peptide lists are shown. Only the initial trace will be labeled with peptide numbers.

When you are comparing homologous sequences, they are best viewed in the 3D display mode:



The picture shows the tryptic digests of four myoglobins, imported into the same graph and displayed in 3D mode. For details of how to control the 3D mode, please see Chapter 11.1.

This graph is intended as a guide only, as the real chromatogram will probably never look like this. The yield of the different peptides by HPLC reversed phase separation will never be identical for all peptides, you will have partial cleavages, autodigest products, the exact position in the elution order will not be precise, the column ages with time etc. Taken with these

## 9 – Cleavages / Coverage map

precautions, the simulated chromatogram has turned out to be a reasonably good guide.

The **edit line** in the toolbar (labeled '**Peptide**') can be used for manual input of a peptide. This peptide will have an id number of 0 (zero). Each time a new residue is entered in the edit line, the graph will be redrawn to reflect the new position and absorption of the changed peptide. This will give the user an indication on how changes in single residues change retention time behavior.

The **Resolution** field can be set to change the peak width (values are 1-6) and can be used to simulate the separation on different systems.

The graph can be scaled, zoomed (right-down to zoom in, left-up to zoom out) etc. like all graphs, for more details please see Chapter 11.1.

### Predict SS cross-links

This command will either:

- 1) if no disulfide bonds are defined the function combines all Cys containing peptides in the digest and sorts them by mass.
- 2) calculate all defined disulfide cross-linked peptides in the digest, even if missed cleavages are defined in the digest.

**Note:** The '**SS**' button in the main toolbar has to be in the oxidized state (SS), not the reduced (SH) state, in order to calculate these values.

## 9 – Cleavages / Coverage map

**Predict Cys cross-links**

#	Av. mass	Mo. mass	Pep.	Cys	Peptides
1	361.49	361.19	1	1	#9
2	1019.16	1018.49	1	1	#10
3	1110.32	1109.55	1	1	#15
4	1362.63	1361.67	1	2	#25
5	1380.65	1379.68	2	2	#9#10
6	1432.64	1431.71	1	1	#13
7	1471.81	1470.74	2	2	#9#15
8	1493.60	1492.59	1	2	#12
9	1496.72	1495.75	1	1	#24
10	1724.12	1722.86	2	3	#9#25
11	1724.98	1723.85	1	1	#31
12	1794.13	1792.90	2	2	#9#13
13	1855.08	1853.78	2	3	#9#12
14	1858.21	1856.94	2	2	#9#24
15	1998.36	1997.03	1	1	#30
16	2086.46	2085.04	2	2	#9#31
17	2129.49	2128.03	2	2	#10#15

Max. # of peptides to combine  
☒ 2 peptides ☐ 3 peptides ☐ 4 peptides ☐ Only even # of Cys

Exclude below  Exclude above

Print
 Copy
 HTML
 171 entries
 Help
 Done

### No disulfide bonds defined:

The Cys-Cys linked list includes average and monoisotopic mass, number of peptides combined and numbers of Cys residues present in the combined peptides. Finally, the peptide numbers from the master peptide list are shown in the last column.

As options, you can choose number of peptides to combine and whether you want only to show an even number of Cys – in most cases it does not have much meaning to show an uneven number of cysteines, as this will leave a free unpaired SH group. Furthermore, you can define upper and lower exclusion mass limits.

When you copy the table, you can choose between either text format ('Copy') or HTML format ('HTML').

### Disulfide bonds defined:

In this case all possible combinations of disulfide-linked peptides are calculated. A maximum of approximately 1500 disulfide bonded peptides can be calculated. Only an approximate number can be given, as for each linkage the following steps are taken:

- 1) All possible extensions are calculated.

## 9 – Cleavages / Coverage map

- 2) The linkages are compared to all previously calculated linkages and deleted if previously calculated (i.e. calculated in a different order).
- 3) Finally, linkages containing overlapping peptides are removed.
- 4) Peptides containing only internal links are not reported as they are listed in the 'standard' list of peptides.

The maximum number of linkages is 1800, but as the interim number during calculations is somewhat higher than the finally reported number, 1500 is a more realistic maximum.

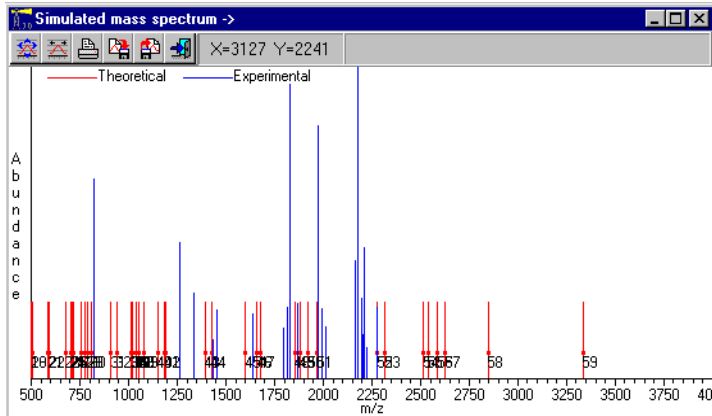
Note that when increasing the number of missed cleavages, the number of potential linked peptides increases hugely: a tryptic digest of BSA (17 disulfide bonds) generates 9 linked peptides. Only 8 are reported in the 'Predict SS cross-links' as a single peptide with internal links is not listed. When having a single missed cleavage the number increases to 145, and with two missed cleavages the number is 972 peptides. With three missed cleavages, the number of potential cross-links is above 1800 and is not reported.

### N-glycosylation


Please see Chapter 7.4 for information on N-glycan determination.

### Simulated mass spectrum

The simulated 'Theoretical mass spectrum' draws the masses of the current peptide list as a stick spectrum with a default mass range of 500 to 4500.



All the 'sticks' of the peptide list are drawn to the same height, 20%, of the window height. The 'sticks' are labeled with the peptide number from the list, and have a small error bar across them.

By pressing the '**Load peak table**' button (  ) you can load an experimental mass list (either a GPMW .PKS, a PerSeptive GRAMS peak list, Bruker peak list, or a Hewlett Packard MALDI-TOF peak list) into the spectrum. The spectrum loaded will be drawn in a different color and will

## 9 – Cleavages / Coverage map

usually have both a mass and intensity defined for each peak. The experimental spectrum will be drawn in a relative scale.

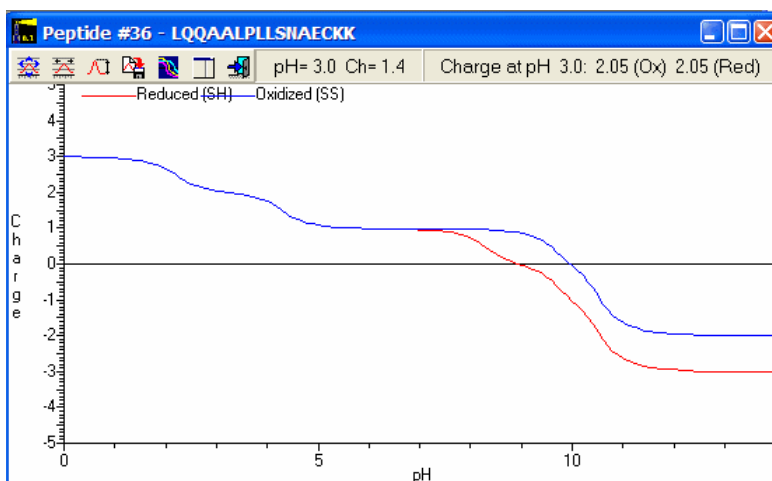
The error bars on the theoretical spectrum allow for easy comparison of the two spectra.

The graph can be scaled, zoomed etc. like all graphs, please see Chapter 11.1 for details.

### Charge vs. pH graph

The Charge vs. pH graph plots the charge of the selected peptide (shown in the window title) versus pH.

The x-scale shows the pH and the y-scale shows the charge.



The point where the graph crosses zero charges corresponds to the pI of the peptide. This value can vary slightly from the value reported in '**Peptide info**' (see above) as the algorithms used for calculations are slightly different. The steepness of the graph as it crosses zero indicates the confidence to put into the theoretical calculations. A peptide with a shallow crossover point is much more sensitive to the surrounding charges than a peptide with a steep crossover point.

Two graphs are displayed, one for the reduced and one for oxidized cysteine (there will only be a difference above the pI of Cys).

When you move the mouse cursor across the graph, the values of the position is continuously updated in the command bar.

You zoom by click and drag the mouse. Double click to un-zoom.

The first three buttons in the command bar




enables you to set full scale, enter


pH	Net charge	
	red	ox
0.0	3.00	3.00
0.5	2.99	2.99
1.0	2.97	2.97
1.5	2.90	2.90
2.0	2.68	2.68
2.5	2.24	2.24
3.0	2.05	2.05
3.5	1.95	1.95
4.0	1.76	1.76
4.5	1.32	1.32
5.0	1.10	1.10
5.5	1.03	1.03
6.0	1.01	1.01
6.5	0.99	1.00
7.0	0.97	1.00

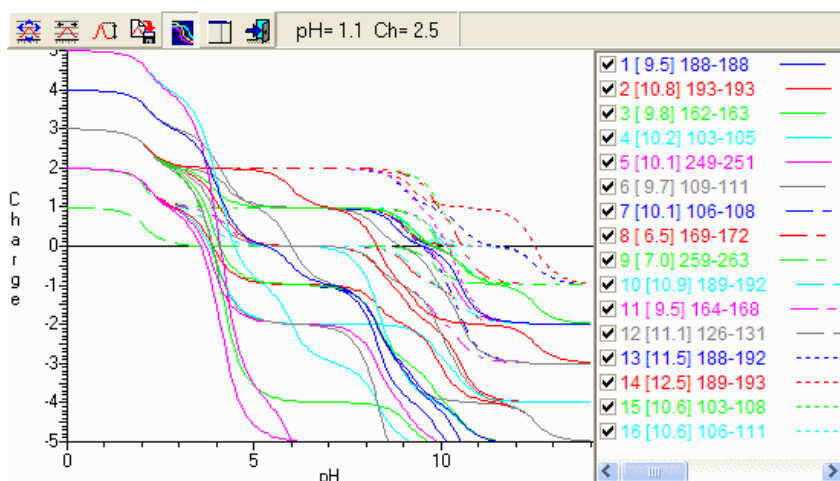
0.5 pH unit

## 9 – Cleavages / Coverage map

values for the x and y scale and automatically set the y-scale to the maximum charge in the graph.

The charge table button  opens a window to show a table of the net charge of the peptide at different pH values. The drop-down box at the bottom of the list enables you to change the list to show values at 1.0, 0.5 and 0.1 pH value difference. The list can be copied to the clipboard through the pop-up menu.

Pressing the 'Multiple graphs' button  change the display to show all peptides in the digest. In the right hand panel is a legend for the peptides indicating color and line style for each peptide. Use the check-boxes to turn the individual peptide graphs on and off.



### Isoelectric focusing (pI strip)

The '**Isoelectric focusing**' button or the pop-up menu in the peptide window will show all peptides in a simulated pI strip, in order to give an idea of the distribution of peptides after isoelectric focusing.

## 9 – Cleavages / Coverage map



Peptides with a missed cleavage will be shown in red, while fully cleaved peptides will be shown in green.

Checking the **'Labels'** button will display labels with the pI of each peptide. The **'pI 0-14'** will expand the x-scale to 0 to 14, this will only be needed if you have defined modifications with particularly high or low pK values. If **'No missed cleavages'** is checked, only fully cleaved peptides will be shown. The **'Copy'** button copies the graphics to the clipboard.

### Import peptide list

9.5

The main menu option **File|Import|Peptide list from file** enables you to read a list of peptides from a text file on disk and display the usual parameters for each peptide.

There are certain advantages to having a peptide list instead of a protein sequence, particularly if the peptides terminate in cleavage points that are not easily defined in the automatic digest (Ch. 9.1). You may want to analyze different variants of the same or similar (synthetic) peptides. Continuously modifying the peptides in GPMAW can be tedious while doing it in a text editor is straightforward. Another possibility is that the peptide list is generated by a different program.

The definition of a peptide list is a text file in ASCII format (e.g. can be edited by Windows Notepad) with a peptide to each line. Residues have to be written in 1-letter code. No headers or other additional information is allowed. E.g.:

```
CGEDYK
HHAISAK
TYFTDK
EEEEEEK
RPDADLK
GIYEE
```

The list will be imported and transformed into a sequence and put into a normal sequence window except that the peptide length information will be retained and a peptide window will be created immediately.

## 9 – Cleavages / Coverage map

The peptide window will be a standard peptide window (Ch. 9.4), and you will be able to perform all the standard operations. You will even be able to save the information in the sequence window as a regular GPMW sequence. However, in this case the peptide cleavage information will be lost.

A peptide list can also be searched for mass values, please see Chapter 12.6.

### Coverage map

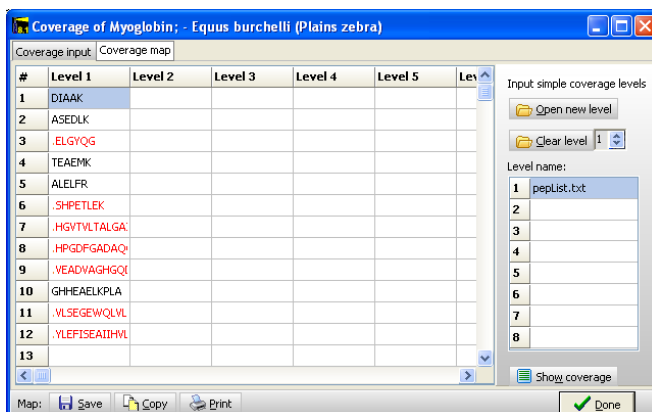
9.6

Coverage map is a graphical presentation of peptides found in a given protein, their location and other properties like e-values (if they have been found in a database search).

Currently you can access the coverage in two ways:

#### Peptide lists:

Given an open sequence window, you can select **Search | Peptide list coverage map**, which will open the coverage window on the Coverage input tab.



Select the **Open new level** button to load a peptide file from disk. This file is a list of peptides in text format. The peptide sequences have to be in 1-letter uppercase code with a single peptide per line.

GPMW now compares each peptide against the sequence, and places the peptides in a new empty level. If the peptide is found in the sequence, it will be listed in black, if not found it will be listed in red, and if multiple copies of the peptide is found, it will be colored yellow.

The name of the peptide file will be listed in the **Level name** box, which can be edited.

You can continue to add peptide files, as long as you have available levels (the maximum number of levels is 8). You can clear a level by entering the number and clicking on the **Clear level** button.

The sequences in the table can be edited, but this will not result in a new search and allocation by GPMW.



## 9 – Cleavages / Coverage map

When you click on the **Show coverage** button, the protein sequence will be shown with boxes representing the peptides found beneath the sequence.

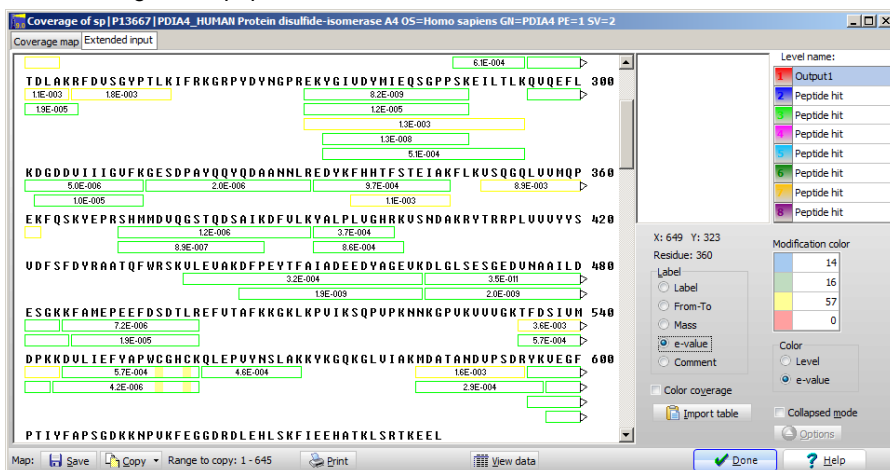
You can switch back and forth between the coverage map and the coverage sequence input through the tabs at the top of the dialog, but the coverage map is only calculated when accessed through the **Show coverage** button.

The **Save**, **Copy** and **Print** buttons at the bottom of the display works only on the coverage map, not on the table.

### Ms/ms coverage map:

The ms/ms coverage map is called from the results tab of the ms/ms search (see chapter 8.12, 8.13). When you have selected a protein in this search, a button labeled **Sequence** will show at the bottom of the window

Pressing this will open a sequence window containing the relevant protein with all identified sequences underlined (see chapter 3.4 for how to handle underlined sequences). The side panel will also open to show 'hit' parameters as you move the mouse cursor across each underlined peptide. Additionally, the coverage map will open in a separate window, showing all the peptides found.



If you close the coverage map window, it can be re-opened by clicking on the **Coverage** button of the sequence window (in the local toolbar)

The display shows the protein sequence in 1-letter code with 60 residues / line. Each peptide is shown in a colored box, initially containing the number of the first and last residue in the peptide. If the window was opened based on the peptide lists, each list will be represented by a single color (note: overlapping peptides are not allowed here). If the coverage was called from the ms/ms mass search, GPMW will try to assign overlapping peptides to different levels. The function can handle up to eight levels. If there are more overlapping peptides than can be accommodated by this, they will be shown overlapping on the bottom level with a red line underneath.

## 9 – Cleavages / Coverage map

If a peptide extends beyond the end of the sequence line, an arrow will be drawn and the rest of the peptide shown in a box on the next line. The label will be shown in the longest part. As a peptide cannot span more than two lines, the longest peptide (optimally) is 120 residues.

The label of the peptide can be changed in the right-hand panel:

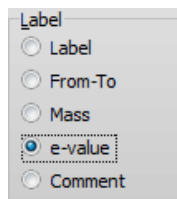
**Label:** A text string (initially this will be 'from-to').

**From-To:** number of first and last residue of the peptide.

**Mass:** Monoisotopic mass (theoretical) of the peptide

**e-value:** The expect value from the ms/ms search

**Comment:** A text string (not yet utilized by GPMW).



When you move the cursor across a peptide box, the information on the peptide will be shown in the right-hand text box. **Residue** is the residue pointed to, **Element** is the peptide; **Mass** is the theoretical

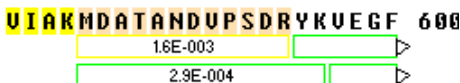
monoisotopic mass of the peptide; **Exp.** is the experimentally observed mass; **Delta** is the difference between the two mass values; **e-value** is the search

expect value; **Modif** is modifications found, shown as modification mass change (in blue) followed by number of modified residue.

Residue: 107  
Element: **100-110**  
Mass: 1429.633 Da  
Exp.: 1429.630 Da  
Delta: -0.003 Da  
e-value: 6.3E-005

Modif: 15.995@109

The colors of the sequence can be made to show the coverage of the given position by checking the **Color coverage** check box. No coverage is shown as a yellow background; single coverage as a weak yellow background. If more than two peptides cover a given position, no background is shown for the peptide.



Each level can have a name, which will be shown in the right-hand table. The number of the table is displayed with the color of that level.

Residues that have been identified as modified in the search are shown as colored blocks in the peptide boxes. They can be displayed in four different colors, depending on the modification. Each color (light blue, green, yellow, red) is assigned to an integer value, which can be set in the **Modification color** table in the right-hand panel. The last value is always regarded as 'the modification different from the other three', and the integer value is thus ignored.

The **Color** options defines the color of the line surrounding each peptide box. If set to **Level**, the colors will be the ones assigned to each level; when set to e-value the color will be dependent of the e-value of the assignment:

> 0.05: Red  
0.01 – 0.05: Orange  
0.001 – 0.01: Yellow  
< 0.001: Green

The data behind the coverage map can be viewed and edited through the

**View data** button  **View data** at the bottom of the window:


## 9 – Cleavages / Coverage map

Coverage of 1277824   Calnexin precursor (Major histocompatibility complex class I - Homo sapiens (Human))									
Coverage map Extended input									
#	Sequence	First	Last	Mass theo.	Mass exp.	Delta ppm	Label	Modif	e-value
1	VTYKAPVPTGEVVFADSF	58	77	2262.111	2262.107	-2			1.4E-000
2	APVPTGEVVFADSFDR	62	77	1770.836	1770.831	-3			4.0E-000
3	APVPTGEVVFADSFDRG	62	87	2813.417	2813.406	-4			1.3E-000
4	GTLGWLISK	78	87	1061.602	1061.599	-3			8.6E-000
5	AKKDDTDDEIAK	88	99	1348.662	1348.659	-2			1.4E-000
6	KDDTDDEIAKYDGK	90	103	1612.737	1612.736	0			1.6E-000

In this view all data can be changed (for good or worse!).






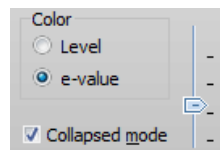
**Notice:** To get the data back to the coverage map you have to press the

**Coverage map** button  Coverage Map at the bottom of the screen. If you switch using the tabs at the top of the window, no exchange of data will take place.

### Collapsed mode (Heat mode)

When you have multiple levels in your coverage map, you can generate a single collapsed map by checking the **Collapsed mode** check-box.

**DYMI EQSGPPSKEILTLK VQEF L 380**  
  
**HHTFSTEIAKFLKUSQGQLUUMQP 360**  
  
**UGHRKUSNDAKRYTRRPLUUVVYS 420**  




The darkness of the color shows the number of 'hits' (layers) at a given site.

By adjusting the right-hand slider, you can change the contrast of the colored sections.

The collapsed mode is called directly from the .mgf file filtering function, and will then show the location of the N- or the C-terminal of potential peptides.



## Fragmentation

Fragmenting peptides based on the Roepstorff/Biemann notation or fragmentation by repeated cleavage of residues from either terminus.

The MS/MS fragmentation can be carried out on peptides selected either in the Peptide window (Chapter 9.4) or on highlighted sequences in a sequence window (Chapter 3.2). Once you have an ms/ms window you can enter a new peptide sequence or change the displayed sequence.

If you are working on *de-novo* interpretation of an ms/ms spectrum, you can use the Fragment analysis module (Chapter 12.3) to quickly scan for sequence tags.

### MS/MS fragmentation

10.1

MS/MS fragmentation - C63 H111 N17 O17									
HGTWLTALGGILK		MH+ 1378.8417		Mo	S	MS	N: Hydrogen - C: Free acid		
Backbone fragments		Fragment losses		Internal fragments		Simple view		Mass	Type
a	b		x	y''	z				
110.072	138.067	1 His 14	-	-	-			73.030	w' 1 +
167.093	195.088	2 Gly 13	12.65.748	1241.783	1222.742			74.025	v' 1 +
268.141	296.136	3 Thr 12	1208.727	1184.762	1165.721			110.072	a 1 +
367.209	395.204	4 Val 11	1107.679	1083.714	1064.673			128.072	z 1 +
466.278	494.273	5 Val 10	1008.610	984.646	965.605			138.067	b 1 +
579.362	607.357	6 Leu 9	909.542	885.577	866.536			147.113	y'' 1 +
680.410	708.404	7 Thr 8	796.458	772.493	753.452			167.093	a 2 +
751.447	779.442	8 Ala 7	695.410	671.446	652.404			171.078	x 1 +
864.531	892.526	9 Leu 6	624.373	600.408	581.367			195.088	b 2 +
921.552	949.547	10 Gly 5	511.289	487.324	468.283			201.124	w' 2 +
978.574	1006.569	11 Gly 4	454.268	430.303	411.262			202.120	v' 2 +
1091.658	1119.653	12 Ile 3	397.246	373.281	354.240			241.156	z 2 +
1204.742	1232.737	13 Leu 2	284.162	260.197	241.156			241.156	z 2 +
-	-	14 Lys 1	171.078	147.113	128.072			260.197	y'' 2 +
- Related immonium ions:									
- 74.061[T] 30.034[G] 44.050[A] 72.081[V] 86.097[I] 86.097[L]									
- 110.072[H]									

The ms/ms fragmentation can be selected

1. From the main menu when a sequence window has the focus (**Cleavage | MS/MS fragmentation**). If part of a sequence has been highlighted, it will be shown in the toolbar and be used for fragmentation. Otherwise the whole sequence will be taken as the basis for fragmentation. If the sequence is longer than 50 residues you will be asked if you want to perform ms/ms on the entire sequence.
2. By selecting ms/ms (button or local menu) for a given selected peptide in the peptide window ('Automatic cleavage', Ch. 9.4). This displays the

## 10 – Fragmentation

peptide in the toolbar of the MS/MS fragmentation window and the fragmentation below.

- From the local pop-up menu of the mass search result window (Chapter 6.1).

The fragmentation pattern is shown using the Roepstorff notation [P. Roepstorff & J. Fohlman, Biomed. Mass Spectrom. 11, 601 (1984)]. Notice that the y ions are shown as their double prime fragments (this can be changed in the ms/ms setup, see below).

The **title bar** of the window shows the elemental composition of the peptide under investigation.

Clicking the label of each column toggles through the three possible prime states (protons added to the base fragment): none, single and double prime.

After the list of fragment ions, the expected immonium ions are listed.



The toolbar at the top of the window contains from left to right:

- The peptide being fragmented is shown in the **edit line**. This line can be edited, and the table is updated whenever a change occurs. **Note:** if the line contains post-translationally modified residues, the edit line is grayed and cannot be edited.
- A panel showing the **mass** of the intact peptide with a button that toggles between average (blue) and monoisotopic masses (red) (please note that this button determines both the intact mass and the table values)
- **Setup** button. Set the chemical composition and number of fragment ions. See below for more details.
- **Ms/ms graph** button to show a graphical representation of the ms/ms fragmentation. This window is similar to the Ms graph in Chapter 6.1 except that each group of sequence ions can be turned on and off.
- **The frames button.** When activated, a size-able frame is displayed in the right hand part of the window listing all fragments from the main table in a numerically **sorted list**. This table is linked to the main fragment table, so when a value is selected in the sorted frame list, the corresponding mass in the main table will be highlighted in yellow. The left-hand part shows the mass, the right-hand part shows the ion-type and charge state. You can sort either column by clicking on the header. The gray arrow shows the sorting direction.
- Partial modifications. If the peptide contains modified residues (see below under posttranslational modifications) then the ms/ms list will be shown without modifications when the button is depressed.

Mass	Type
73.030	w' 1 +
74.025	v' 1 +
110.072	a 1 +
128.072	z 1 +
138.067	b 1 +
147.113	y" 1 +

## 10 – Fragmentation



- 'Graphical fragment mapper' . Opens a dialog window showing the peptide fragment as a picture with fragment marks and labels, see chapter 10.3.
- The '++' button toggles the multiply charged fragment ions on and off.
- The drop-down selection box determines the maximum number of charges that will be displayed when the '++' button is activated.
- Modification state of the **peptide terminals** is listed in the right hand panel. You can change the terminals either by double clicking on the panel or by right clicking in the window and choose 'Set terminals' from the local menu.

Masses are by default only shown with two decimals. However, in the local menu (right-click the mouse) you can toggle **4-decimal mass display** on and off.

### Fragment types / internal fragments

The most common fragmentation seen in ms/ms analysis is the fragmentation of the backbone (i.e. a,b,c,x,y,z ions). However, fragmentation resulting in loss of certain side chains and multiple cleavages of the backbone may also occur.

These alternative fragmentations are displayed on separate pages of the multipage dialog box. Clicking on the '**Fragment losses**' page shows the fragments that are the result of the loss of H<sub>2</sub>O from Ser or Thr and the loss of NH<sub>3</sub> originating from Arg.

'**Internal fragments**' lists the fragments that arises from double cleavages of the backbone. As this results in a

huge number of possibilities, only the most common fragments are shown (those directed by Pro, His, Lys, Asp and Glu). Along with the mass of the fragment is shown the commonly seen –28 Da mass.

Backbone fragments		Fragment losses		Internal fragments	
b-H <sub>2</sub> O	Res.	y-NH <sub>3</sub>	y-H <sub>2</sub> O		
-	1 Lys 14	-	-		
-	2 Asp 13	-	1354.528		
-	3 Gly 12	-	1239.501		
-	4 Leu 11	-	1182.480		
-	5 Gly 10	-	1069.396		
-	6 Glu 9	-	1012.374		
-	7 Tyr 8	-	883.332		
846.400	8 Thr 7	-	720.269		
948.401	9 Cys 6	-	619.221		
1049.449	10 Thr 5	-	517.220		
1151.450	11 Cys 4	-	-		
1264.534	12 Leu 3	-	-		
1407.592	13 Glu 2	-	-		
-	14 Gly 1	-	-		

### Simple view

This view shows a slightly more simplistic view of the table. In the right-hand side of the

Backbone fragments		Fragment losses		Internal fragments		Simple view	
a	b	c	Res.	x	y	y-17	
110.072	138.067	157.108	1 H 14				<input checked="" type="checkbox"/> a
167.093	195.088	214.129	2 G 13	1265.748	1241.783		<input checked="" type="checkbox"/> b
268.141	296.136	315.177	3 T 12	1208.727	1184.762		<input checked="" type="checkbox"/> b-18
367.209	395.204	414.245	4 V 11	1107.679	1083.714		<input checked="" type="checkbox"/> c
466.278	494.273	513.314	5 V 10	1008.610	984.646		<input checked="" type="checkbox"/> x
579.362	607.357	626.398	6 L 9	909.542	885.577		<input checked="" type="checkbox"/> y
680.410	708.404	727.446	7 T 8	796.458	772.493		<input checked="" type="checkbox"/> y-17
751.447	779.442	798.483	8 A 7	695.410	671.446		<input type="checkbox"/> y-18
864.531	892.526	911.567	9 L 6	624.373	600.408		<input checked="" type="checkbox"/> z
977.615	1005.610	1024.651	10 P 5	544.325	520.356		

## 10 – Fragmentation

table, the columns can easily be turned on and off, and the number of decimals can be changed between 2 and 4.

**Printing** is done from the local menu, main toolbar or main menu (**File|Print**). **Copy to clipboard** is done in a similar manner (menu command: **Edit|Copy to clipboard** or <Ctrl+C>). Through the pop-up menu you can further select to copy an individual column (e.g. **Copy special | C ions**).



**Note:** The length of the sequence to be fragmented is limited to 400 residues!

## Posttranslational modifications

You can modify individual residues in the peptide under fragmentation by double clicking on the relevant residue. In the resulting 'Insert modification' dialog box, you either enter the modification name and composition or you select the relevant modification from a modification file. For more details see 'Amino acid modifications', Chapter 3.6.

When a residue is modified, the sequence edit line in the toolbar will be grayed, and it will not be possible to edit the sequence. Modifications will be displayed in an information line at the bottom of the display.

Modifications: Glu6 Methylation /



**Info:** If your sequence have modified residues when selected (e.g. from the sequence window) you will get the message “Sequence contains modifications. Edit line not active” which means that you cannot edit the sequence in the edit line (grayed), but the modification will be incorporated in the mass calculations. You should beware of possible fragmentation of the side chain (e.g. loss of the modification).

The partial modifications button in the toolbar will remove the masses of the modifications from the mass list when depressed (i.e. the peptide will be calculated as un-modified).

## Linked peptides

The linked peptides option is a display similar to 'Simple view' (see above) except that it enables two peptides to be linked by a disulfide bond.

One of the peptides to be linked is the one in the toolbar, the other is to be entered (or pasted) into the edit box of the toolbar of the Linked peptides panel:

a	b	b-18	c	Seg	x	y	y-17	y-18	z
---	---	------	---	-----	---	---	------	------	---

The linkage position have to be defined, either by manually entering the positions in the '1<sup>st</sup> link' and '2<sup>nd</sup> link' edit boxes, or more simply, by pressing the **'Find'** button, which will search and enter the last Cys in each of the two sequences.



## 10 – Fragmentation

Press the '**Refresh**' button to update the fragment grid.

1447.683	1475.678	1457.667	1492.704	9 G 3	459.257	433.277	415.266	416.250	417.258
1548.730	1576.725	1558.715	1593.751	10 T 2	402.235	376.255	358.245	359.229	360.237
1676.825	1704.820	1686.810	1721.846	11 K 1	301.188	275.208	257.197	258.181	259.189
				12 K 0	173.093	147.113	129.102	130.086	131.094
88.040	116.035	98.024	133.061	1 D 11					
201.124	229.119	211.108	246.145	2 I 10	1761.878	1735.898	1717.888	1718.872	1719.879
357.225	385.220	367.209	402.246	3 R 9	1648.794	1622.814	1604.804	1605.788	1606.795
1676.825	1704.820	1686.810	1721.846	4 C 8	1492.693	1466.713	1448.702	1449.686	1450.694
				5 K 7	173.093	147.113	129.102	130.086	131.094

The fragmentation of the sequence in the main toolbar will be shown first followed by an empty row and the fragmentation of the second sequence.

The Linked peptide fragmentation can of course also be used with links other than Cys-Cys links. In this case you cannot use the '**Find**' button but have to enter the positions manually. If the linkage has a mass, you enter this in the '**Linkage map**' edit field.

The actual fragments to display, number of decimals and charge are defined by the right-hand panel like for 'Simple view'.

### MS/MS setup

Pressing the setup button in the MS/MS window enables you to edit the ms/ms table, mass of fragment, fragment displayed, and default prime state.

**Edit MS-MS parameters**

N-terminal fragments					C-terminal fragments						
	Name	Compos	Protons	OK	Fragment type		Name	Compos	Protons	OK	Fragment type
1	a	-C1O2H1	0	<input checked="" type="checkbox"/>	Backbone cleavage	x	-H1+C1O1	0	<input checked="" type="checkbox"/>	Backbone cleavage	
2	b	-O1H1	0	<input checked="" type="checkbox"/>	Backbone cleavage	y		1	<input checked="" type="checkbox"/>	Backbone cleavage	
3	c	-O1+N1H1	1	<input type="checkbox"/>	Backbone cleavage	z	-N1H2	0	<input checked="" type="checkbox"/>	Backbone cleavage	
4	d	-O1+C2N1H3	1	<input type="checkbox"/>	d,w type chain loss	v	C2N1H2O1	1	<input checked="" type="checkbox"/>	v type chain loss	
5			0	<input type="checkbox"/>	Backbone cleavage	w	C3O1H3	1	<input checked="" type="checkbox"/>	d,w type chain loss	
6			0	<input type="checkbox"/>	Backbone cleavage			0	<input type="checkbox"/>	Backbone cleavage	

Formula    Default    Display 4 decimals    OK    Cancel    Help

The name column gives the name of the fragment (1 character).

In the composition column you enter the composition which may be both positive and negative (i.e. additions and removal of atoms). When the cursor is in the 'Compos' field, the Formula button (bottom left corner) will be active, allowing you to enter the composition through the 'Elemental composition calculator' (Chapter 12.2).

In the protons columns you enter the numbers of primes (0, 1 or 2 protons that are added during fragmentation). If the '**OK**' column is checked the corresponding fragment column will be displayed.

## 10 – Fragmentation

In the last column you enter the fragmentation type, 'Backbone cleavage', 'v type chain loss' or 'd, w type chain loss'.

If the 'Display 4 decimals' is not checked, the results will be displayed with 3 decimals.

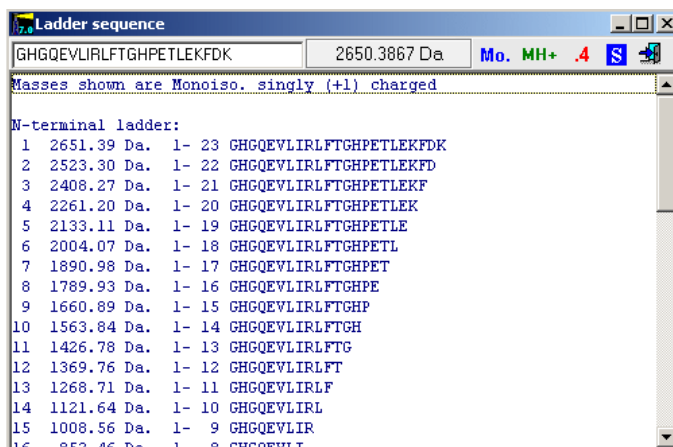
The 'Default' button will reset the table to its default values (as shown in the picture above).

### References:

P. Roepstorff & J. Fohlman, Biomed. Mass Spectrom., 11, 601 (1984)

## Ladder sequencing

10.2



Ladder sequencing shows the sequence and masses of the selected (or manually edited) sequence after sequential removal of residues from either the N- or the C-terminal end. The ladder sequence can show a maximum of 100 residues.

The function is selected in a manner similar to ms/ms fragmentation.

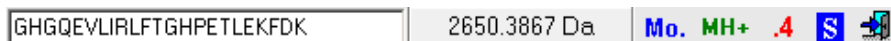
- Ladder sequencing can be selected when a sequence window is open by selecting **Cleavage | Ladder sequence**. If part of the sequence is highlighted (Chapter 3.2), the selected sequence will be displayed in a list with successive removal of a residue from the C-terminus followed by a list of peptides with successive removal of residues from the N-terminus. If no sequence is selected, the display will initially be clear and you will have to enter a sequence manually.
- When you have a peptide list (from automatic fragmentation) you select a peptide (highlight) and choose 'Ladder sequencing' from the main or the pop-up menu.

The ladder sequence can be edited in the edit box of the local toolbar. Changes are immediately shown on screen.

## 10 – Fragmentation

The first line of the list tells you the mass type (ave./mono.) and charge state of the peptide masses. Then follows the N-terminal sequence ladder followed by the C-terminal ladder.

### Toolbar



From left to right the toolbar shows:

**Edit box** for the base sequence. The ladder sequences are updated whenever changes are made to the base sequence.

**Mass** of base sequence.

**Av./Mo.** Toggles between average and monoisotopic masses.

**M / MH+ / MH++ / M-H** Charge state button. Toggles between no charge, singly charged positive ions, doubly charged positive ions and singly charged negative ions.

**.4/.2** toggles between displaying the mass using four and two decimals.

**‘S’ Setup:** Enables you to specify whether the N- and/or C-terminal modification of the parent window should be used when generating each step of the ladder sequence.

**Exit:** Close the ladder sequence window.

**Printing** is selected either from the main toolbar (<Ctrl+P>) or the pop-up menu.

### Pop-up menu

The pop-up menu contains, in addition to the commands in the toolbar, the option to set the **charge state**. The available options are: No charge (default), singly charged (MH+), doubly charged (M2H++) and negative singly charged (MH-). In addition you can select Average/Monoisotopic, 2/4 decimals, Setup and Printing.




Like for the ms/ms fragmentation above, modified residues in your sequence you will give you the message “Sequence contains modifications. Edit line not active”. This means that you cannot edit the sequence in the edit line (grayed), but the modification will be incorporated in the mass calculations.

### Graphical fragment mapper

10.3

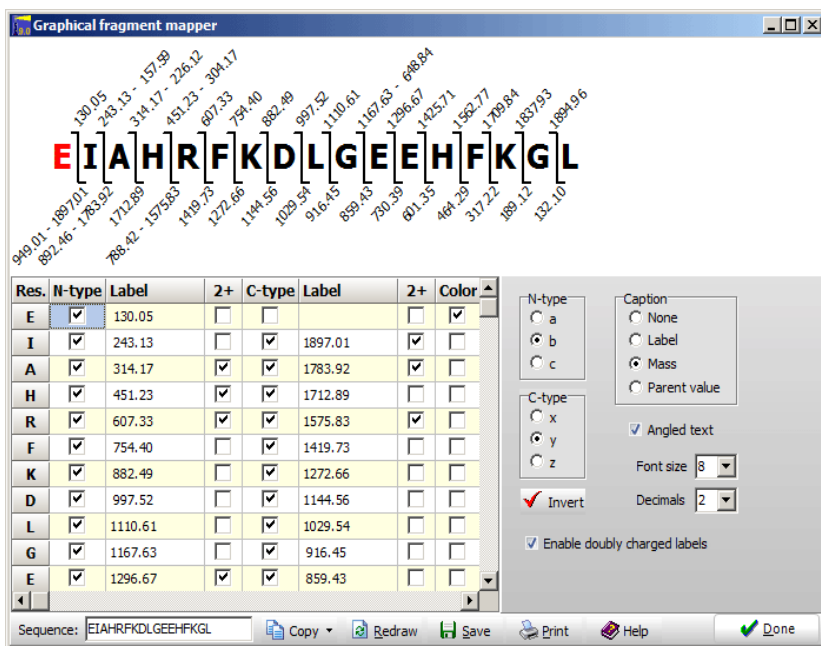
The ‘Graphical fragment mapper’ is an option to generate nice-looking peptide fragments for publications, reports or documentation in spectra.

This dialog can be opened from the main menu (Utilities | Graphical fragment

mapper), from the ms/ms fragment window (see 10.1) through the  button and from a number of other windows showing fragment patterns of peptides.

From the main window the mapper will open without a sequence, while if called from the ms/ms window, the working peptide will be transferred.

## 10 – Fragmentation



The sequence to show is displayed in the bottom left corner labeled 'Sequence', and can be freely edited. The fragment-mapped sequence is displayed in the top box and is controlled by the table below.

The left-hand column shows the residue for the corresponding position, the second and fourth columns show a check-box for displaying the corresponding fragment, while the third and fifth column shows the text to display. If the last column is marked, the residue will be colored red.



**Note:** The label can be freely edited in the table.

The initial caption of the fragments is controlled by the 'Caption' radio buttons in the right-hand toolbar.

**None:** no caption;

**Label:** fragments will be labeled with the N- and C- fragment type selected in the right-hand radio buttons;

**Mass:** the mass of the corresponding fragment will be displayed.

**Parent value:** if the fragment mapper is called from the ms/ms window, this option loads the mass values from the 'Simple ms/ms table'. This enables to get modified mass values (e.g. PTM's).

The **N-type** and **C-type** radio buttons determines what kind of labels will be displayed above and below the sequence line, i.e. a, b or c ions for N-type and x, y and z ions for C-type fragments.

**Angled text:** The text is rotated 45 degrees.

## 10 – Fragmentation



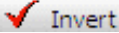
**Note:** Doubly charged mass values are only shown if you have checked this box.


**Font size:** Determines the size of the label.

**Decimals:** Number of decimals (0-2) when the label has been chosen to show the mass.





**Note:** The fragment mass displayed is calculated based on the rules set out in the ms/ms fragment window (see 10.1).


The **'Invert'** button  inverts the selection made in the N-term and C-term columns. If the image is not updated, you can press the Refresh

button  to force an update (this is necessary whenever you check or un-check a box in the table).

**Enable doubly charged labels:** Check this to expand the graphical display area vertically.

The **'Copy'** button  copies the fragment pattern to the clipboard as a meta-file, ready for pasting into your report (ideal for Word and PowerPoint as the graphics can be scaled without loss of resolution). Through the associated drop-down menu, you can select to copy either as a metafile or as a bitmap file. If you copy as bitmap, it will be the exact size of the displayed graphical area (resize the dialog to fit the sequence).

The **'Save'** button  saves as a meta-file picture to the disk (.emf –

Windows enhanced metafile), while the **'Print'** button  sends an image to the currently selected printer.



## Graphs

Graphs are used in GPMW to display a number of parameters from secondary structure prediction to user-defined graphs.

Several graphs have been used as daughter windows of other windows. They behave in a manner similar to the windows described in this chapter. This concerns composition search (Chapter 7), simulated HPLC chromatogram, simulated mass spectrum, and charge / pH graph (Chapter 9).

### Common graph commands

11.1

Most graphs in GPMW, whether they are line or stick graphs, share some common commands:

#### Toolbar



The layout of the toolbar varies slightly from window to window depending on the capabilities of the current graph.



Reset graph to full scale.



Set scale. Opens a dialog box that enable you to specify x- and y-scale.



Save graph to disk as a file in Windows metafile format (vector format). A vector format can be rescaled in the target application (e.g. Word) without loss of resolution.



Toggle the horizontal/vertical grid on and off.



Only for sequence related graphs. When depressed, you can draw a horizontal line (the selection bar) on the graph, that will highlight the corresponding sequence in the sequence window. See an example below in section 'Hydrophobicity', 11.3.



Toggle the **'toolbox'** on and off. The toolbox contains controls for adjusting the 3-dimensional 'look' of the graph, setting the colors etc. for more details please see below.



Exit. Close the graph window.

x: 270 y: 180.2

Current positions of the mouse cursor in graph x- and y- values.

## 11 – Graphs

### Mouse commands

**Zooming:** You can zoom (expand) part of the graph by pointing to the top left corner of the area to be zoomed, press and hold the left mouse button while dragging the mouse cursor down and right, covering the area to be zoomed. You can zoom several times in order to zoom into small details.

The selected area will be 'grayed' in order to give you a clear indication of what part of the graph will be zoomed. The zoom function also works in 3-D.

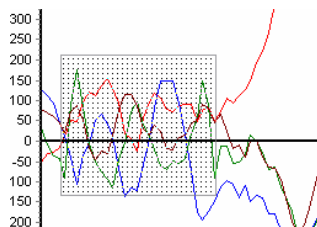
You unzoom by click and drag the mouse cursor left and up or you left double-click anywhere in the graph area.

**Pop-up menu:** Right-click anywhere in the graph area opens the local pop-up menu enabling you to choose between the following commands:

**Full scale, Set scale, Import sequence, Copy as bitmap, Copy as metafile, Save to file, Print, and Exit.**

**Full scale, Set scale, Save to file** and **Exit** are identical to the toolbar buttons described above.

A graph may additionally have the **Save to file (text)** option. This will save the graph as x/y values in a text file, that can be taken into a spreadsheet for additional editing. The **Import sequence** command is specific to certain graphs, please see the respective graphs. **Copy as bitmap** and **Copy as metafile** copies the current graph to the clipboard in either bitmap or metafile (vector) format. You should copy as metafile whenever you need to scale the graph. However, some programs do not accept metafiles, and in these cases you need to copy as bitmap. You can use the keyboard shortcuts indicated to control the commands from the keyboard. The **Edit | Copy to clipboard** command in the main menu will copy the graph to the clipboard in bitmap format.



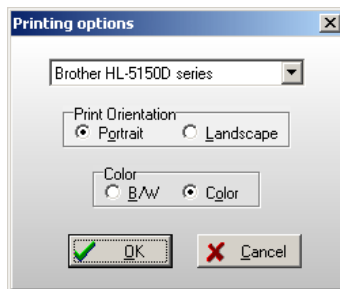
Full Scale	
Set scale	
Import sequence	
Copy as bitmap	Ctrl+C
Copy as metafile	Ctrl+Alt+C
Save to file	Ctrl+S
Print	Ctrl+P
Exit	

### Print

When selecting '**Print**' (through the main menu, the main toolbar or the pop-up menu) you have to acknowledge a dialog box with printing options (right).

From the '**Printer**' drop-down box you can select any printer installed on your system. The system default printer will always be the one initially selected.

The Graphs are always printed in a pre-determined ratio of width to height, but utilizing the full width of the paper. This means that you can get a considerably





## 11 – Graphs

larger graph by setting the printer to print in landscape format.

Graphs can be printed in either black and white (B/W) or in color. If you have a monochrome (laser) printer you may have to experiment which selection is best as different printers translate the color table in different ways.

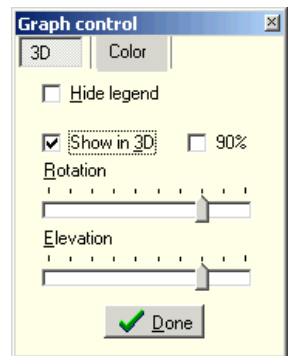
### Toolbox (graph control)

The toolbox button toggles the display of the graph toolbox on and off. The box is a two-page control. When selecting the **‘Done’** button, the toolbox will be hidden, but will not be deleted. This means that all settings will be retained when the dialog box is reopened. The toolbox will always stay on top of the GPMW program.

#### 3D:

The 3D page controls the display properties of the graph. By default the graph is shown in two dimensions (flat), but by checking the **‘Show in 3D’** check box, the graph changes to a three-dimensional display. The two sliders below the check box can be used to rotate the graph around the vertical axis (**Rotation**) or the horizontal axis (**Elevation**).

**Hide legend:** When checked, the legend in the graph will be hidden (can be useful when rotating the graph).

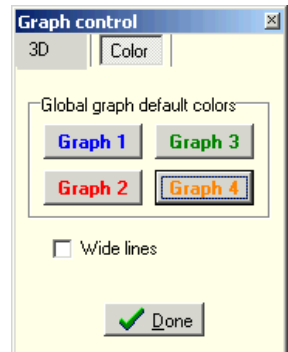


**90%:** When checked, the size of the graph will be reduced to 90% of the normal size. This feature is most useful when the graph is shown in 3D.

#### Color:

The color page of the dialog enables you to set the color of the first four graphs displayed in the graph. If more than four graphs are displayed at any time, the color of the additional graphs will be determined automatically.

**Wide lines:** The width of the lines used to draw the graph is by default 1 pixel. When the 'Wide lines' check box is checked, the lines will be drawn with a width of 2 pixels.

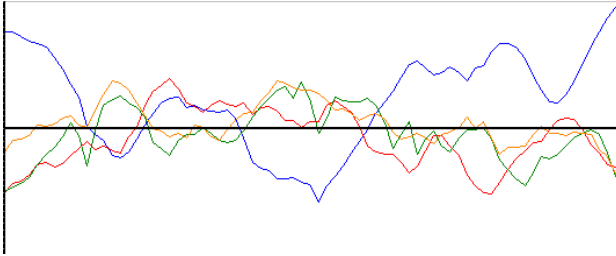


## 11 – Graphs

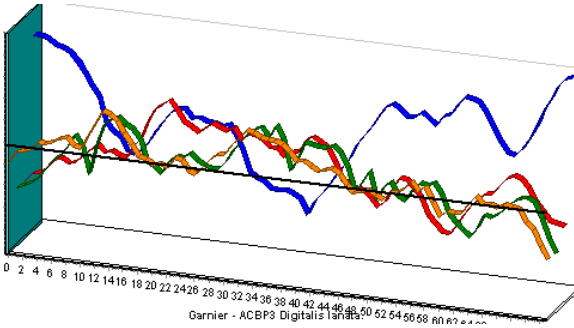
### 3-D effects

The three-dimensional effect is immediately apparent when you select the 3-D option in the toolbox.

Two-dimensional display:

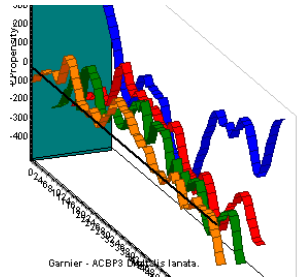
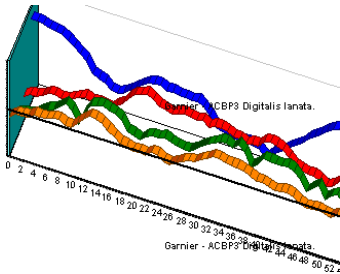


Turning on the three-dimensional effect will produce this display




Using the '**Rotate**' sliding bar will turn the graph around the vertical axis (right).

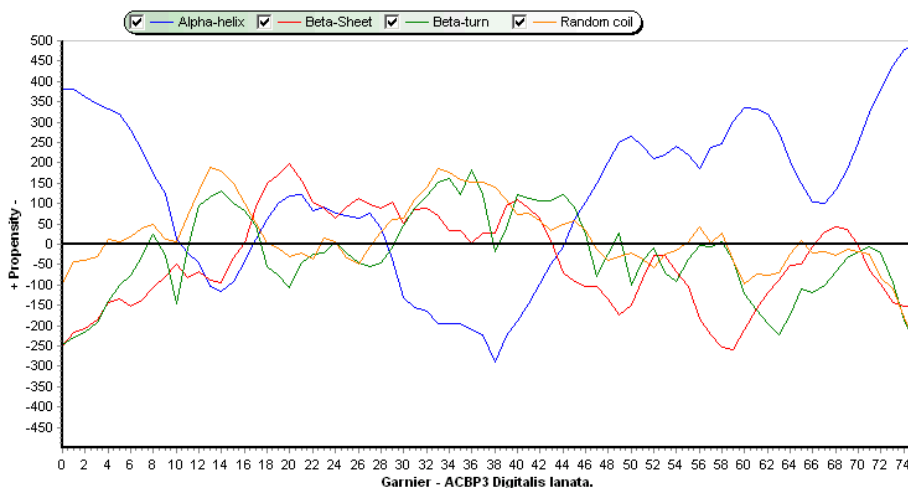
Using the '**Elevation**' sliding bar will turn the graph around the horizontal axis (below).



## Secondary structure prediction

11.2

The secondary structure prediction  is based on the values of Garnier [J. Garnier, D.J. Osguthorpe & B. Robson, J. Mol. Biol. 120, 97-120 (1978)] – also known as the GOR secondary predictions.



When the command is selected the daughter window opens immediately showing the calculated propensity graphs for alpha helix, beta sheet, beta turn and random coil. Each graph has its own color, which can be set in Setup | Colors (Ch. 5.3, all system colors) or through the toolbox (Ch. 11.1, just graph colors).

Values well above zero, isolated from other curves, give a good indication for a particular structure. Other things to look out for is that alpha helices are usually of a length of 10-15 residues and beta sheets are never single, but always occur in multiples, often separated by beta turns in the case of anti-parallel beta sheets.


The different curves can be toggled on and off in the check-boxes in the legend.

Please note that the predictions are far from accurate, but only give an indication of a given structure. The proposed structure should always be checked by other means: alpha helical wheels (Ch. 11.7), homologous structures, circular dichroism measurements etc.

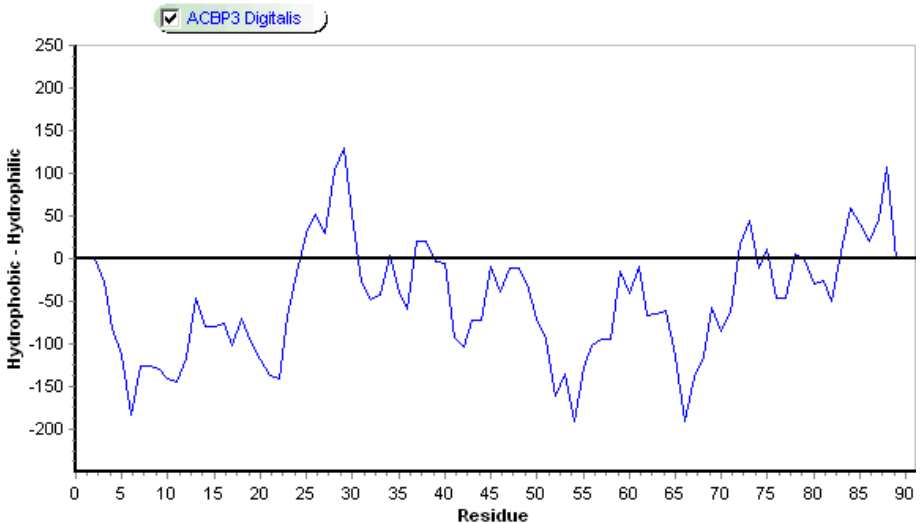
You can compare predicted alpha-helical sections with the alpha-helical wheel, Chapter 11.7.

## Hydrophobicity

11.3

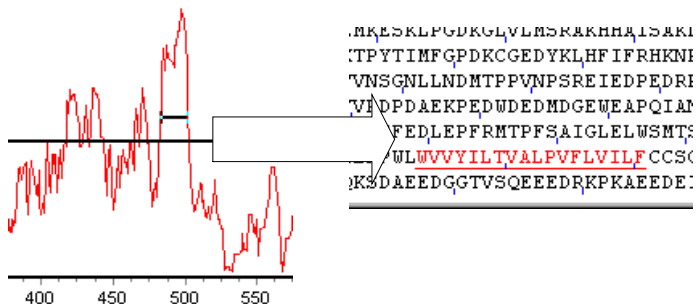
Selecting **Graph | Hydrophobicity** or the main toolbar button  opens a daughter window with a hydrophobicity graph based on the hydrophobicity values of [J. Kyte & R.F. Doolittle, J. Mol. Biol. 157, 105-132 (1982)].


## 11 – Graphs



The graph shows hydrophobic regions above the X-axis and hydrophilic regions below. The graph is particularly good at detecting transmembrane regions of a protein (around residue 500 in the above picture, an alpha-helical transmembrane section will normally have a width of 20-22 residues) and you can also often locate the activation peptide (left side of the graph).

Additionally, the hydropobicity graph may indicate areas that can generate 'sticky' peptides that are difficult to handle – or you may just want know where in the sequence the transmembrane region is located.

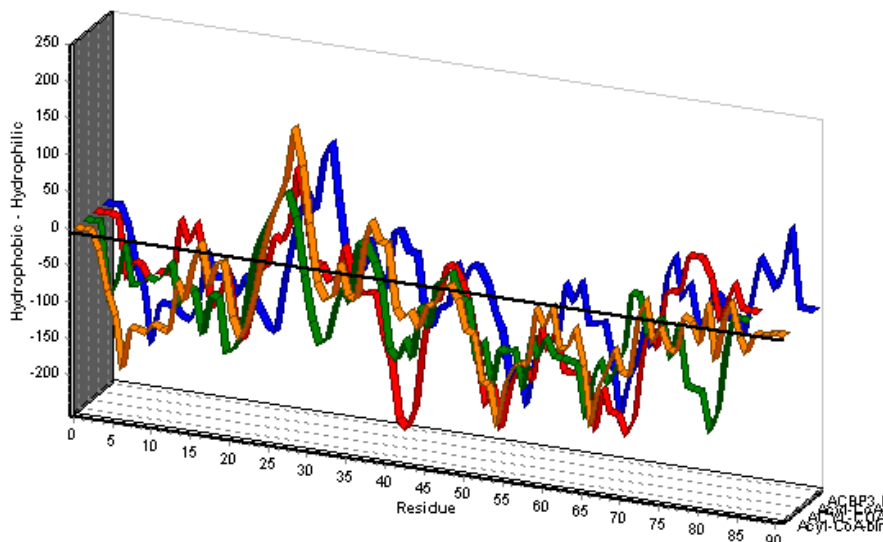


When the **selection bar**  in the toolbar is depressed, you may draw a horizontal bar across important regions of the graph and dynamically see the same region underlined in the protein sequence. The same region (residue numbers) will be shown in the toolbar in sharp brackets

[483-499] x: 580 y: 214

## 11 – Graphs

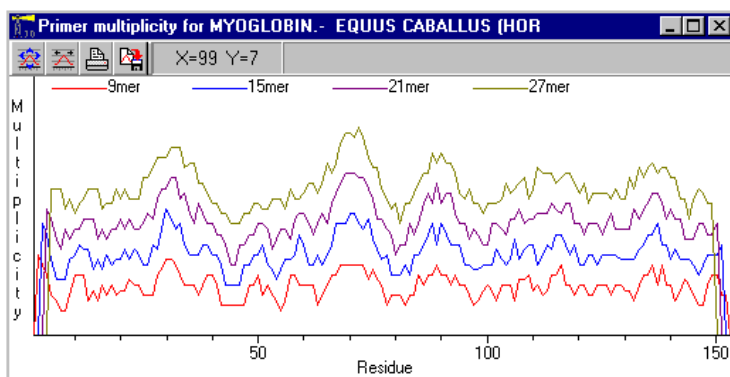
Compared to the standard graph, an additional menu item is available in the pop-up menu: **Import sequence**. Activating this command will import up to three additional sequences currently opened in GPMW into the same window for comparison. Below an example of four sequences is shown displayed in 3-D:



In the pop-up menu there is an option 'Save to file (text)' which enables you to save the hydrophobicity graph as x/y values to a text file for import into other programs like Excel.

### Primer multiplicity

11.4



If you want to generate a primer from a protein sequence the **Graph | Primer multiplicity** generates a graph showing various multimers based on the protein sequence

## 11 – Graphs

Low points in the graph show the lowest multiplicity, that is the best position for generating primer. Point the mouse cursor at the points of interest and read the cursor position in the toolbar (see general graph information above). Please note that all graphs go to zero at the ends - this does not mean that the terminals are the best sequences for basing a primer.

### User graph

11.5

Using the **Graph | User graph** option it is possible to customize a graphical display of various protein features.

In the Residue graph parameters, you enter values for each residue depending on the feature you would like to emphasize. The example above displays the distribution of positive, negative, hydrophobic, and small residues along the polypeptide chain by assigning each of the residues in the group a value of 10. You do not have to assign the same value to every residue, but you should keep the average value in the range of 10-20, as the height of the graph otherwise will be too small or too large to show any details.

File name: RESIDUE.UGF

Graph name: Residue character

Graph comment: Chemical character

Residue	Graph 1	Graph 2	Graph 3	Graph 4
N-term.	10	0	0	0
C-term.	0	10	0	0
Xxx	0	0	0	0
Ala	0	0	0	10
Cys	0	0	0	0
Asp	0	10	0	0
Glu	0	10	0	0
Phe	0	0	0	0

Graph ID: Positive Negative Hydrophob Small

Reset

Load

Save

Graphs to display: 4

Scaling factor: 2

Initial scale: 50

Smoothing: 1

OK Cancel Help

#### Parameters:

**File name:** If the data are read from a file, this field shows the file name (8 characters)

**Graph name:** Title of the graph.

**Graph info:** Supplemental information.

## 11 – Graphs

**Graph 1..Graph 4:** Values can be specified for each amino acid residue for up to four graphs (remember to enable the graph in the 'Number of graphs' box).

**ID:** Label for each graph.

**Number of graphs:** Number of graphs to display.

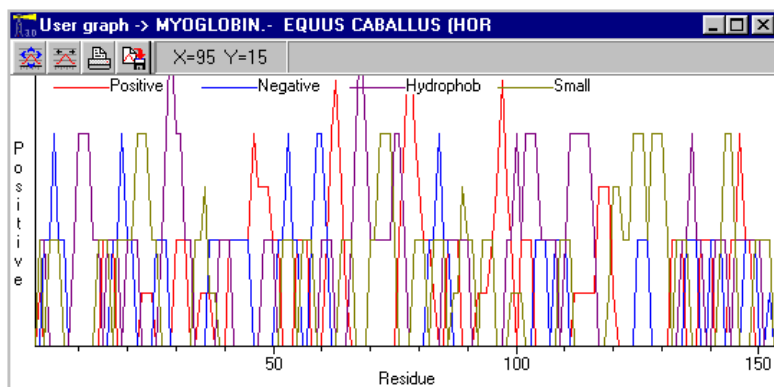
**Scaling factor:** Not used at present.

**Scale:** Initial y-scale of the graph.

**Smoothing:** Number of values (x-axis) that are averaged in the display.

**Reset:** Clears the table.

**Load / Save:** Loads and saves a table to a disk file.

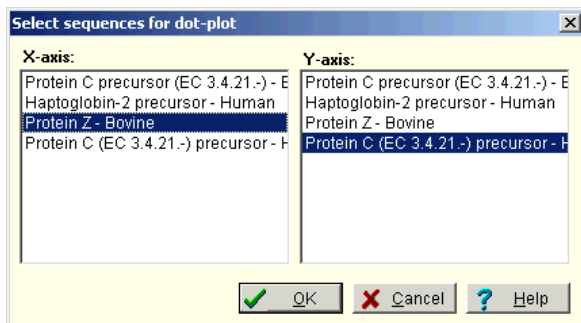


The common graph commands are available for the user graphs, please see above.

## Dot-plot

11.6

The dot-plot graph is used to compare two protein sequences by plotting identical or homologous residues in a two-dimensional pattern. Unlike the other graphs discussed, the dot-plot graph cannot be zoomed with the mouse, and the local menu is also slightly different. The function can also be used for comparing a sequence against itself in order to check for internal repeats (select the same protein for both axes).

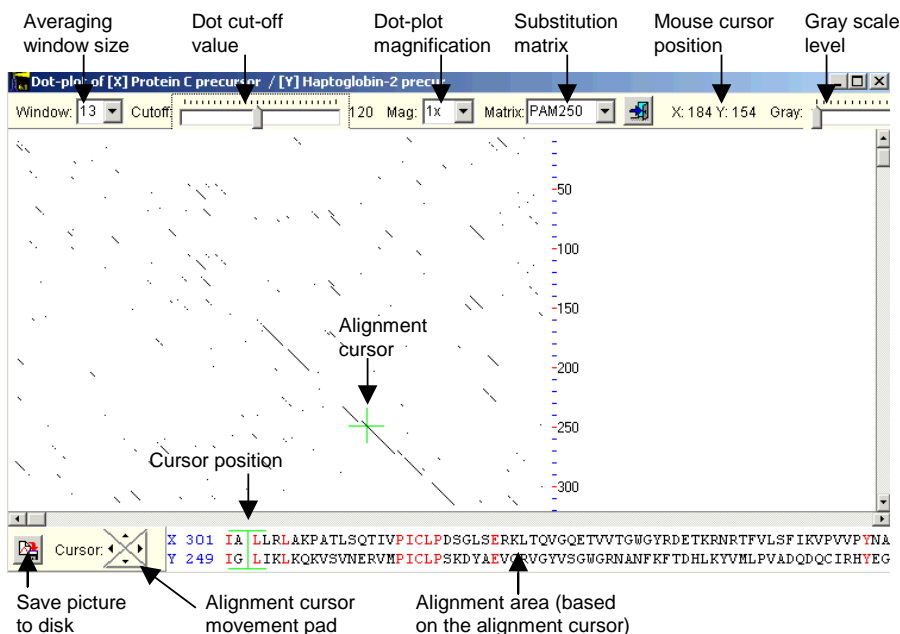


## 11 – Graphs

When selecting **Graph | Dot-plot** you open a dual list dialog box. Each of the list-boxes presents a list of all sequences opened on the desktop. You select one sequence from each list-box and press **'OK'**. You can select the same sequence in both list-boxes if you want to make a dot-plot of a sequence against itself.



**Note:** Only sequences opened on the desktop can be selected for display in the Dot-plot window.



The dot-plot is an arrangement of one sequence along the top and one down the left side. In the graph the two sequences are compared, and wherever the alignment between the two sequences is 'good enough' a dot will be marked on the graph. If the two sequences then show an alignment along part of their sequences, it will show up on the graph as diagonal lines.

Residue numbering is shown along the right and bottom edges of the dot-plot.

As the level of similarity may be quite weak, a number of options are available to fine-tune the comparison, all placed along the top of the window:

**Window:** Determines how many residues (1-29) are compared. Use a low value for very similar sequences and high values for divergent proteins. You can only choose uneven values as the dot is placed in the center of the comparison.



## 11 – Graphs

**Cutoff:** When comparing sequences, the matches are scored according to the substitution matrix used. Above the given cut-off value the dot is set (black), below this value no dot is placed. The value of the threshold is shown to the right of the slider.

**Magnification:** Enables to see a 2x magnified view of the alignment.

**Substitution matrix:** Select either the PAM250 or Identity matrix. PAM250 is based on the observed substitution of closely homologous proteins. The identity matrix scores 1 for each identical residue in the comparison.

**X-Y position:** Position of the mouse cursor in the dot-plot.

**Gray scale level:** When the slider is moved from the left-hand position, the scoring system changes from an on/off to a gray scale value. In some cases this yields a clearer diagonal. Use it in combination with the 'Cut-off' value.

The alignment of the two sequences along a given diagonal can be displayed by clicking on the graph. This will place a green cursor on the dot-plot that can be moved either with the cursor keys, the mouse (click) or the 4-way movement pad at the bottom of the window. The actual alignment is shown in the 'Alignment area'. Identical residues are colored red and the alignment position is shown as a green vertical bar. The residue numbers to the left of the alignment is the first position to the right of the green bar. The alignment can be copied to the clipboard by right-click in the alignment and select 'Copy'.

The graph can be saved to disk in bitmap (.bmp) format through the bottom left-hand button.

The **pop-up** menu enables you to move the dot-plot to either top, bottom, left or right edge of the alignment. This can be very handy when comparing two large proteins.



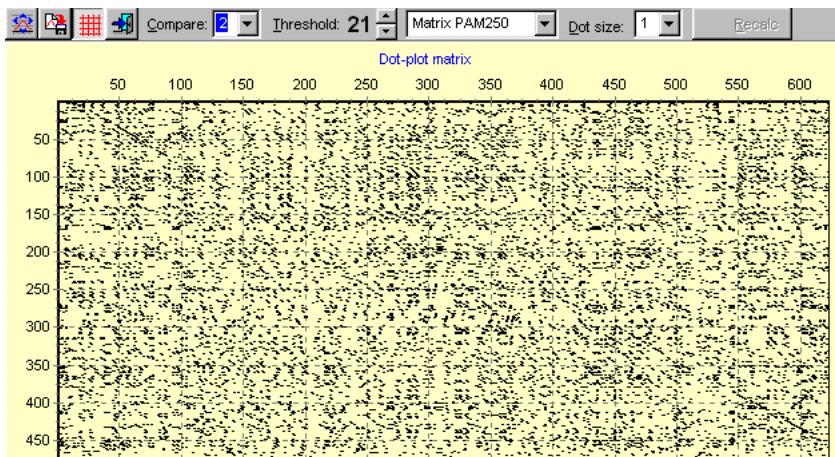
**Hint:** In the normal mode (x1) it can be difficult to point exactly to the first residue in a diagonal line, but it is usually easier to get an overview of the alignment. Switch to enlarged mode (x2) to make it easier to position the mouse.

### Dot-plot chart

The dot-plot chart is a variant of the dot-plot discussed above. The algorithm is the same, but the display does not have a 1:1 ratio to the screen pixels. This means that you can show the complete sequence in a small graph.

However, the calculations for the bit-map is much slower, and for large sequences on a slow computer it may turn out to be unrealistic to perform a thorough analysis. In this case you should get the right parameters from the dot-plot above and transfer the parameters to the dot-plot graph afterwards.

## 11 – Graphs



The graph functions slightly different from the other graphs. The toolbar shows from left to right:



- Set full graph.
- Save graph to disk.
- Toggle grid.
- Close dot-plot chart window.
- Number of residues to compare (e.g. window size).
- Threshold level for setting a dot (approx. 12 for each residue to compare depending on comparison method).
- Comparison method – direct; PAM250 and chemical similarity.
- Dot size (1, 2 or 3 pixels wide).

The **'recalc'** button will be disabled until at least one parameter is changed, then it will show a green light indicating that you need to force a recalculation of the dot-plot in order to see the effect of the changed parameters. This is done so you can change several parameters for each recalculation.

You can zoom parts of the dot-plot by click and drag the mouse over the part you want to zoom in on. You can go back to full size by clicking on the left-most button in the toolbar.

The dot-plot can be copied to the clipboard by right-clicking in the graph and select copy to clipboard from the pop-up menu, either as bitmap (copy to bitmap graphics program, e.g. paint) or as metafile (copy to vector graphics program, e.g. Corel Draw) – Word accepts both forms.

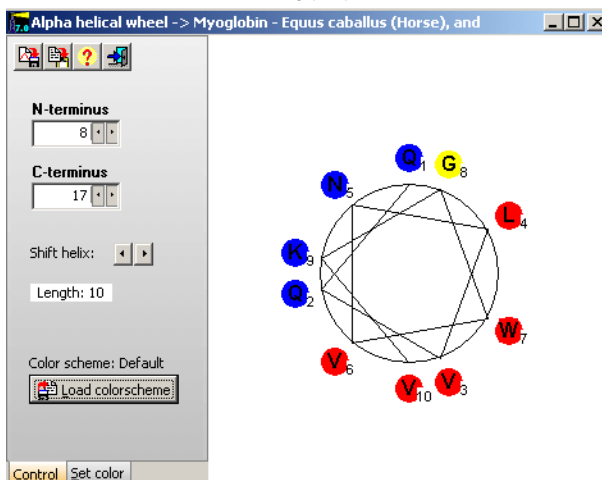
As the interior of soluble proteins is hydrophobic, alpha helices that are located on the surface of a protein will in most cases be amphiphatic (i.e. one side will be hydrophobic, interacting with the interior of the protein, and the other side will be hydrophilic, interacting with the environment). By plotting the amino acid residues in the helix onto a circle with a spacing of  $100^\circ$  around the periphery it is possible to represent the helix as viewed end-on, and a possible amphiphaticity can be estimated.

This graph can with advantage be used in combination with the 'Secondary structure prediction' graph, Ch. 11.2.

The **alpha helical wheel** function of GPMW is linked to a protein sequence and can be selected from the Graph section of the main menu. The function will draw a schematic representation of a section (or all) of a protein as an alpha helical wheel.

The window opens with a control panel on the left and the helical wheel on the right. In order to more clearly show hydrophobicity, the different residues are colored according their hydrophobicity (default color scheme):

- K, R, D, E, Q, N : Dark blue (charged and very hydrophilic residues).
- H, T, S : Light blue (hydrophilic residues)
- G, P, M, C : Yellow (neutral residues)
- A : Magenta (slightly hydrophobic residue)
- V, L, I, F, Y, W : Red (strongly hydrophobic residues)



The 'helix' will be numbered with the first residue in the helix as number one and located at the top of the circle.

The figure below illustrates the f-helix from sperm whale myoglobin and clearly shows a hydrophobic part (right-hand side) and a hydrophilic part (left-hand side).

## 11 – Graphs

### Commands:

**N-terminus** and **C-terminus**: The start and end of the helix. Values can be entered directly, or the mouse can activate the up/down arrows. If you by accident enter a number in the N-terminus field that is larger than the value in the C-terminus field, the helix will be drawn from the C-terminus value to the N-terminus value.

**Shift helix**: The left/right arrows will shift the N- and C-terminus synchronously up and down one residue, keeping the length of the helix constant.

The current length of the helix is shown below the 'Shift helix' command.

**Color scheme**: Different color schemes (tables containing the color of each residue) can be implemented. Click on the 'Set color' tab at the bottom of the command panel to get a table of residues colored according to the current color scheme (right). Double-click on a residue name to edit the color through a standard color dialog. A modified color scheme can be saved to disk for later use.



Save the alpha-helical graph to disk as a bitmapped file (bmp format). All graph-handling programs that can read bitmapped files can read this file format.



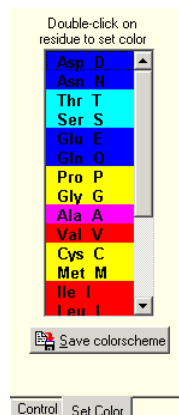
Copy to clipboard. The alpha helical graph is copied as a bitmapped image.




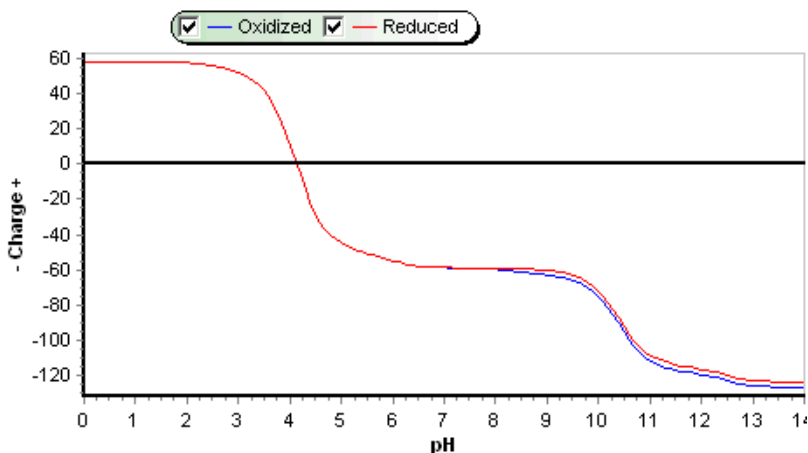
On-line help



Exit. Close the alpha helical wheel window.



The Charge vs. pH graph  is similar to the graphs of the same name generated from the peptide list (Ch. 9.4). The graph shows the charge of the protein at given pH values.



Two graphs are displayed, one shows the charge of the protein with oxidized cysteine residues (i.e. as disulfide bridges) and the other graph shows the charge of the reduced protein. The difference of the two graphs is due to cysteine having a pI of approximately 8.3. Below this value the two graphs are identical. Either graph can be turned off by un-checking the relevant graph in the legend.



**Autoscale:** When the graph opens, the scale is set to +/- 25 charges. Activating the autoscale button will rescale the graph to the maximum and minimum of the complete graph.



**Import other sequences.** This function will import other sequences from the GPMW desktop into the same graph window.



**Charge number list.** When the button is activated a frame is opened in the right-hand side of the window with the charge vs. pH as a list of numbers (see right). The 'granularity' of the pH values can be set to 0.1, 0.5 and 1.0 pH units, either through the pop-up menu or the drop-down list at the bottom the list. You can copy the list to the clipboard through the pop-up menu.

pH	Net charge red	- ox
0.0	85.00	85.00
0.5	84.99	84.99
1.0	84.97	84.97
1.5	84.90	84.90
2.0	84.68	84.68
2.5	82.21	82.21
3.0	77.96	77.96
3.5	65.03	65.03
4.0	26.31	26.31
4.5	-22.07	-22.07
5.0	-43.77	-43.77
5.5	-51.67	-51.67
6.0	-56.80	-56.80
6.5	-60.64	-60.56
7.0	-61.79	-61.55
7.5	-62.47	-61.83
8.0	-64.03	-62.03
8.5	-67.67	-62.23
9.0	-71.15	-63.95
9.5	-75.14	-67.38
10.0	-85.62	-77.70
10.5	-110.47	-102.47
11.0	-131.69	-123.69
11.5	-138.95	-130.95
12.0	-143.10	-135.10
12.5	-151.00	-143.00

0.5 pH unit

## 11 – Graphs

The main functions of the graph is to inform the user of the behavior of the protein during ion exchange and isoelectric focusing. Furthermore, the slope of the graph when it passes through zero (i.e. the isoelectric point) will indicate the 'stability' of the isoelectric point. With a weak slope, very minor changes in the three-dimensional structure of the protein can be expected to change the pI while a steep slope indicates that major changes in the protein (i.e. neutralization of several charges) need to be take place before a change in pI is experienced.

### DigestAlyzer

11.9

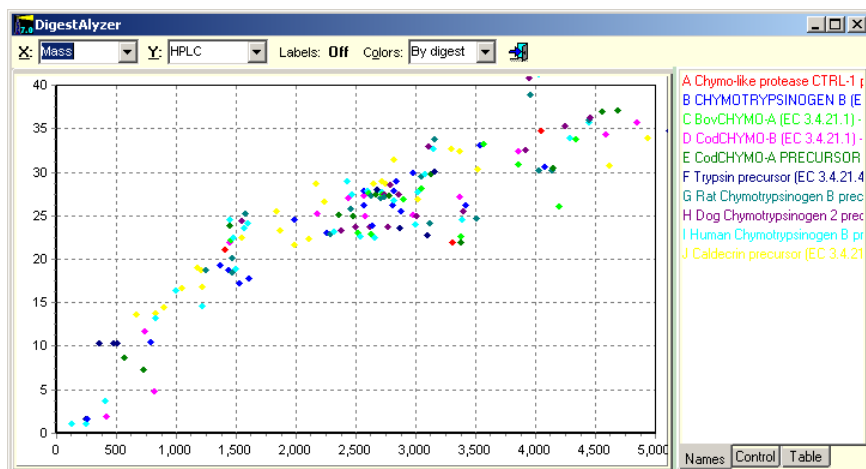
The DigestAlyzer is a graphical way of comparing a number of protein digests (e.g. peptide lists) for two parameters that can be either of: Mass, HPLC index, pI or hydrophobicity. It is particularly useful if you have a number of closely similar proteins where you need to separate the peptides.

Start by opening all the relevant sequences on the GPMaw desktop.

The most common procedure is then to select the 'Digest all sequences' form the 'Quickdigest' menu (if relevant you should first check the '1 missed cleavage') then you select the appropriate enzyme from the 'Quickdigest' menu (Chapter 9.1).

Alternatively you may cleave each protein in the usual way, using the same or different enzymes and parameters).

To activate the DigestAlyzer, you select any of the peptide windows, and in the '**Peptide list**' menu you select the '**DigestAlyzer**' option, or you may right-click and select it from the pop-up menu.



The graph will initially be displayed with mass along the x-axis and HPLC retention index along the y-axis. In the toolbar you can select the X- and Y-drop-down selection boxes to change either axis to show mass, hplc index, pI or hydrophobicity (the pI calculation method is chosen in Setup, Chapter 5.1).

When in the 'By digest' mode, each peptide will be color-coded according to the name label shown in the right-hand box (cannot be user-defined). In the

## 11 – Graphs

'By count' mode, the color will be dependent of the number of peptides present in each spot (see below).

The graph can be zoomed in the way normal for graphs, Chapter 11.1.

Each dot representing a peptide can be labeled by clicking the '**Off**' button in the toolbar. Turning the labels on is usually only preferred when showing only a small portion of the graphs.

**By digest:** Each protein digest is shown in each own color.

**By count:** Identical peptides will be colored an increasing shade of red depending on the number of peptides in a given spot. This means that if you are comparing proteins that only vary in a few positions, these positions will be black while the identical ones will be red (and perhaps have a different shape, see below).

The right-hand panel have three pages

**Names** **Control** **Table**, of which the first two controls the x/y graph and the third controls a list of peptides:

**Names:** Display the names of the analyzed proteins.

**Control:** Modify the display of the dots (see right):

**Expand X-/Y-axis:** Expands the relevant axis by 50% (i.e. mass from 5000 to 7500).

**Point size:** Select the point size in pixels (2-5).

**Color change:** How fast does the color change from black to red when in 'By count' mode. The color bar above the selection shows the color at any given identical peptides.

**Dot change:** How many identical dots before changing dot shape from diamond to square. The control is most easily controlled by selecting it by mouse and then use the right/left arrow keys to change the value.

Whenever you make a change to any of the parameters on the 'Control' page, you have to press the 'Update' button for the changes to take effect. This is indicated visually, as the button becomes active (changes color).

The pop-up menu gives you access to the **Print** command and copying of the graph to clipboard in either bitmap (Ctrl-C) or vector format. Furthermore, you can copy the mass list to the clipboard, sorted by protein, mass or hplc index.

**Table:** The 'Table' page displays a list of all the peptides participating in the graph. Based on the mass of the peptide it is possible to select either 'Different peptides' or 'Identical peptides', the differentiation is based on the value entered in the 'Max. diff.'

Show  
☒ Different peptides  
☐ Identical peptides

Max. diff. 0.010 Da.  
Total: 1012

A 119-E9  
B 119-B9  
C 119-B11  
D 119-C7  
E 119-C10  
F 119-D2  
G 119-D11  
H 119-E8

☐ Expand X-axis  
☐ Expand Y-axis

Point size: 3

By count:  
0 5 10 15 20 25

Color change  
☐ Slow  
☒ Medium  
☐ Fast

Dot change: 2

Update

## **11 – Graphs**

edit box. 'Total' shows total number of peptides in the lists (i.e. 'different' + 'identical').



## Utilities

This section contains a few utilities that do not fit into other categories.

Several of the functions presented here (e.g. mass comparison and composition calculator) are usually called as auxiliary functions from other windows, but as each function can sometimes be useful on its own, they are presented here as individual windows.

### MS peak analysis - sequence tag and mass difference

12.1

The **MS peak analysis** command contains two functions that both centers around the analysis of peptide mass differences, either in the form of residue differences, or in the search for modifications. Both functions have the same input, a mass table data input.

#### Data input

You start by entering the peak data (mass list) MALDI ms peak file format (among others) or you can paste most formats from the clipboard. GPMW's own PEP format is of course also supported.

Once the data is entered you can edit the numbers (see Chapter 4.1 for an explanation of the **'Edit'** button), copy to clipboard or save to a disk file in PEP format just like data entry in the 'Mass search' (Chapter 6.1) and 'Digest database search' (Chapter 8).

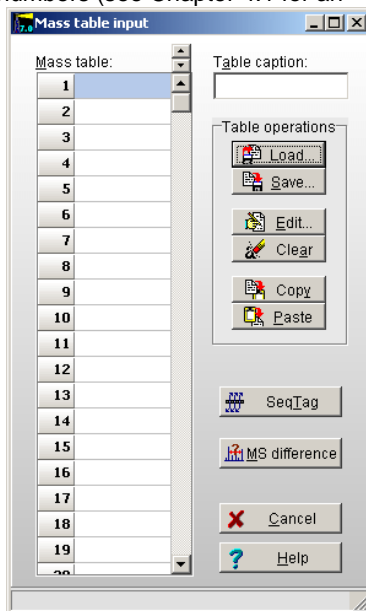
The table information is automatically set to the file name of the most recent file read.

The local pop-up menu supports the same functions as the right-hand buttons.

The entered mass list can be used either for extracting sequence tags (the **SeqTag** button) or to look at mass differences (the **MS difference** button).

The table accepts a maximum of 150 entries.

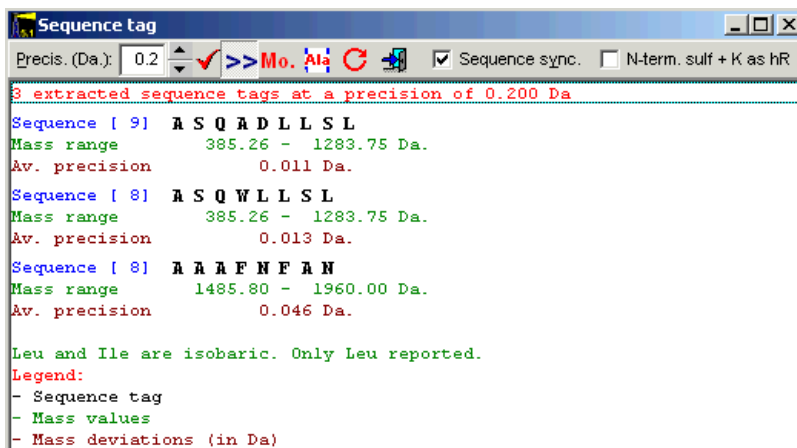
#### Sequence tag



## 12 – Utilities

The sequence tag window will extract sequence tags from the mass list. All sequence tags that are different will be displayed.

Two display formats are possible, the compact format



and the expanded format

21 extracted sequence tags at a precision of 0.200 Da

9	Ala	Ser	Gln	Ala	Asp	Leu	Leu	Ser	Leu	
	385.26	456.29	543.33	671.40	742.43	857.46	970.53	1083.60	1170.66	1283.75
	0.007	-0.008	-0.011	0.007	-0.003	0.014	0.014	-0.028	-0.006	
8	Ala	Ser	Gln	Ala	Asp	EP	Ser	Leu		
	385.26	456.29	543.33	671.40	742.43	857.46	1083.60	1170.66	1283.75	
	0.007	-0.008	-0.011	0.007	-0.003	-0.045	-0.028	-0.006		
8	Ala	Ser	Gln	Ala	MP	Leu	Ser	Leu		
	385.26	456.29	543.33	671.40	742.43	970.53	1083.60	1170.66	1283.75	
	0.007	-0.008	-0.011	0.007	-0.007	0.014	-0.028	-0.006		
8	Ala	Ser	Gln	Trp	Leu	Leu	Ser	Leu		
	385.26	456.29	543.33	671.40	857.46	970.53	1083.60	1170.66	1283.75	
	0.007	-0.008	-0.011	0.019	0.014	0.014	-0.028	-0.006		
8	Ala	Ser	Gln	Trp	Leu	EP	Ser	Leu		
	385.26	456.29	543.33	671.40	857.46	970.53	1083.60	1170.66	1283.75	
	0.007	-0.008	-0.011	0.019	0.014	0.014	-0.028	-0.006		

Both formats lists each sequence tag on a three-line display. The compact format shows:

- Line 1 : The sequence in 1-letter code.
- Line 2 : Mass range (lowest to highest mass).
- Line 3 : Average precision of the mass differences.

The expanded format shows:

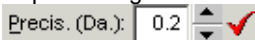
- Line 1 : The sequence tag in 3-letter code.
- Line 2 : The mass values used for extracting the sequence tag.
- Line 3 : Either the mass difference or the difference between the residue tag and the mass difference.

Each sequence line starts with the length of the sequence.

## 12 – Utilities

The bottom of the display shows a legend for the color code of the sequence tag lines.

The precision for the comparison of mass values for the extraction of sequence tags is controlled by the **Precision field** in the toolbar



The value can be changed by using the mouse to change the up/down arrows or by directly entering a new value in the numeric field followed by >Enter< or click on the red **V**. The value of the precision field is saved by the program between runs.

The remainder of the **toolbar** controls the following:



Toggle between compressed and expanded mode.



Toggle between monoisotopic (red) and average (blue) mass calculations.



Toggles between display of residue tag difference and mass value differences (only valid in expanded mode).



Redo analysis – switches back to the data input dialog.

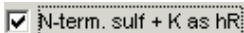


Close the window.



Sequence synchronization. When checked, the

topmost sequence window will be searched for any occurrence of the sequence tag lines which currently has the focus. The search will be performed in both sequence directions (N -> C and C -> N) and isobaric residues (e.g. Ile and Leu) will be considered.



N-term. sulf + K as hR

When this option is checked the N-terminal of the peptide will be considered as sulfonated and lysine residues will be considered derivatized to homo-arginine (e.g. derivatized according to the method of [Keogh et al.]). Furthermore, as the occurrence of Proline and Glycine in the sequence may lead to loss of sequence information (depending on type of instrument and sample amount), the program will try to fill out 'blanks' in the sequence tag with double residues containing Gly or Pro. These changes leads to the following changes in the above sequence tag extraction:

## 12 – Utilities

```

21 extracted sequence tags at a precision of 0.200 Da
Sequence [ 9] A S Q A D L L S L
Mass range      385.26 - 1283.75 Da.
Av. precision    0.011 Da.
Sequence [ 8] A S Q A D E P S L
Mass range      385.26 - 1283.75 Da.
Av. precision    0.014 Da.
Sequence [ 8] A S Q A M P L S L
Mass range      385.26 - 1283.75 Da.
Av. precision    0.011 Da.
Sequence [ 8] A S Q W L L S L
Mass range      385.26 - 1283.75 Da.
Av. precision    0.013 Da.
Sequence [ 8] A S Q W L E P S L
Mass range      385.26 - 1283.75 Da.

```

Notice that some 'double residues' have been inserted in the sequence. When viewed in the expanded mode, the double residue EP will be displays as E/P (e.g. not in 3-letter code).



**Note:** Isobaric residues may lead to problems in the identification.

**Isoleucine** is always shown as **Leucine**. Lys and Gln may be mixed up (not if Lys have been modified or trypsin is used as cleavage agent – except that the bond Lys-Pro is not normally cleaved by trypsin). Other residues are also close to each other: GG ~N, GE ~W, GA ~Q, GV ~R.

**Note:** If the precision is worse than ~0.5 Da you will also run into problems with residues separated by 1 Da.

The pop-up menu (right-click in the window) contains a few extra commands:

**Copy to clipboard (Ctrl+C):** Copies the whole table to the clipboard.

**Copy tag:** Copies only the currently highlighted tag to the clipboard. If you are in the compact mode display, you will get a 1-letter code sequence on the clipboard, which you can use to paste into search programs (see also below) or into the GPMW highlight residues (Ch. 3.3).

**Perform BLAST:** This command is present as two different commands: **as y-ion series** and **as b-ion series**. Both commands opens the local **BLAST search** dialog (see Ch. 7.2) with the selected sequence tag. When selected as y-ion series, the sequence will be inverted (read from right to left) while as b-ion series it will be read from left to right. The sequence can be edited before running the BLAST comparison.

**Print (Ctrl+P):** Prints the table in the display format presented on screen. Alternatively, you can use the corresponding print command from the menu or the main toolbar.

Copy to clipboard	Ctrl+C
Copy tag	
Perform BLAST (as y-ion series)	
Perform BLAST (as b-ion series)	
Print	Ctrl+P

### MS difference - X/Y table

Once the data has been accepted by pressing the **'OK'** button, a new dialog opens that shows the mass difference data.

## 12 – Utilities

	676.63	679.70	704.51	709.50	718.48	733.43	810.40	832.40	842.48	856.49	
679.70	3.07										
704.51	27.88	24.81									
709.50	32.87	29.80	4.99								
718.48	41.85	38.78	13.97	8.98							
733.43	56.80	53.73	28.92	23.93	14.95						
810.40	133.77	130.70	105.89	100.90	91.92	76.97					
832.40	155.78	152.71	127.89	122.91	113.93	98.98	22.00				
842.48	165.85	162.78	137.97	132.98	124.00	109.05	32.08	10.07			
856.49	179.87	176.79	151.98	147.00	138.02	123.06	46.09	24.09	14.02		
864.45	187.83	184.75	159.94	154.96	145.98	131.02	54.05	32.05	21.98	7.96	
870.51	193.89	190.82	166.00	161.02	152.04	137.08	60.11	38.11	28.04	14.02	
887.46	210.83	207.76	182.95	177.96	168.98	154.03	77.06	55.05	44.98	30.97	
932.46	255.83	252.76	227.95	222.96	213.98	199.03	122.06	100.05	89.98	75.97	
998.53	321.90	318.83	294.02	289.03	280.05	265.10	188.13	166.12	156.05	142.04	
1014.51	337.89	334.82	310.00	305.02	296.04	281.08	204.11	182.11	172.04	158.02	
1045.55	368.92	365.85	341.04	336.05	327.07	312.12	235.15	213.14	203.07	189.06	


Print    Precision: 1.6    Look for: 0.00 Da    Mo.    Amino acid    Sugar    Modificat.    ? Help    ✓ Done

The table shows the left hand column minus the top row.

The table by itself can be interesting, as it is relatively easy to spot repetitive occurrences of the same or similar mass. The real advantage comes when comparing to predefined mass tables as defined in the bottom status line.

Activation the **'Amino acid'** button highlights all mass differences that correspond to amino acid residue differences. The amino acid masses are taken from the currently loaded mass file. When you move the mouse cursor on to a highlighted field, a fly-by hint will open showing the name of the amino acid residue. The local pop-up menu allows you to switch the mass value of the highlighted fields to three-letter residue names. In cases where you have isobaric residues (Leu/Ile, Lys/Gln) both names will be shown.

The **'Precision'** box defines the precision by which the highlights are found. The box is in Dalton and can be edited directly or you may use the mouse to activate the up/down arrows.

The sequence tag button  will draw lines between successive amino acid residues, i.e. a sequence tag:

93,39											
LysGln	34,92										
138,29	44,90										
169,40	76,01	41,08	31,11								
185,46	92,06	Gly	47,17	16,06							
273,42	180,02	145,10	135,13	104,02	87,96						
284,45	191,05	Arg	146,16	Asp	Val	11,03					
299,36	205,96	171,04	161,07	129,96	Asn	25,94	14,5				
312,46	219,07	184,14	174,17	143,06	127,01	39,05	28,0				
322,42	229,02	194,10	184,13	153,02	His	49,00	37,5				

The **'Sugar'** button highlights mass differences corresponding to sugar (carbohydrate) residue differences. The mass values are taken from a

## 12 – Utilities

modification file (in the 'System' directory) called 'SUGARS.MOD'. The user can modify this file like all other modification files (see Chapter 4.3) and can in principle contain any data. In order to stay with the label of the button, you should keep it this way.

The **'Modificat.'** button uses the currently loaded modification file (Chapter 4.3) to search for mass differences.

The three mass difference types use the same precision but use different background colors for highlighting. Only the amino acid differences can be switched to show names in the table, the two others only show the names in the fly-by hint.

The bottom left button **'Print'** prints the table (across several pages if needed) and the button to the right of this toggles between showing average and monoisotopic mass values.

### Pop-up menu

The local pop-up menu has four entries:

- |                          |  |
|--------------------------|--|
| <b>Draw aa links:</b>    | Same as the 'Sequence tag' button.   |
| <b>AA as text:</b>       | Toggles between showing highlighted amino acid masses as text and as masses.         |
| <b>Compressed table:</b> | Toggles the displays the table into a compressed format in order to show more cells. |
| <b>Print:</b>            | Same as the 'Print' button.  |

### Composition calculator

12.2

The composition calculator is usually called from dialog boxes that demand input of compositions (i.e. 'Edit mass file', 'Edit modification file' etc.). Through the 'Utilities' section you have direct access to the dialog for calculating the mass of a given composition or to get the composition string in GPMW format.

For more information on the GPMW composition formula strings, see Chapter 4.4.

The Elemental composition dialog contains an edit dialog with a corresponding up/down spin control for each atom define in the 'Atomic masses' section of the 'Edit mass' dialog (see Chapter 4.2).

## 12 – Utilities

Elemental composition C2H4O1

Carbon C	2	Fluoride F	0
Hydrogen H	4	Iodide I	0
Nitrogen N	0	Potassium K	0
Oxygen O	1	Lithium Li	0
Phosphor P	0	Sodium Na	0
Sulphur S	0	Selenium Se	0
Bromide Br	0	Zink Zn	0
Chloride Cl	0	Iron Fe	0

Composition: C2H4O1 Clear

Mass average/monois. 44.0532 / 44.0262

OK Cancel Help

The number of each atom is controlled by entering a number in the edit boxes or by using the spin control with the mouse. Only integer values can be entered. The composition box is updated for every change made in the composition. The composition line can be copied (highlight and press <Ctrl+C> or use the pop-up menu) but cannot be edited directly.

Negative compositions or compositions that are part negative are accepted.

The **'Clear'** button zeroes both the composition and the edit boxes.

Both the average and monoisotopic mass is reported.

### Fragment analyzer

### 12.3

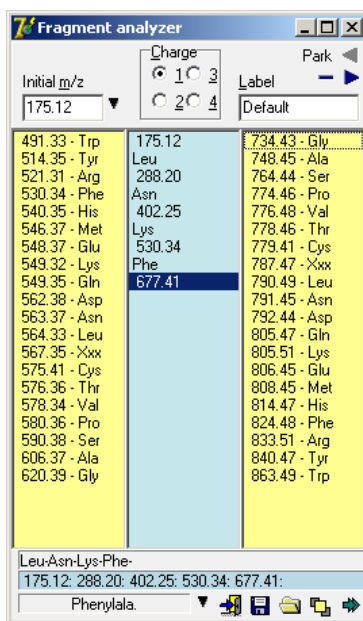
If you try to manually *de-novo* interpret an ms/ms spectrum of a peptide, you will usually be in for a lot of typing on your calculator. Even if your spectrum software is able to show mass differences on the screen, there can be a lot of mouse movements and calculations.

The Fragment Analyzer is a tool, which will help you manually calculate (assign) a series of fragment ions in a simple and flexible way.

The 'Fragment Analyzer' is a separate program ('Fragment.exe') called from GPMW. If the menu item (Utilities | Fragment analyzer) and the toolbar icon is grayed and inaccessible, GPMW has been unable to locate the file. Please copy the file 'Fragment.exe' to the \gpmw\bin\ directory and restart GPMW:

Working with the Fragment analyzer is quick and easy:

## 12 – Utilities



- 1) Start by selecting the charge state (top center) that you are working with. You can only work with a single charge state.
- 2) Enter the starting mass for the analysis in the 'Initial m/z' edit box (top left). This does not need to be the first or last mass value in the series, but is typically one of the major peaks that look to be part of a series. Press >Enter< to start the analysis.  
If you start with the y1 mass of a tryptic peptide, you can select m/z 147.1 (K) or 175.1 (R) through the drop-down arrow next to the input box.
- 3) The left-hand box will now show the mass values of all residues subtracted from the currently select value in the center list, and the right-hand box will show mass values of all residues added to this value. You now double-click on the mass value that fits with the next peak in your spectrum.

As you select a new mass value, the selection in the center box automatically jumps to that selection, going up or down.

If you want to cancel part of a series, just select the last entry you want to keep, and then click on the new next entry (up or down) – the rest of the series in the relevant direction will be deleted.

The bottom part of the dialog box shows the fragment series, the m/z series, the full name of the currently selected residue and buttons for 'Options menu', 'Exit', 'Save fragment', 'Load fragment' and 'Stay on top'.

The 'Options menu' is also available as pop-up menu (right-click) and contains the following commands:



## 12 – Utilities

*Stay on top:* The Fragment analyzer will stay on top of GPMaw, even when working in other parts of the program.

*Save fragment.* The fragment is saved in a file with the same name as the label.

*Load fragment.*

*Display in 1-letter code.*

*Display mass list.*

*Load Alt. mass list.* Enables you to load an alternative mass list among the ones used by GPMaw (by default located in the \gpmaw\system\ directory).

*Copy sequence.*


*Copy mass/peak list.*

*Print.*

The **label** is used as filename/title when saving the fragment to disk and when 'Parking' the fragment.

**Park:** When trying to interpret a complex ms/ms spectrum, you often get partial sequences in different parts of the spectrum. By using the 'parking' area, you can save the list temporarily by pressing the little right-arrow in the top corner of the dialog. This will store the present fragment in a list and make the left-arrow active. Press the left-arrow to open the list and retrieve a previously stored fragment. Click on the minus '-' button to clear the fragment list.

### Extended display mode

Clicking on the bottom right button  'Extended display mode' will enlarge the 'Fragment Analyzer' window to show the b/y equivalent column and the 'Extend sequence' list box.

**b/y ion equivalent:** An additional column with an input field labeled 'Parent ion' at the top is displayed to the right of the 'up mass residues'. Upon entering the parent ion mass of the peptide in question, the b/y ion equivalent series of mass values relative to the mass values in the center column will be displayed. Whenever a change occurs in the mass fragment column or the parent ion field, the list will be updated.

The b/y ion equivalent is the series that is complementary to the one entered in the center column. I.e. if the series entered in the center column originates from the y ion series, the list presented here is the corresponding b ion series. Seeing both the y and the b ion in a spectrum gives more confidence to the assignment.

**Extend sequence:** When you try to extend a sequence tag, you often encounter a gap in the sequence information, i.e. the distance to the next residue is larger than the mass of a single residue, and you may have to resort to a table of multiple residues

## 12 – Utilities

The 'Start value' box contains the currently selected mass value from the mass fragment column. When you enter a value into the 'Extend to' box, all possible combinations of amino acid residues that fit to this difference within the precision defined below the central table will be listed in the table. Each suggestion of residues will be listed with calculated difference, deviation from the actual difference and composition. The suggestions will be listed relative to the deviation. If you double click on a suggestion or press the 'Transfer' button, the selected residues will be inserted into the main fragment analyzer.



**Note:** The residues will be inserted in the order they are shown in the 'Extend sequence' box, which may not correspond to the actual sequence.

If you click on a mass value in the main Fragment analyzer, the value will be transferred to the 'Start value' box. The 'Start' and 'Extend' will also be calculated if the 'Extend value' is lower than the 'Start value'.

When you have a suggestion for a complete sequence, you can enter it in the ms/ms fragmentation dialog to get a list of all the normally seen fragmentation mass values (Chapter 10.1).



**Note:** Isoleucine (Ile) is not present in the up/down mass lists as it is isobaric with leucine. When doing sequence/homology searches you have to be aware that a hit on Ile is equal to that of Leu. As the two residues have similar physical/chemical properties it is usually not of great concern.

**Database indexer (DBindex)****12.4**

The database indexer, called DBindex, is a separate program that is either bundled with GPMaw or can be freely downloaded from the Lighthouse data website (see Chapter 1.9). As the program is separate from GPMaw it means that once called you can run it independently (i.e. switch back and forth without one program interfering with the other). The version numbers are separate from GPMaw and you should consult the web site or Lighthouse data for the latest version. Current version, October 2002, is 1.20. The program (DBINDEX.EXE) has to be present in the same directory as the GPMaw executable file (GPMaw3.EXE) in order to be called from the menu.



**Note:** If the menu entry **Utilities | Call DBindex** is disabled (grayed) it is because GPMaw could not find the database indexing utility. You should then copy the program and the help file (DBINDEX.HLP) to the GPMaw\BIN directory.

When the program opens, you are greeted with a dialog stating copyright and version number.



From the copyright dialog you can select what section of the DBindexer to enter. Independent of the section you choose here, you can switch between sections on different tabs in the running program.

**Quick conversion:** This function will convert, name index and create BLAST indices in a single operation. This option is recommended for most users when indexing a database. The combined functions available here can be found as individual functions on the 'Index' and 'Convert' pages.

**Index:** Create indices from a FastA formatted database that enables GPMaw to search the database on the basis of protein name or accession number.

## 12 – Utilities

Combine databases (i.e. add one database to the end of another).

**Convert:** Convert the Swiss-Prot database to FastA format and still enable retrieval of the full Swiss-Prot entry through GPMW. Reduce the complexity of some non-redundant databases that create very long (>255 character) name lines (NCBI-nr and EMBL-nr).

Convert files in VMS format to DOS (Windows) format. The VMS format is used on most UNIX systems and needs to be converted to for GPMW to access the databases.

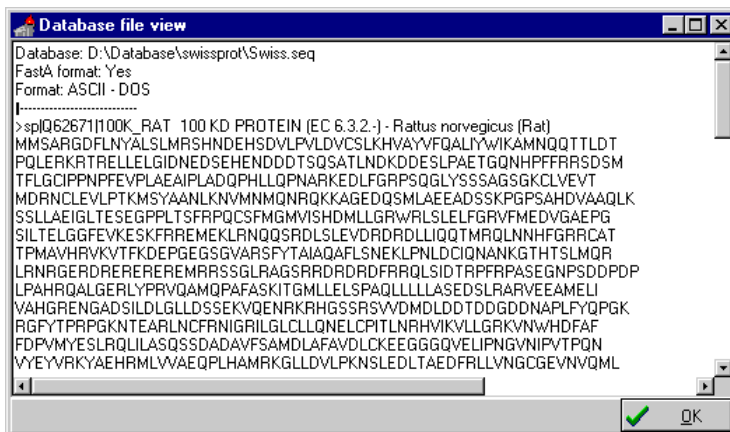
**Filter:** Filter a FastA formatted database with regard to amino acid composition and molecular size.

**Other:** The program opens on the first page of the window without asking for a database.

The 'Index', 'Convert' and 'Filter' commands starts by asking for a database file to work on. If you later need to open a new database, just click on the 'Load database' button in the bottom command line.



When you have opened a database, the next button in the command line, 'Db Peek', becomes active. Pressing this button opens a text dialog with the first couple of thousand characters of the database. This enables you to check the format and content of the database.



In the database file viewer (above) you are only able to only view the content of the file, you cannot edit it.

The view starts with the full file name, then comes whether the file conforms to the GPMW FastA format, and finally what type of file is present (ASCII vs. Binary format, DOS vs. VMS format).

After the stippled line, comes the first records of the database.

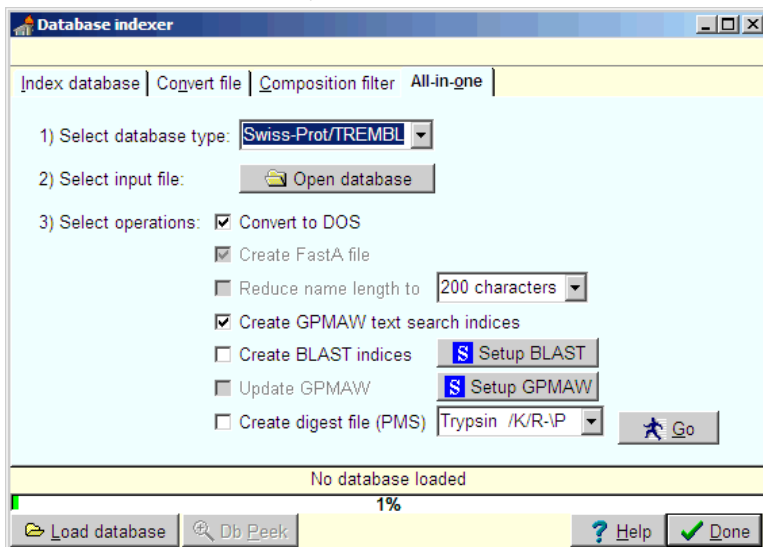
Click on 'OK' to return to DBindex.

## 12 – Utilities

### Quick conversion (All-in-one).

The All-in-one conversion page enables you to

- 1) convert a database from most FastA formats to a general format accepted by GPMW (Swiss-Prot formatted files are first copied to a FastA format).
- 2) index the database for word searching by GPMW
- 3) create indices for use by BLAST homology searching – the database index can be automatically added to GPMW.



You perform the operation in four steps:

- 1) Select database type. If the database is FastA but in unknown secondary format, you select 'Other FastA'.
- 2) Open database. By default the Open dialog looks for files with the extension .seq, but database files often have the extension .dat.
- 3) Select operations.  
The first two options **Convert to DOS file format** and **Create FastA file** cannot be selected by the user, but are dependent on the choice made in 1). If Swiss-Prot format is selected here, a FastA format is automatically created and used for the later operations. If a FastA file format has been chosen, the file will be converted to DOS and GPMW accepted format.  
**Reduce name length to xxx** makes sure that the length of the name line does not exceed 250 characters (the limit for GPMW). This is often the case for the NCBI and EMBL non-redundant databases. In order to save space on your hard drive you can set this to a lower value (in most cases it is the same name repeated a large number of times with different accession numbers).  
**Create GPMW text search indices** will create indices that enables

## 12 – Utilities

you to search the database from GPMaw based on words in the name line.

**Create BLAST indices** creates index files for homology searching by BLAST. In order for this function to be enabled, the DBindex program needs to be informed of the location of the BLAST indexing program, formatdb.exe. This should be located in the \gpmaw\bin\ directory, but you have to press the **'Setup BLAST'** button and navigate to the correct directory. If the GPMaw program (gpmaw3.exe) is located in the same directory, the **Update GPMaw** option will also be enabled, otherwise you will have to locate the gpmaw3.exe program in a similar way.

**Update GPMaw** will enter the information of the BLAST database location in the GPMaw ini file so the program will know about it when it is next started. **Note:** GPMaw must not be running while you index the database, otherwise this information will be lost!

Create digest file for Peptide Mass Searching. In addition to checking this option, you have to select the enzyme for which to make the digest.

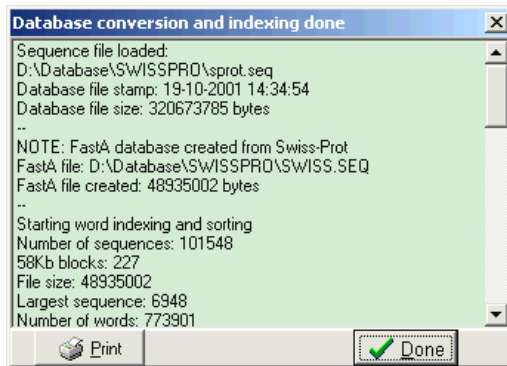
- 4) Press the **'Go'** button to start the conversions. Particularly the creation of search indices takes time, and for a large database the whole operation may take ~10 minutes.

The operation of the individual functions can be followed on the status bars at the bottom of the window. **Note:** The BLAST indexing is done by a console application, you should not close the program, or perform other operations while it is in action.

Upon completion of all operations you will be presented with a dialog detailing the conversion operations along with some statistics. The list may be printed for your records. Part of the information can also be found in the \*.FAC file created with the database.

The operations performed during **'Quick**

**conversion'** are the same functions as can be performed from the **'Index database'** and **'Convert file'** pages of DBindex as detailed below.



### Indexing a FastA formatted database

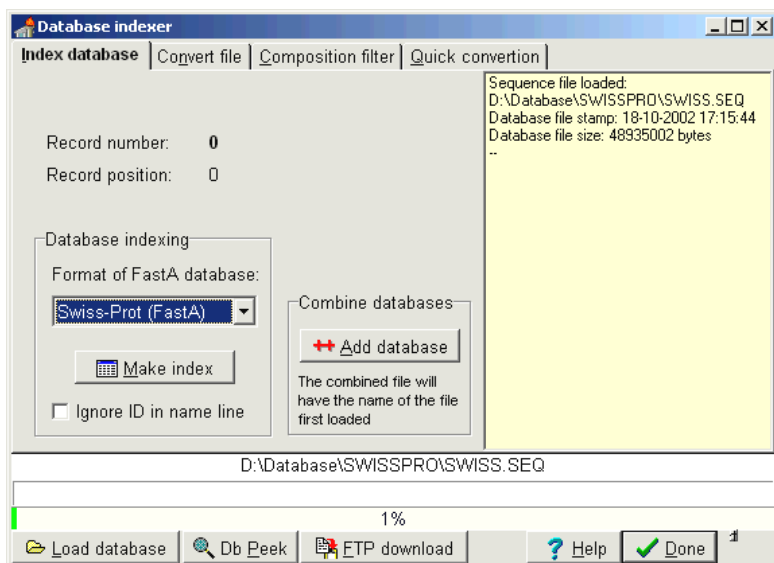
This function generates indices used by GPMaw for searching for sequences based on words in the name line of the FastA records.

Start by loading a FastA formatted database. If you have not selected one when opening the program, or if you want to select a different database, press the **'Load database'** button and select a new database. The database loaded will be shown in the yellow list box in the right hand part of

## 12 – Utilities

the dialog and in the status bar above the progress bar at the bottom of the display.

In the drop-down list 'Format of FastA database' you select the kind of database loaded.



Press the '**Make index**' button and database indices will be generated. The progress of the index creation can be followed in the progress bar, the 'Record number' and the 'Record position'. The yellow list box will show statistics and files generated.

When the '**Ignore ID in name line**' option is checked, the indexing function will ignore accession numbers and treat the name line as a single line. This is most useful when indexing 'homemade' database that do not contain accession numbers.

The index files generated will be placed in the same directory as the database.



**Note:** Make sure that you have sufficient space on your harddisk as the index files a quite voluminous. E.g. the current version of EMBL-nr takes up 280MB of space (after conversion from VMS to DOS and reduction of name lines) while the indices additionally take up almost 30 MB of space.

Four files are generated, each characterized by its extension:

- .ACC      Accession number
- .FAC      Facts file. This is a text file that contains general information on the database and the indices.
- .NDX      Index into the main database.
- .TRG      Target database. Contains the search words.

## 12 – Utilities

The user should not modify any of the files generated. Only the FAC file can be of any use and can be read into any text editor.

When the program finishes indexing (a rather lengthy procedure ~20 min.) you can copy all databases to a different drive or medium (e.g. CD-ROM) for easier access. You may choose only to copy the indices and leave the database as such on a networked or slow drive. In this case you should modify the reference to the original database in the FAC file using Notepad.

### Combine databases

The command '**Add database**' enables you to combine two databases (or any files) into a single database. This can typically be used with the PIR database (comes as 4-5 separate files), genome databases (are often present as a database for each chromosome), the TREMBL database (a database for each species) or the Swiss-Prot and the GenPept that are published as a main database with an update.

Start by loading a database. Press the 'Add database' button and you will be asked for a database to add. When you select a file, it will be added to the original database (file).



**Note:** The combined database will have the name of the database first loaded. If you want to preserve this file you may make a copy of the first database and rename it to reflect the nature of the final database. The files added to the first file are not changed in any way.

When all the files have been added, you can proceed with indexing the database.

As the addition of files is a straight binary combination of files, it does not matter whether you convert from VMS to DOS before or after combining the files.

### References:

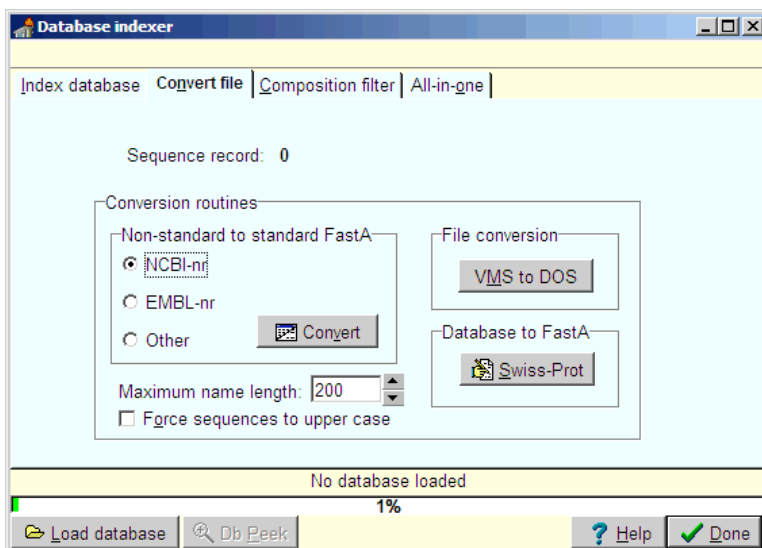
Keough, T., Lacey M.P., Youngquist, R.S. (2002) *Rapid Commun. Mass Spectrom.* Solid-Phase Derivatization of Tryptic Peptides for Rapid Protein Identification by [MALDI] Mass Spectrometry.



### Converting files

Several databases on the Internet are in a format that is not directly accessible by GPMW. In most cases the transformation is rather trivial (e.g. converting from VMS to DOS) and is used mainly to speed up access and simplify coding, while in the case of the Swiss-Prot, TREMBL and IPI databases, the information herein is so valuable, that it is of great value to access directly from GPMW.

Use the '**Convert file**' tab of the program to access these conversion routines.



### Reducing the complexity of a database

A few of the non-redundant databases (NCBI-nr and EMBL-nr) are created with name lines that exceed 255 characters, which is the limit for accepted names in GPMW.

You have to reduce these databases by selecting the relevant radio button in the '**Non-standard to standard FastA**' panel, and then press '**Convert**'.

You are then asked to open a database, and then to give a name to the new database. GPMW suggests a name depending on the selection of the radio-buttons above.

The new database will be placed in the same directory as the original database.



**Note:** Make sure you have enough disk space for the operation, as the new database will only be about 10% smaller than the original.

The '**Convert**' command also takes care of converting from VMS to DOS format if necessary.

## 12 – Utilities

When you are through converting the database, you can delete the old (original) database.

If you want to create indices from the newly created database you have to re-load it through the **‘Load database’** button.

### VMS to DOS conversion

Text files, and thus also flat file databases, are internally differently represented in UNIX and DOS (Windows). Where the DOS format specifies that each line ends in carriage return and line feed characters (#13#10), the VMS file system only specifies a line feed character (#10). Using the VMS to DOS file conversion routine takes care of this.

When pressing the ‘VMS to DOS’ button you are asked for the file to convert, and when accepted, the file is converted.

The new file replaces the old one (unlike the ‘Convert’ command above).

You can follow the file conversion process in the ‘File position’ label and in the progress bar.

### Convert Swiss-Prot to FastA

Press the **‘Swiss-Prot’** button and select the Swiss-Prot sequence database (typically named sprot37.seq for the main database or new\_seq.seq for the update). You are then asked whether you are converting the main, the update or another database. The default names for the converted database is SWISS.SEQ for the main and SWISSNEW.SEQ for the update.

If you keep all the files in the same directory, GPMW is able to search the FastA indexed database and retrieve the full entry in the original Swiss-Prot database, please see ‘Reading CD-ROM based sequences - FastA’ in Chapter 2.6 for details.



**Note:** The TREMBL and IPI databases are constructed similarly to Swiss-Prot and can be indexed identically (i.e. start by using the ‘Swiss-Prot’ button to create a FastA version of the database).



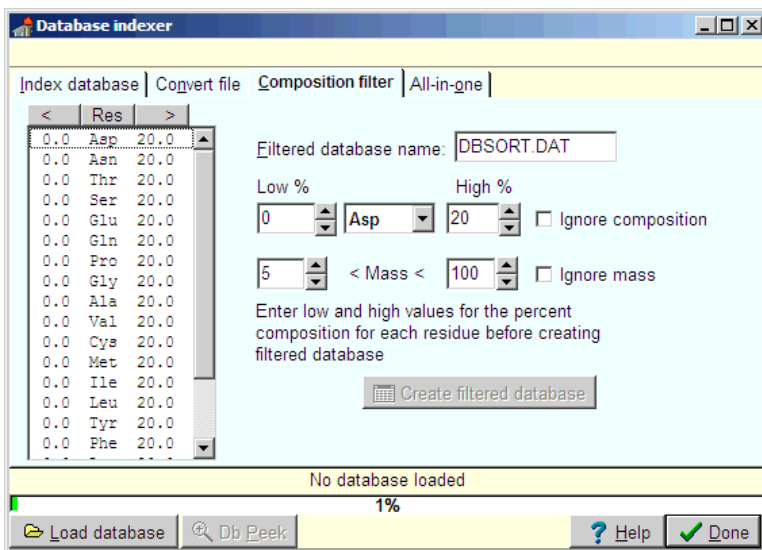
**Note:** The Swiss-Prot database is not free-ware. If you are a non-commercial organization you need a licensing agreement. Please see <http://www.expasy.hcuge.ch/announce/> for further details.

## 12 – Utilities

### Composition filter a FastA database

The '**Composition filter**' page of DBIndex enables you to filter a FastA formatted database based on amino acid composition and/or molecular mass.

The input is a database that has to be in FastA format. The operation is based on a composition range (in %) for each amino acid residue and/or a mass range specified by a lower and an upper limit. The result is a new database where each entry conforms to the filter specification.



- Select a database if the proper database has not been selected already (the currently opened database is shown in the bottom status bar).
- Enter minimum and maximum % values for each residue in the top edit boxes. The amino acid residues are selected from the left-hand list box whereupon the low and high % can be changed. The edit box is updated whenever the focus changes from an edit box. By default all residues are set to a composition between 0 and 20%.
- Enter low and high mass values (in kDa) for the proteins to be selected.
- You may check either of the 'Ignore composition' or 'Ignore mass' to generate a database that does not take composition/mass into consideration when filtering
- The name of the resulting database can be edited (by default DBSORT.DAT).
- Click on the '**Create filtered database**' to start the conversion process.

The filtering process can be followed on the progress bar.

## 12 – Utilities

After the filtered database has been generated it can be searched, indexed and viewed like a normal FastA formatted database. If the file is small enough (< 32.000 bytes for Win95/98) it can be viewed directly in Notepad, if larger it can be viewed in a word processor.

### Simulated 2-D gel

12.5

The simulated 2-D gel shows a graphical representation of a number of proteins presented as dots in a graph where the X-axis is the pI of the protein and the Y-axis is the mass. This is the typical setup when using 2-D gel electrophoresis.



**Note:** The pI calculated by GPMW is a theoretical calculation based on the input sequence and as such is quite approximate. You should be aware that the trimmings of signal sequences and other post-translational modifications have a considerable influence the pI. The mass of the protein is influenced to a lesser degree by modifications. Even when the protein contains no modifications the three-dimensional structure can influence both the pI and electrophoretic mobility in the gel.

The initial window of the 'Simulated 2-D gel' shows a display with a mass range of 10 – 200 kDa and a pI range of 3 – 10. Green lines show each pI value and 25 kDa mass. The 100 kDa mass line and the pI 7.0 line are shown in dark green.

The '2D-gel' can show either

1. the proteins present in a FastA formatted database
2. the proteins opened on the desktop
3. a combination of 1. and 2.

The parameters can be accessed through the toolbar and the right mouse popup menu.



The toolbar shows from left to right:



**Open database:** Select a database in FastA format. As each protein has to be read and the pI calculated the time to read a large database takes quite a time. The database proteins are shown as blue dots.



**Save graph to disk:** The graph ('2-D gel') is saved to disk in bitmap format. This can then be imported in a report or further modified in a graphics program. You can **copy to clipboard** through the main menu (**Edit | Copy**) or use the Ctrl+C keyboard shortcut.



**Import sequences:** Load all protein sequences that are opened on the desktop into the '2-D gel'. The imported proteins are shown as red dots.



**Set scale:** Enables you to redefine the display limits. This dialog can also be invoked by double-clicking in the 'gel' display area. You can zoom in

## 12 – Utilities

to part of the display by 'click-and-drag' the mouse cursor across the required part of the graphics.



**Display grid:** Turns the green grid on and off. The lines are drawn for every 25 kDa and 1 pI unit. 100 kDa and pI 7 are drawn in a darker color. The grid is on by default.



**Show tails as lines:** Shows N- and C-terminal trimmings as lines. N-terminal trimmings are green and C-terminal trimmings are red. Up to 300 residues are trimmed from either terminal. If you combine the lines with trimming dots (see below) you will be able to navigate by positioning the mouse cursor above the dots. This function works only on imported sequences.




**Exit.** Close the '2D-gel' window.

To the right of the command buttons you can select various options for the display



**Dots:** The dots are either single pixels or 2x2 pixel dots. The large dots are on by default. You will probably only need the small dots when displaying a very large database.

The following options will only work on imported sequences (**not** on database proteins) and are only enabled when sequences have been imported from the desktop. Each options works through a drop-down menu activated by

pressing on the down-arrow . You can get information on the resulting trimmed or 'phosphorylated' dots by resting the mouse cursor on top of the dot in question (see 'Information panel' below).

**Trim size:** Defines the number of residues cleaved from either end of the imported protein. The 'trim' parameter is combined with the 'Numbers' parameter below. Trim size can be defined as 1, 2, 3, 5, 10 or 20. The last option on the menu, '**Labels**', will turn name labels on and off for proteins imported from the desktop.

**Numbers:** The number of 'tails' shown, i.e. the number of 'trimmed' spots generated for each protein. You can specify 1, 5, 10, 20, 30 or 50 tails. With a 'trim size' 2 and 'tail numbers' of 5, the N-terminal tails of a 400 residue protein will be 2-400, 4-400 ... 10-400, and the C-terminal tails will be 1-398, 1-396 ... 1-390.

**Phospho:** Lets you simulate the addition of phosphorylations to the imported proteins. You can specify 1-4 phosphorylations. These will show up as dots trailing towards the acidic part of the '2D-gel'. Only the charge is considered in the 'gel' representation, as the mass will usually be insignificant. The 'phosphorylations' will only be carried out on the intact protein, not on the 'trimmed' proteins.

**Spot colors:**

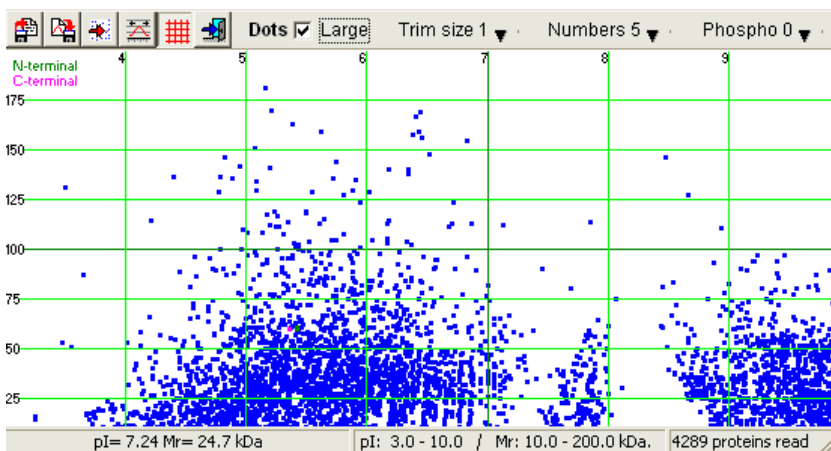
## 12 – Utilities

- **Blue:** Database proteins
- **Red:** Proteins imported from the desktop
- **Green:** N-terminal 'tails'
- **Pink:** C-terminal 'tails'
- **Light blue:** 'Phosphorylations'

**Information panel:** The bottom of the '2D-gel' window contains three information panels that show from left to right:

1. **pI and mass** of the mouse cursor position. If the mouse cursor points to an imported protein or one of the tail dots, the name and modification of that protein will be shown. If the cursor rests for a couple of seconds on the spot, the name will also be shown in a fly-by help window.
2. **pI and mass range** of the entire window.
3. **Number of proteins** imported from a database

You can **zoom** the view of the '2D-gel' in the usual way for graphs by 'click-and-drag' the mouse cursor across the required part of the picture in order to enlarge that portion. If you double-click in the graph area, you reset the graph to the default values.



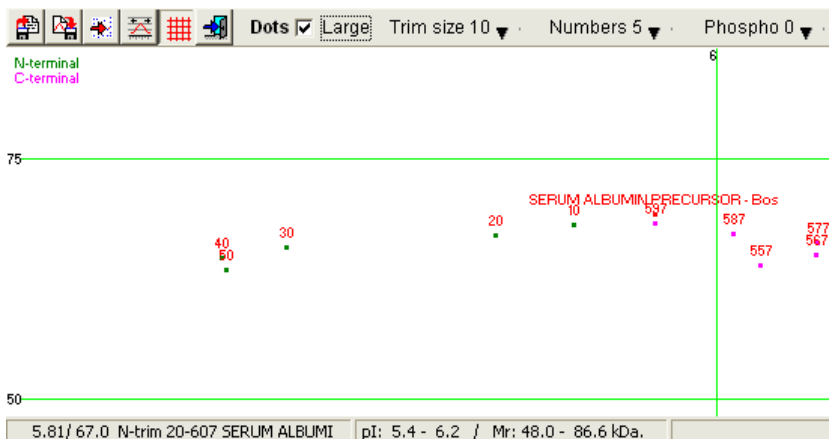
The picture shows the '2-D gel' window with the proteins from the E. coli protein database.

Any protein database can be displayed in the '2D-gel' as long as it is in FastA format. If you have problems reading the database, you may have to convert it from VMS to DOS format or the name lines may be too long (please see 'Database indexer' in section 12.4 and Appendix B for more information on how to treat FastA formatted databases).

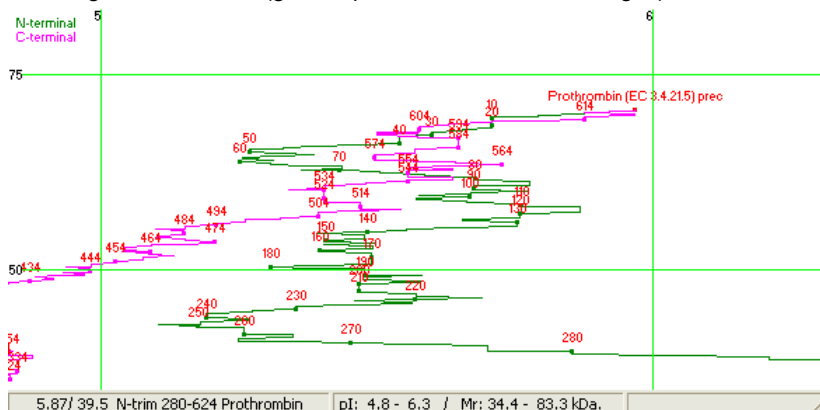


**Note:** Both the mass and the pI of each spot in the '2D-gel' are the result of theoretical calculations based on the sequence in the database. In vivo a large part of the proteins are likely to be post-translationally modified, either by trimming and/or by chemical addition/removal of groups, which can affect both pI and mass.

## 12 – Utilities



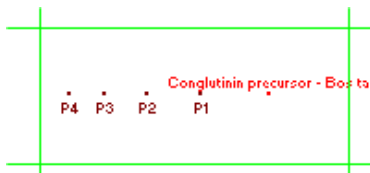
A zoomed view of the '2D-gel' after import of the prothrombin sequence. 5 N-terminal and 5 C-terminal trimmings of 10-50 residues are shown in the graph. N-terminal trimmings result in a move of the protein towards the acidic part of the 'gel' and C-terminal trimming results in a, smaller, move towards the basic part of the '2D-gel'. The legend in the lower left panel shows that the mouse cursor rests on top of the spot representing the N-terminal trimming of 20 residues (green spot labeled 20 in the '2D-gel').



Combining the above view with the 'Tails as lines' option gives you a view of how the protein 'travels' through a 2D-gel when being trimmed from either end. The dots of the tails can be seen as bulges on the lines. When the mouse cursor points to a dot, the trimming and sequence will be shown as fly-by help and in the bottom left-hand panel. The legend to each dot can be turned on and off through the 'Trim size | Labels' menu option.

If you **combine** imported sequences with a database, you should read the database as the last operation, as every other operation re-draws the database on the screen which can be quite time-consuming for a large database.

## 12 – Utilities



Using the **'Phospo'** button enables you to simulate up to four negative charges (phosphorylations) on the intact protein. These dots will show up as dots towards the acidic end of the '2D-gel'. As these spots can lie quite close to the intact proteins, their label will be printed below the dots (all other labels will appear just above their respective dots). The spacing of these dots relative to the 'native' protein gives a good indication of how 'resistant' the protein is towards single charge changes.

### Print

You can print the '2D-gel' by selecting **'File | Print'** in the main menu, pressing the 'Print' button in the main toolbar or selecting 'Print' from the pop-up menu. Only the displayed part of the '2D-gel' will be printed.



**Note:** Printing complex '2D-gels', particularly with N- and C-terminal tails should be done on a color printer, as a black and white print can be very confusing.

## Coverage analysis

12.7

Sequence coverage is a typical way of presenting sequence data, particularly when doing bottom up analysis, but other applications may demand the display of sequence features.

The coverage analysis of GPMaw has the following features:

- Up to eight different levels can be defined.
- Each level may contain up to 2000 peptides, each defined with first and last position in the sequence, a label (16 characters) and a comment (40 characters). Mass values can be displayed, but are calculated based on the currently selected mass file. Each level also has its own color for display.
- Coverage files may be saved and loaded from disk.
- Levels can be edited individually.
- If a level contains overlapping peptides, it can be split into two levels.

The sequence coverage analysis can be found in GPMaw Utilities|Coverage analysis. This opens a blank window, where you can either start to define the coverage manually, by defining the sequence and entering values manually. However, transferring data from other parts of the program is faster and usually very convenient. Transfer of data can take place in the following cases:



## 12 – Utilities

- 1) From the mass search results window (Ch. 6.1), you have a coverage on the 'Report' page. This coverage can be saved to disk or copied into memory for pasting directly into the coverage window.
- 2) From the peptide window (Ch. 9.3) you can copy the entire peptide list to clipboard as a coverage file. This can be a handy way of getting all peptides into the coverage file, as it is much faster to delete peptides, than it is to enter values manually.
- 3) If you have a table of peptide from-to values (e.g. a Mascot search) you may copy the entire table onto the clipboard and through a special import table feature, you can easily paste it as a new level in the GPMW coverage analysis.
- 4) The cleavage analysis routing (Ch. 9.3) use the same display engine and allows for transfer to the coverage window, either through a file on disk or through the clipboard.

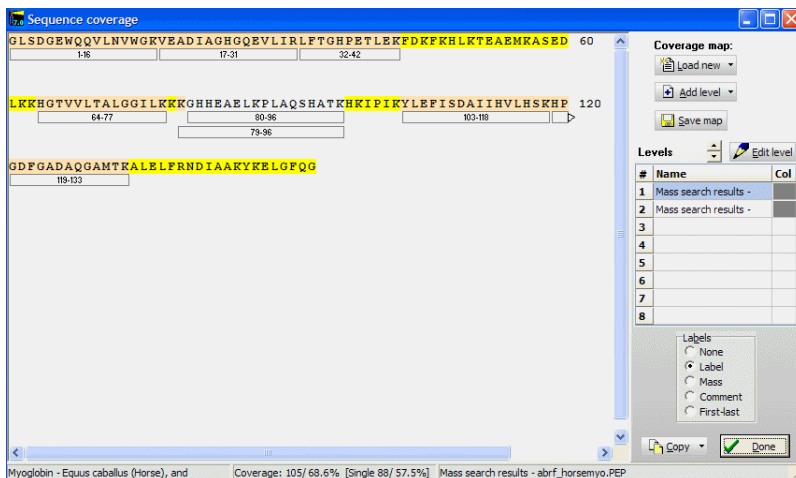
When you open the coverage analysis, you are greeted with a layout having a left-hand sequence window, a right-hand control panel and a status line along the bottom. The window is resizable, but the sequence is fixed at 60 residues per line.

You start by clicking the '**Load new**' button to read a coverage file from disk. If you want to paste a coverage map, you click the down-button and select '**From clipboard**' from the drop-down menu.

Once a coverage map is loaded (or pasted from the clipboard), you can add additional levels through the '**Add level**' button. If you click on the down-arrow, you are given the choice of adding through from a file on disk or from clipboard. The combined coverage map can be saved through the '**Save**' button.

The maximum number of levels that can be displayed is 8 (eight). Parts of the sequence which has no sequence coverage has a yellow background, single coverage is indicated by light orange and double-coverage (or more) is with a white background. If multiple peptides in a given level cover the same sequence, the bottom line of the colored box will show a red line.

## 12 – Utilities



The bottom status line shows from left to right:

*Name of sequence:* Name copied from the sequence header.

*Coverage:* first is shown coverage in residues and percent (e.g. 105/68.6%) then in parenthesis is shown the single coverage (e.g. how many residues are only covered once).

*Name of level:* Each level has a name that can be edited through the Edit level

Activating the **'Copy'** button will copy the graphical coverage map onto the clipboard in vector format (Windows meta file). This can then be copied into Word, Excel, PowerPoint and other programs capable of handling vector graphics. Vector graphics has the advantage that it can be scaled without loss of resolution. The graphical coverage map cannot be saved directly to disk, you will have to go through another program.

In the right-hand panel is a list-box 'Levels', showing the name and color of each level. The order of the levels can be changed through the up/down arrow above the list-box.

You set the label of each peptide in the coverage map through the group of radio buttons below the 'Levels' box. The choices are

*None:* No label is shown in the peptide boxes.

*Label:* Show the label. This is usually transferred from the originating function.

*Mass:* Monoisotopic mass calculated by GPMW based on the 'from-to' of each peptide.

*Comment:* The comment line for each peptide. This is not normally set initially.

*First-Last:* The position of the peptide in the sequence.

## 12 – Utilities

### Edit level

You can edit each level either by double-clicking on the level name in the 'Levels' box, or you can click on the **'Edit level'** button after first selecting the correct line. This opens the 'Edit coverage' dialog box:

In this box you can edit for each peptide:

*From:* The beginning of the peptide

*To:* The end of the peptide

*Label:* The peptide label (16 characters)

*Comment:* The peptide comment (40 characters)

The label and comment are independent of each other.

The currently active row (peptide) can be cleared by pressing the **'Clear'** button. The **'Paste'** button pastes a level on the clipboard into the current level thus combining them. You are asked for confirmation before the pasting operation.

**'Paste table'** will paste a table from the clipboard, see below 'From other programs'. The **'Import as multiple levels'** option will split the pasted coverage into different levels, if empty levels are available.

**'Clear all'** clears the entire level.

The level name can be edited up to 40 characters.

Clicking on the small **'Color'** panel opens the standard color dialog box, enabling you select a different color for the box around each peptide label in the main figure.

### Getting data into the sequence coverage

#### Manually

Start by opening the correct sequence on the GPMW desktop. Open the Coverage analysis window and click on the arrow in the **'Load new'** button. From the drop-down menu, you select 'Sequence from desktop' and select the correct sequence. You can then click on the **'Edit level'** button to manually edit each level, or you can paste levels from other parts of the program.

#### Mass search report:

The primary input for the sequence coverage is the

#	From / To	Label	Comment
1	97	102	97-102
2	134	139	134-139
3	140	147	140-147
4	146	153	146-153
5	32	42	32-42
6	134	145	134-145
7	64	77	64-77
8	119	133	119-133
9	17	31	17-31
10	1	16	1-16
11	80	96	80-96
12	32	47	32-47
13	79	96	79-96
14	78	96	78-96
15			
16			

Current row: Clear Paste level Paste table

☐ Import as multiple levels [1]

Level name:  Color: Clear all

OK Cancel Help

## 12 – Utilities

mass search report (Ch. 6.1). In the right-hand panel on the report page you have buttons for saving the report, and for copying to the clipboard.

The **'Save report'** button activates a small panel giving you the option to name the coverage map (default is the title of the window), save coverage with multiple levels (if the coverage contains overlapping segments – the number of levels is shown in parenthesis) and to select a color for the coverage map. A default color is shown, which can be changed by clicking on the colored panel. The default color is chosen randomly among 16 colors.

If you select **'OK'** you can save the report to a file on disk, if you click on **'Clipboard'** it will be copied to the clipboard in text format. If you want to copy the coverage as a graphic, you have to click the **'Copy to clipboard'** button in the right-hand panel.

### Cleavage analysis

From the cleavage analysis window you can save the coverage map to a file on disk or to the clipboard through the appropriate buttons in the right-hand panel.

### Peptide window

From the peptide window you can also copy a coverage map to the clipboard (not to file). Right-click in the window and select **'Copy special'** from the pop-up menu. In the left-hand list of radio-buttons you select **'Copy as coverage file'**. If **'Copy selected items only'** is checked; only the peptides selected in the list will be copied.

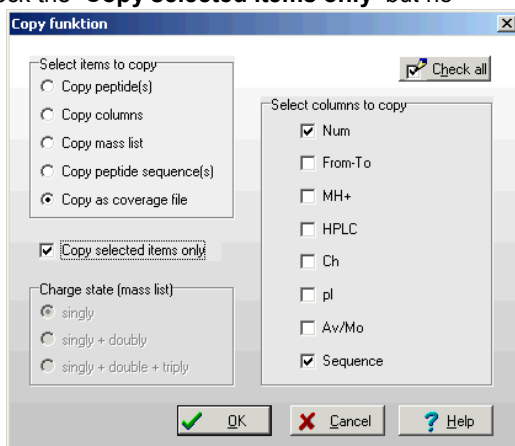
Copying a peptide coverage list from the peptide window is usually much faster than entering the peptide values in the **'Edit level'** window. This is also a way of getting a sequence without any coverage levels into the **'Sequence coverage'** window: If you check the **'Copy selected items only'** but no peptides are selected; only the protein sequence will be copied onto the clipboard.

### From other programs (reports)

If you have performed a Mascot search, you will obtain a sequence coverage as part of the details of the search. You can transfer that to a GPMW 'sequence coverage' by:

Load the appropriate sequence into GPMW

(either load the sequence through the annotation number on the web (Ch. 2.7), or copy and paste it into the **'Import ASCII dialog'** (Ch. 2.5)), followed by loading it into the **'Sequence coverage'** window as described above.



## 12 – Utilities

On the web page with the Mascot search report, copy the 'Detailed information' table to the clipboard (highlight and Ctrl + C).

Switch to GPMaw and in the 'Sequence coverage window' you select the arrow on the '**Past level**' button and select '**Paste table**' from the menu. This opens a window, where the different columns have been parsed into a spreadsheet-like grid.

Import text to column

	1	2
1	1	-
2	17	-
3	32	-
4	51	-
5	64	-
6	79	-
7	80	-
8	103	-
9	119	-
10	119	-
11	134	-

Matched peptides shown in Bold Red

1 GLSDGEWQQV LIFWGKVRAD IAGHQGEVLI RLFTGHPEPL EKFDKFFHLK  
51 TEARMGEASD LKRRDTVLT ALGGILRRGQ HHEAEKLPLA QSHATKHP  
101 IYLFETSDA IYHFLSHKIF GDPGADAGGA NTKEALELRY DIARYKELG  
151 FQG

Show predicted peptides also

Sort Peptides By: ☒ Residue Number ☐ Increasing Mass ☐ Decreasing Mass

Start - End	Observed	Mr (expt)	Mr (calc)	Delta	Miss	Sequence
1 - 15	1815.8490	1814.8417	1814.8951	-0.0534	0	- GLSDGEWQQV/LIFWGK.V
17 - 31	1606.8630	1605.8557	1605.8474	0.0083	0	K-VEADIAAGHQGEVLI.L
32 - 42	1271.6340	1270.6267	1270.6557	-0.0290	0	R-LFTGHPEPLEK.F
51 - 62	1351.5680	1350.5607	1350.6337	-0.0729	1	K-TEARMGEASDILK.K
64 - 77	1378.7980	1377.7907	1377.8343	-0.0436	0	K-RDTVLTALGGILK.F
79 - 96	1882.0440	1881.0367	1881.0493	-0.0126	1	K-RHHEAEKLPLAQSHATK.H
80 - 96	1853.9450	1852.9377	1852.9543	-0.0166	0	K-GHHEAEKLPLAQSHATK.H
103 - 118	1884.9890	1883.9817	1884.0145	-0.0328	0	K-YLEFISDALIYHFLSHK.H
119 - 133	1502.6190	1501.6117	1501.6619	-0.0502	0	K-HPGDPAADAGQAMTK.A
119 - 133	1518.6130	1517.6057	1517.6568	-0.0511	0	K-RPGDPAADAGQAMTK.A Oxidation
134 - 139	748.3840	747.3767	747.4279	-0.0512	0	K-ALELFLR.N

GPMaw will try to 'guess' which columns are the 'from' and 'to' columns and will highlight these in green. In the right-hand panel you can select the correct columns. Furthermore, you can select a column to be 'Label' and another to be 'Comment' – typically you will want one of them to be the determined peptide sequence.

Selecting '**OK**' will transfer the table to the Coverage window.

The same operation can be performed through the '**Edit level**' dialog box. Here you can check the option '**Import as multiple levels**', when this is checked, overlapping peptides will be imported into separate levels (if empty levels are available).

### File format

The file format of the coverage map is quite simple and is in plain text:

First line is: COVERAGE MAP

Second line is the name of the coverage map (40 character limit)

Third line is the protein sequence in a single line one-letter code (no line breaks!!).

Fourth line is the word 'LEVEL' (upper case!) followed by a tab character, the name of the level (max 40 characters), another tab; level type (12 characters) – not used at present, another tab; finally the color of the level a number between 0 and 64000.

The following lines (max 500) contains the peptides in the level and starts with the hash character (#) immediately followed by a number, a tab, the starting point of the peptide, a tab, the end point, a tab, fragment label (16

## **12 – Utilities**

characters), a tab, comment (40 characters). The first three values and all the tabs have to be entered, but the label and comment are optional.

The next level is a line starting with the word 'LEVEL'. Up to eight (8) levels can be accommodated in a coverage map.

---

## File formats

Specifications of the various file formats used by GPMaw.

Most of the file formats written by GPMaw are in plain text (called ASCII files). While the file can be edited with a text editor like Notepad, you should only do so if it is absolutely necessary, as the consequences of doing anything wrong will most likely be that the program ceases to function. The only exception to this rule is sequence files that are often obtained from the most unlikely places. Do not use a word processor like Word or WordPerfect - you'll most likely end up with an unreadable non-ASCII file.

### Directory structure

Due to the large number of files that can be generated during the use of GPMaw, the various files are located in different directories under the main installation directory (by default called C:\GPMaw, but can be changed during installation):

**\BIN:** Contains the binary files: the main executable (gpmaw3.exe), help file, and check files as well as the INI file. Several additional helper programs will also be located here.

**\DATABASE:** The digest mass databases generated by the user. It is not necessary to place databases here; quite often it is more convenient to position them on another disk, or if you are a member of a networked group you may locate these files on a central server, as they can be quite large.

**\SYSTEM:** Shared files that can be generated by the user, but are useful for other users (using the same computer) or other sessions. The mass files (MSS), the highlight profiles (HPF) and the modification files (MOD) are among the files saved in this directory.

**\USER:** The default user directory containing the sequence files (SEQ) and the other files resulting from daily use: peptide lists (PEP) and peak lists (PKS). Other user directories can be created and used as default, please see Chapter 5.4, Setup).

**\BINDATA:** This directory contains various files for use by the local BLAST engine. Please refer to chapter 7 for details.

**\BINCLUSTALW:** If you have installed ClustalW on the computer (see Chapter 7.3 for details), this is where the executable and accessory files will be located.

**\BACKUP:** This directory contains backup copies of files replaced during an upgrade of the program. If an upgrade fails (e.g. due to license problem), the relevant files can be retrieved from here.

Two files are essential for starting GPMW: GPMW3.INI (the initialization file) and AA\_MASS.MSS (the default mass file). If the files are missing the program will generate default values, but particularly for the INI file a number of parameters will not be convenient, see below. Most of the other files used by the program are generated during use.

The files can be recognized by their 3-character extension (after the last dot in the name):

**MSS:** Mass files.

**SEQ:** Sequence files (single sequences or small databases).

**PEP:** Peptide list.

**PKS:** Peak lists.

**HPF:** Highlight profiles (motifs).

**MOD:** Modification file.

**DAT, NAM, and INF:** Digest mass database files.

### GPMW3.INI

The initialization file is read during the start of GPMW and contains all the customizable information used by the program. The file follows the standard Windows INI file format with a section header (in square brackets) followed by setup data. The format of some of the data is obvious, but the user should not modify this file.

#### General content (sections):

[DISPLAY]	general layout of the display, sequence window, mass format, highlight etc.
[PRINTER]	format of print layout for peptide lists.
[DRIVES]	directory location of databases, default and system directory (see Chapter 5.4, Setup).
[ENZYMES]	names and cleavage specifications for auto digests (see Chapter 9.1, Cleavage).
[COLOURS]	default colors described as integer values (see Chapter 5.3, Setup Colors).
[FASTFILES]	a list of the five most recently used files.
[ATOM MASS]	abbreviation, name and mass of the atoms used to construct amino acid residues and modifications.
[MSMS]	abbreviation, type and composition for ms/ms fragmentation.



## A - File formats

[DIGESTSEARCH] default values for digest mass search (see Chapter 5.5, Setup and Chapter 8, Digest mass search).

### Sequence files (.SEQ)

A sequence file in GPMW can contain several sequences (acts like a small database) and will often contain information in addition to the basic name and amino acid sequence. The maximum size of a .SEQ file is 250 sequences or 62.000 bytes, whichever is smallest.

Although most sequence views in GPMW can be switched between 1- and 3- letter display, amino acid sequences are always saved in their 1-letter code in order to save space, increase speed and increase compatibility with other programs.

The basic sequence format consists of a line containing the name of the sequence terminated by a dollar sign (\$). The amino acid sequence then follows on the following lines until terminated by a star (\*). The first two entries in the example sequence file 'Blood' (can be found in the USER directory after installation):

```
Haptoglobin-2 precursor - Human$
MSALGAVIALLLWGQLFAVDSGNDVTDIADDGCPKPPEIAHGYVEHSVRYQCKNYYKLRT
EGDGVYTLNDKKQWINKAVGDKLPECEADDGCPKPPEIAHGYVEHSVRYQCKNYYKLRT
GDGVYTLNNEKQWINKAVGDKLPECEAVCGKPKNPANPVQRILGGHLDAGKSFQAKMV
SHHNLTTGATLINEQWLLTTAKNLFNLHSENATAKDIAPTLTLYVGKKQLVEIEKVVLP
NYSQVDIGLIKQKQVSVNERVMPICLPKDYAEVGRVGYVSGWGRNANFKFTDHLKYVM
LPVADQDQCIRHYEGSTVPEKKTTPKSPVGVQPILNEHTFCAGMSKYQEDTCYGDAGSAFA
VHDLLEEDTWYATGILSFDKSCAFAEYGVYVKVTSIQDWVQKTIAEN*
Coagulation factor XI (EC 3.4.21.27) precursor - Human$
MIFLYQVVFILFTSVSGECVTLLKDTCEGGDITTVFTPSAKYQCVVCTYHPRCLLFT
FTAESPSEDPTWFTCVLKDSVTETLPRVNRATAISGSYFQCSHQISACNKDIYVDLDM
KGINYNSSVAKSAQECQERCTDDVHCHFFTYATRQFPSLEHRNICLLKHTQTGTPTRITK
LDKVVSGFSLKSCALSNLACIRDIFPNTVFADSNIDSVMAPDAFVCGRICTHHPGLFFFT
FFSQEWPKESQRNLCLLKTSEGLPSTRIKSKALSGFSLQSCRHSIPVFCHSSFYHDTD
FLGEELDIVAAKSHEACQKLCTNAVRCQFFTYTPAQASCNEGKGCYKLKLSNGSPTKIL
HGRGGISGYTLRLCKMNECTTKIKPRIVGGTASVRGEWPWQVTLHTTSPTQRHLGGSI
IGNQWILTAACHFYGVESPKILRVYSGILNQSEIKEDTSFFGVQEIIHDQYKMAESGYD
IALLKLETTVNYTDSQRPICLPSKGRNVIYTDWVGTGWGYRKLKRDQNTLQKAKIPLV
TNEECQKRYRGHKITHKMICAGYREGGKDACKGDSGGPLSCKHNEVWHLVGTISWEGECA
QRRPVGVTNVVEYVDWILEKTQA*
```

The sequence name may contain any character except '\$', '\*', '-' and '@'. The sequence may similarly contain any character, only characters specified as 1-letter codes in the present mass file will be accepted when reading a sequence. The length of the sequence lines as well as spaces etc. has no influence. This has the following implication:

Always load the correct mass file before loading the sequence if our sequence contains non-standard 1-letter codes.

## A - File formats

When downloading sequences from other sources remember to put a '\$' character after the name (one line only) and a '\*' character after the sequence in 1-letter code (use Notepad not Write or a word processor).

If your downloaded sequence contains formatting characters like dots '.', numbers, spaces or newlines, these will be ignored when GPMW reads the sequence and does not have to be removed. Dashes '-' are read as chain delimiters and should be removed. They can be removed by editing the sequence after being read into the program, but GPMW has a limit of six chains (five delimiters) after which no further sequence will be read.

When GPMW saves a sequence, the name will be in the first line and the sequence will follow with 60 characters per line.

### Additional sequence information.

When the sequence has been modified in GPMW, the additional information is saved in lines between the name of the sequence and the actual sequence in lines starting with '\' (backslash) and a character specifying the kind of information that follows. Remember that sequences are not saved automatically when changes have been made, but you have to save each file specifically.

#### Information saved:

<b>\MR</b>	Modified residue: position, name and composition.
<b>\SS</b>	Cross-links: position-position linked.
<b>\NT</b>	N-terminal residue: name and composition.
<b>\CT</b>	C-terminal residue: name and composition.
<b>\OF</b>	Sequence offset.
<b>\CL</b>	Color residue list (marked residues, Ch. 3.3).
<b>\AN</b>	Annotation: following lines of annotation start with '\' (backslash and a space) until terminated by a line containing only <b>\END</b> .
<b>\ID</b>	ID (accession number).
<b>\HU</b>	Highlight (underline) regions.

### Mass files (.MSS)

These files contain the 1- and 3-letter code, full name as well as the composition of each residue. The file contains 30 residues followed by 2 time 6 modified residues for the N- and C-terminal modifications, respectively. You can make other mass files in the program (see Edit mass files, Chapter 4.2), but the file AA\_MASS.MSS will always be loaded when the program starts.

Mass file (only part of the file is shown, parts omitted are indicated by .....):

```
+
H2O
Base mass
H2O1
X
Xxx
Unknown
```

## A - File formats

```
C6H8N1O1
A
Ala
Alanine
C3H5N1O1
....
B
Bbb
Unknown

.....
-
---
Split

Hydrogen
H1
Free acid
H1O1
Pyroglutamic acid
C5H6N1O2
.....
```

### Peptide list (.PEP)

Peptide lists are text files listing a series of real values, one to each line, each representing a mass. The first line has to be 'GPMW PEPTIDE' and a maximum of 100 masses can be accepted. The files can easily be edited using Notepad, but it is easier and safer to edit in GPMW (load into Mass search (Ch. 6.1) or Digest mass search (8.3)):

```
GPMW PEPTIDE
1078.1900
727.7800
3498.0200
0.0000
0.0000
.....
```

### Peak files

In addition to reading and saving mass values in the peptide list (above) the GPMW program can **read** peak lists (i.e. files containing mass values) created by a number of other programs. These peak lists usually have extension like PKM, PKS, LST or PEP. Depending on the input dialog GPMW reads either both mass and intensity values or only the mass values. The format of the peak file is recognized by GPMW based on the following criteria:

**HP MALDI-TOF** is a binary file format that cannot be modified or viewed by the user. It is recognized by GPMW by containing the word 'TOF\_CATALOG\_VER2' in line 3. Only the mass values are read.

## A - File formats

**PerSeptive GRAMS TOF** is an ASCII (text) file generated by the GRAMS software package. The file is recognized by the first line ' "PEAK TABLE" '. The information in the beginning of the file is skipped, until a line starting with 'Center X' is encountered. Then all peak mass values and intensities are read into GPMaw.

**Bruker TOF** peak files are recognized on the first line starting with '####'. Then the content of the file is skipped until the line '/\*\*\*\*\*' Peak List Report \*\*\*\*\*/' is encountered after which the mass values and intensities are read.

**ASCII peak table** is a simple format that contains both mass and intensity values. The first line has to be 'ASCII PEAK TABLE', then follows three lines that contains free text (skipped by GPMaw) and finally the list of mass and intensity values separated by either a space or a tab character (ASCII #8). If you do not have the intensity value, you have to insert a zero '0'. Both mass and intensity values can be real numbers (e.g. they do not have to be integers). Example:

```
ASCII PEAK TABLE
My own mass data
Saved from an Excel spreadsheet
Date: 05-15-99
  727.7800 20154
1078.1900 42500
3498.0200 12514
.....
```

**Mass/intensity table** is similar to the table above, except that the first line has to be 'MASS/INTENSITY TABLE' and a variable number of free text lines are allowed, but they have to be terminated by a line starting with '----'. Then follows the mass and intensity values separated by space or tab characters (ASCII #8) like in the above table.

### Highlight profiles (.HPF)

The highlight profile files are text files consisting of 13 lines with the text 'GPMaw highlight profile' in the first line and one highlight motif on each of the following lines (four for each color), each line can have up six residues (1-letter code) with '?' substituting for any residue:

```
GPMaw highlight profile
N?S
N?T
N?C

K
R
.....
```

### Digest mass files (.DAT, .NAM and .INF)

When digest mass files are created, three files are generated, all with the same name but having different extensions. The DAT and the NAM files are binary files and should never be tampered with, as all access to these files

## A - File formats

will most likely be destroyed. The INF file is in ASCII and can be edited by Notepad. The only lines in the file that are likely to need to be modified are the lines specifying the location of the original databases. This is necessary if you move the databases from one drive to another (across a network, to a central server or from CD-ROM) and if you access the databases from different computers.

### Modification files (.MOD)

The modification files are text files with 30 entries, each entry consists of four lines: Name (10 characters), composition, valid residues (1-letter code, no residue specified means that all are valid), 1 for enabled - 0 for disabled. The first line has to be 'MODIFICATION V.3.01':

```
MODIFICATION V.3.01
Methylation
C1H2
DE
1
Oxydation
O1
M
1

0
.....
```

### The INF file format:

```
GPMW 3.0
OWL                                     {database}

1                                     {number of databases}
154601                               {number of sequences}
D:\DATABASE\OWL.SEQ                 {location of database}
TRYPOWL                             {digest database name}
/K/R-\P                             {cleavage parameters}
Average                             {mass type}
4
```

## Mass list input files

A.2

### .pkl file

The pkl file format is a simple text file containing a list of mass/intensity values. The first line of each spectrum data contains the parent ion mass, intensity and charge. Each spectrum is divided from the next by an empty line:

```
741.130 42954.00 1
267.0100 2649.0000
```

## A - File formats

```
283.0700 3869.0000
286.1000 514.0000
...
...

361.090 51286.00 1
101.1300 255.0000
121.2900 538.0000
...
...
```

### .mgf file

The mgf file format (Mascot generic file) has each spectrum contained in a BEGIN IONS and END IONS lines. In addition there are lines defining parent mass, charge and a title:

```
SEARCH=MIS
REPTYPE=Peptide
BEGIN IONS
PEPMASS=415.07863610848
CHARGE=2+
TITLE=Elution from: 0.725 to 0.725 period: 0 experiment: 1 cycles: 1
preIntensity: 10144.0 FinneganScanNumber: 37 MStype: enumIsNormalMS
rawFile: Ob00349.RAW
147.06349 2.1
160.03842 0.5
...
...
END IONS

BEGIN IONS
PEPMASS=415.07863610848
CHARGE=1+
TITLE=Elution from: 0.725 to 0.725 period: 0 experiment: 1 cycles: 1
preIntensity: 10144.0 FinneganScanNumber: 37 MStype: enumIsNormalMS
rawFile: Ob00349.RAW
147.06349 2.1
160.03842 0.5
192.86951 0.3
...
...
END IONS
```

### .dta file

The dta file format is the simplest, as it just contains mass / ion pairs, each spectrum starting with the parent ion followed by fragment mass / ion pairs. Spectra are separated by an empty line.

```
267.0100 2649.0000
283.0700 3869.0000
286.1000 514.0000
```

## A - File formats

...

...

101.1300 255.0000

121.2900 538.0000

...

...

### **.gpm file**

The gpm file format (gpmaw mass file) is a simple way of presenting list files as a single file. First line of each spectrum is a title (TITLE=) and then follows the fragment spectra, either just mass values or mass / intensity values:

TITLE=A1\_MS\_1.list

734.2588

837.4823

853.4734

...

...

3451.0166

TITLE=A3\_MS\_1.list

731.9337

742.1808

...

...

### **.list file**

List files are just the mass values of single fragment spectra without any header.





---

## Databases

How to handle protein databases to be used by GPMW.

Protein databases are used by GPMW for two purposes:

- 1) Direct search by name to retrieve a sequence.
- 2) As a source to generate digest mass databases. For this purpose the database has to be in FastA or PIR/NBRF format.

### Databases for sequence retrieval

### B.1

#### SWISS-PROT

EMBL Outstation  
Hinxton Hall  
Hinxton  
Cambridge CB10 1RQ  
U.K.

(Fax: +44 1223 494468; E-mail: [datalib@ebi.ac.uk](mailto:datalib@ebi.ac.uk))

The Swiss-Prot database is the best annotated protein database available. You can access it directly on the Internet (<http://www.ebi.ac.uk> or <http://www.expasy.ch>) or download it for local access on your computer. If you download it, you will have to index the database before use, please see Chapter 12.4.

The version of Swiss-Prot delivered with GPMW is already indexed and ready for use after installation.

#### IPI

International Protein Index databases.

IPI provides a top level guide to the main databases that describe the **human, mouse and rat** proteomes: Swiss-Prot, TrEMBL, RefSeq and Ensembl.

IPI:

- effectively maintains a database of cross references between the primary data sources
- provides minimally redundant yet maximally complete sets of human, mouse and rat proteins (one sequence per transcript)
- maintains stable identifiers (with incremental versioning) to allow the tracking of sequences in IPI between IPI releases.

IPI is updated monthly in accordance with the latest data released by the primary data sources.

The databases can be downloaded from this address:

## B - Databases

<ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/>

Each database (human, mouse, rat) is available in Swiss-Prot and FastA format – having the extension .dat and .fasta respectively. The FastA formatted version can be indexed right after download and decompression, while the .dat file has to be converted into FastA before indexing. However, if both the .dat and the converted files are in the same directory, GPMW will load the complete IPI database record instead of only the limited FastA record.

### PIR

The PIR database is maintained by:

National Biomedical Research Foundation  
3900 Reservoir Road, NW  
Washington, DC 20007  
USA

E-mail: [pirmail@gunbrf.bitnet](mailto:pirmail@gunbrf.bitnet)

&

Martinsried Institute for Protein Sequences  
Max-Planck-Institute for Biochemistry  
82152 Martinsried  
Germany

E-Mail: [mewes@ehpmic.mips.biochem.mpg.de](mailto:mewes@ehpmic.mips.biochem.mpg.de)

You can search the PIR databases through <http://pir.georgetown.edu/> where you can also download the databases in FastA format.

The PIR and SWISS-PROT databases downloaded from the Internet are not accessible for searching as the index files are missing. See Chapter 12.4 for details on indexing.

After indexing, most of the databases in section B.2 can also be used for data retrieval, but you will only be able to retrieve the protein name and sequence (due to the limited amount of information saved in the FastA format).



**Note:** PIR has recently joined forces with EBI and SIB to create the UniProt (United Protein Databases). GPMW supports this database also.

## Databases for digest mass search

### B.2

#### Internet

Databases downloaded from the Internet can usually be used as input. As they are stored in various variations of the FastA format, compatibility cannot be guaranteed. If you experience serious problems with an important database, please contact Lighthouse data.

Databases on the Internet are usually compressed in the Z format (not compatible with ZIP) and have to be decompressed with a utility like GZIP, GUNZIP, 7-ZIP or similar before use. The decompression utilities are usually available free of charge on the Internet. Newer versions of popular commercial compression packages like WinZip and ZipMagic can also handle Z compressed files.

## B - Databases

If you are unable to find a copy of your favorite database try the GPMaw home page (at present <http://www.gpmaw.com>). Here you can also find references to some of the more recent databases. If the page is unavailable you can send a request to [php@bmb.sdu.dk](mailto:php@bmb.sdu.dk).

FTP sites for downloading databases (not an exhaustive list):

EBI:	<a href="ftp://ftp.ebi.ac.uk/pub/databases/">ftp://ftp.ebi.ac.uk/pub/databases/</a>	(Swiss-Prot, TrEMBL)
NCBI:	<a href="ftp://ftp.ncbi.nlm.nih.gov/repository/">ftp://ftp.ncbi.nlm.nih.gov/repository/</a> <a href="ftp://ftp.ncbi.nlm.nih.gov/blast/db/">ftp://ftp.ncbi.nlm.nih.gov/blast/db/</a>	(OWL, GenPept) (NCBI-nr)
Expasy:	<a href="ftp://www.expasy.ch/databases/">ftp://www.expasy.ch/databases/</a> <a href="ftp://www.expasy.ch/databases/sp_tr_nrdb/">ftp://www.expasy.ch/databases/sp_tr_nrdb/</a>	(Swiss-Prot) (S-P, TrEMBL, nr)
EMBL:	<a href="ftp://ftp.embl-heidelberg.de/pub/databases/nrdb/">ftp://ftp.embl-heidelberg.de/pub/databases/nrdb/</a>	(EMBL-nr)
PIR:	<a href="ftp://nbrfa.georgetown.edu/pir_databases/psd/fasta/">ftp://nbrfa.georgetown.edu/pir_databases/psd/fasta/</a> <a href="ftp://nbrfa.georgetown.edu/pir_databases/nref/">ftp://nbrfa.georgetown.edu/pir_databases/nref/</a>	(PIR) (NREF)
IPI:	<a href="ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/">ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/</a>	(IPI)

### File formats

### B.3

#### FastA

Each entry in the FastA format begins with a '>' followed by the entry number, a delimiting character (usually '|' or ';') followed by the name of the sequence entry. On the next line(s) follows the sequence in 1-letter code until terminated by '\*' or a new entry (recognized by the line starting with '>').

```
>P1;CBRT
Cytochrome b - Rat mitochondrion (SGC1)
MTNIRKSHPLFKIINHSFIDL PAPS
VTHICRDVNYGWLIRY
TWIGGQPVEHPFIIIGQLASISYFSIIL8ILMPISGIVEDKMLKWN*
PIR/NBRF
```

The PIR/NBRF format is similar to the FastA format except that the first line only contains the '>' and entry number and the name follows on the second line. The following lines then contain the sequence.

```
>P1;CBRT; Cytochrome b - Rat mitochondrion (SGC1)
MTNIRKSHPLFKIINHSFIDL PAPS
VTHICRDVNYGWLIRY
TWIGGQPVEHPFIIIGQLASISYFSIIL8ILMPISGIVEDKMLKWN*
```

#### Indexing FastA databases

An indexing utility, 'DBIndex', is available from Lighthouse data that enables searching by name or accession number on databases downloaded from the Internet or elsewhere (see Chapter 12.4 for more information). If the database indexer is not part of your package, please contact Lighthouse data ([e-mail php@bmb.sdu.dk](mailto:php@bmb.sdu.dk)) for further details.



## Tables

Names, formulas and molecular masses of standard amino acid residue, modified residues, secondary modifications, carbohydrate units and pI values.

### Mass types

C.1

When you calculate molecular masses, the mass type you calculate will depend upon the resolution of the mass spectrometer used.

**Monoisotopic mass:** If you have sufficient resolution to distinguish the individual isotopes you can calculate the masses based on  $C = 12.0000$ . This is usually the case for sector instruments and MALDI instruments equipped with an ion mirror. Samples determined as monoisotopic masses are generally more precise than when they are determined as average masses. As you move towards higher masses ( $> 4\text{-}5\text{ kDa}$ ), the determination of the monoisotopic peak gets more difficult and it often makes sense to calculate your sample using smoothing and averaging.

Monoisotopic mass is also called exact mass.

**Average mass:** If your resolution is not high enough to distinguish the individual isotopes, you calculate the mass based on the natural mixture of isotopes. This is usually the case for time-of-flight instruments without ion mirror and high masses (e.g intact proteins).

**Integer mass:** This is based on the integer value of each amino acid residue. The integer mass of a protein does not have relation to a measured value but is included in the program for comparison because many databases still reports this value.

### Atomic masses

C.2

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	35.452737
		36.96590262	
Iodine	I	126.904473	126.904473

## C - Tables

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

## C - Tables

### Masses of less commonly occurring and modified residues C.4

Aminobutyric acid	Aba	C4H7NO	85.05276	85.1057
Aminoethylcys.	AECys	C5H10N2OS	146.05138	146.2132
Aminoisobutyric acid	Aib	C4H7NO	85.05276	85.1057
Carbamidomethylcys.		C5H8N2O2S	160.03065	160.197
Carboxymeth. cys	CMCys	C5H7NO3S	161.01466	161.1815
Dehydroalanine	Dha	C3H3NO	69.02146	69.0629
Homoserine	Hse	C4H7NO2	101.04768	101.105
Homoserine lactone	Hsl	C4H5NO	83.03712	83.090
Hydroxylysine	Hyl	C6H12N2O2	144.08988	144.1735
Hydroxyproline	Hyp	C5H7NO2	113.04768	113.1161
IsoValine	Iva	C5H9NO	99.06841	99.1326
NorLeucine	nLeu	C6H11NO	113.08406	113.1594
Ornithine	Orn	C5H10N2O	114.07931	114.1472
2-Piperidinecarb. acid	Pip	C6H9NO	111.06841	111.1436
Pyridylethylcysteine	PECys	C10H12N2OS	208.06703	208.284
Pyroglutamic acid	pGlu	C5H5NO2	111.03203	111.1022
Sarcosine	Sar	C3H5NO	71.03711	71.0788

### Mass changes due to post-translational modifications C.5

The following table shows only a small representation of the modifications possible for amino acid residues. Some modifications are also shown in the table above and sugar residues are shown in the table below.

Modification	Composition	Mass (av)	Residues
Pyroglutamic ac.	-N1H2	-17.0306	Gln (N-terminal)
Disulphide bond	-H2	-2.0159	2 x Cys
Dehydro alanine	-H2	-2.0159	Ala
C-term. Amide	-H1	-0.9847	Gly (C-terminal)
Deamidation	-H1	-0.9847	Asp, Glu
Methylation	C1H2	14.0269	Asp, Glu, C-term. + various
Hydroxylation	O1	15.9994	Pro, Lys + various
Oxidation	O1	15.9994	Met
Proteolys	H1O2	18.0153	Peptide bond
Formylation	C1O1	28.0104	Gly, Met, Trp, amino
Methylation (x2)	C2H4	28.0540	Lys, Arg
Acetylation	C2H2O1	42.0373	N-term, Lys, Ser
Carboxylation	C1O2	44.0098	Asp, Glu
Methylation (x3)	C3H7	43.0889	Lys

## C - Tables

Pyrovoyl	C3H2O2	70.0477	N-term
Phosphorylation	P1O3H1	79.9799	Arg, Asp, Cys, His, Lys, Ser, Thr, Tyr
Sulphation	S1O3	80.0642	Tyr
Trifluoroacetyl	-H1+C2O1F3	96.0086	N-term
4-nitrophenyl	C6H3N1O2	121.0955	C-term
Farnesylation	C15H24	204.3556	Cys-SH
Myristylation	C14H26O1	210.3598	N-term
Biotinylation	C10H14O2N2S1	226.2994	Lys
Dansylation	C12H9O2N1S1	233.29	N-term
Palmitoylation	C16H30O1	238.4136	Cys-SH
Glutathionylation	C10H15O6N3S1	305.3117	Cys-SH
5'-adenosyl	C10H12O6N5P1	329.2091	Tyr

## Carbohydrates

## C.6

Proteins are very often modified by the addition of carbohydrates onto certain residues. Usually only asparagine residues (N-glycosylation) or serine/threonine (O-glycosylation) are modified, but a number of other residues are also potential targets (hydroxylated lysine, glycation of lysine etc.).

As it is usually not possible to differentiate between different stereo isomers using mass spectrometry, the following table lists the various common monosaccharide residues. Please note that the formulas and molecular masses are the residue masses. In order to obtain the composition and mass of an isolated sugar you have to add water (H<sub>2</sub>O).

Name	Abbr.	Formula	Monoiso.	Average
<b>Deoxypentose</b>	<b>DeoxyPent</b>	<b>C<sub>5</sub>H<sub>8</sub>O<sub>3</sub></b>	<b>116.0473</b>	<b>116.1167</b>
Deoxyribose (dRib)				
<b>Pentose</b>	<b>Pent</b>	<b>C<sub>5</sub>H<sub>8</sub>O<sub>4</sub></b>	<b>132.0423</b>	<b>132.1161</b>
Arabinose (Ara), Ribose (Rib), Xylose (Xyl)				
<b>Deoxyhexose</b>	<b>DeoxyHex</b>	<b>C<sub>6</sub>H<sub>10</sub>O<sub>4</sub></b>	<b>146.0579</b>	<b>146.1430</b>
Fucose (Fuc)				
<b>Hexosamine</b>	<b>HexN</b>	<b>C<sub>6</sub>H<sub>11</sub>NO<sub>4</sub></b>	<b>161.0688</b>	<b>161.1577</b>
Galactosamine (GalN), Glucosamine (GlcN)				
<b>Hexose</b>	<b>Hex</b>	<b>C<sub>6</sub>H<sub>10</sub>O<sub>6</sub></b>	<b>162.0688</b>	<b>162.1424</b>
Galactose (Gal), Glucose (Glc), Mannose (Man)				
<b>Hexuronic acid</b>	<b>HexA</b>	<b>C<sub>6</sub>H<sub>8</sub>O<sub>6</sub></b>	<b>176.0321</b>	<b>176.1259</b>
Glucuronic acid (GlcA)				



## C - Tables

<b>Heptose</b>	<b>Hep</b>	<b>C<sub>7</sub>H<sub>12</sub>O<sub>6</sub></b>	<b>192.0634</b>	<b>192.1687</b>
<b>N-acetylhexosam.</b>	<b>HexNAc</b>	<b>C<sub>8</sub>H<sub>13</sub>NO<sub>5</sub></b>	<b>203.0794</b>	<b>203.1950</b>
N-acetylgalactosamine (GalNAc), N-acetylglucosamine (GlcNAc)				
<b>Muramic acid</b>	<b>Mur</b>	<b>C<sub>11</sub>H<sub>17</sub>NO<sub>7</sub></b>	<b>275.1005</b>	<b>275.2585</b>
Sialic acids (SA):				
<b>N-acetylneuram.a.</b>	<b>NeuAc</b>	<b>C<sub>11</sub>H<sub>17</sub>NO<sub>8</sub></b>	<b>291.0954</b>	<b>291.2579</b>
<b>N-glycolylneua.a.</b>	<b>NeuGc</b>	<b>C<sub>11</sub>H<sub>17</sub>NO<sub>9</sub></b>	<b>307.0903</b>	<b>307.2573</b>

### N-linked glycosylation

N-linked glycosylations all contain a common pentasaccharide core:

Man $\alpha$ 1-6

Man $\beta$ 1-4GlcNAc $\beta$ 1-4GlcNAc-Asn(peptide)

Man $\alpha$ 1-3

The outer mannose residues are then further derivatized depending on the glycosylation type:

**Complex:** One to five arms made of Gal $\beta$ 1-4GlcNAc $\beta$ 1-2/4/6. Each arm is usually terminated by sialic acid (Sia $\alpha$ 2-6).

**High mannose:** one to four Man $\alpha$ 1-2Man $\alpha$ 1-3/6 units.

**Hybrid:** One arm derivatized with a complex type arm and the other with a high mannose arm.

**Bisecting:** The complex and hybrid type can be further derivatized by a GlcNAc $\beta$ 1-4 saccharide on the inner mannose unit of the core.

**Fucose:** The inner GlcNAc unit of the core is often derivatized with a fucose residue (Fuc $\alpha$ 1-6).

The structural diversity mentioned above are the ones most commonly found, but a large number of other outer chain variations are found.

## C - Tables

### pKa values

**C.7**

The pKa values are used throughout the program for calculating pI and charges. In the sequence information window (Ch. 3.8) values for all three tables are given. In all other situations only the value from the column selected in Setup (Ch. 5.6) will be shown.

	N-terminus			C-Terminus			Side-chain		
	1	2	3	4	5	6	7	8	9
<b>Xxx</b>	9.5	9.51	8.8	2.2	2.21	3.13			
<b>Asp</b>	9.8	9.85	8.6	2.1	2.02	2.75	3.9	3.82	3.5
<b>Asn</b>	8.8	8.82	7.3	2.1	2.06	2.75			
<b>Thr</b>	9.1	9.10	8.2	2.1	2.09	3.2			
<b>Ser</b>	9.2	9.18	7.3	2.2	2.20	3.2			
<b>Glu</b>	9.7	9.57	8.2	2.2	2.15	3.2	4.3	4.18	4.5
<b>Gln</b>	9.1	9.13	7.7	2.2	2.17	3.2			
<b>Pro</b>	10.6	10.62	9.0	2.0	1.98	3.2			
<b>Gly</b>	9.8	9.78	8.2	2.4	2.35	3.2			
<b>Ala</b>	9.9	9.87	8.2	2.4	2.34	3.2			
<b>Val</b>	9.7	9.68	8.2	2.3	2.30	3.2			
<b>Cys</b>	10.8	10.40	7.3	1.7	1.93	2.75	8.3	8.26	10.3
<b>Met</b>	9.3	9.24	9.2	2.1	2.28	3.2			
<b>Ile</b>	9.8	9.72	8.2	2.3	2.34	3.2			
<b>Leu</b>	9.7	9.67	8.2	2.3	2.35	3.2			
<b>Tyr</b>	9.1	9.11	7.7	2.2	2.20	3.2	10.1	10.11	10.3
<b>Phe</b>	9.2	9.21	7.7	2.2	2.37	3.2			
<b>Lys</b>	9.0	9.06	7.7	2.2	2.17	3.2	10.5	10.66	10.3
<b>His</b>	9.2	9.18	8.2	1.8	1.79	3.2	6.0	6.08	6.2
<b>Trp</b>	9.4	9.42	8.2	2.4	2.40	3.2			
<b>Arg</b>	9.0	9.02	8.2	2.2	1.91	3.2	12.5	12.48	12.5

**Column 1, 4, 7** are from B. Skoog & A. Wichman, *Trends Anal. Chem.* **3**, 82-83 (1986). **Column 2, 5, 8** are free amino acids. **Column 3, 6, 9** are from Rickard, Strohl & Nielsen, *Anal. Biochem*, **197**, 197-207 (1991)

## C - Tables

### Peptide residue mass values < 304 Da

**C.8**

The following peptide masses are residue mass (i.e. without water). Cysteine is calculated as reduced (SH) and isoleucine is omitted as its mass is identical to that of leucine. First column is average mass, second column is monoisotopic mass.

G	57.052	57.021	EG	186.167	186.064	SK	215.252	215.127
A	71.079	71.037	SV	186.211	186.100	DT	216.194	216.075
S	87.078	87.032	W	186.213	186.079	SE	216.194	216.075
P	97.117	97.053	TS	188.183	188.080	CL	216.304	216.093
V	99.133	99.068	GM	188.250	188.062	GGC	217.249	217.052
T	101.105	101.048	SC	190.223	190.041	NC	217.249	217.052
C	103.145	103.009	GH	194.193	194.080	DC	218.233	218.036
L	113.159	113.084	PP	194.233	194.106	AF	218.255	218.106
N	114.104	114.043	PV	196.249	196.121	SM	218.277	218.073
GG	114.104	114.043	TP	198.222	198.100	GY	220.228	220.085
D	115.089	115.027	VV	198.265	198.137	SH	224.219	224.091
GA	128.131	128.059	GAA	199.210	199.096	QP	225.247	225.111
Q	128.131	128.059	QA	199.210	199.096	PGA	225.247	225.111
K	128.174	128.095	AK	199.253	199.132	PK	225.291	225.148
E	129.115	129.043	EA	200.194	200.080	EP	226.232	226.095
M	131.199	131.040	TV	200.238	200.116	LL	226.319	226.168
H	137.141	137.059	SL	200.238	200.116	QV	227.263	227.127
AA	142.158	142.074	PC	200.261	200.062	GAV	227.263	227.127
SG	144.130	144.053	NS	201.182	201.075	NL	227.263	227.127
F	147.177	147.068	SGG	201.182	201.075	GGL	227.263	227.127
PG	154.169	154.074	DS	202.167	202.059	AR	227.266	227.138
GV	156.184	156.090	TT	202.210	202.095	VK	227.307	227.163
R	156.187	156.101	AM	202.277	202.078	NN	228.208	228.086
SA	158.157	158.069	VC	202.277	202.078	NGG	228.208	228.086
TG	158.157	158.069	GF	204.228	204.090	EV	228.248	228.111
GC	160.197	160.031	TC	204.250	204.057	DL	228.248	228.111
Y	163.176	163.063	CC	206.290	206.018	PM	228.315	228.093
PA	168.195	168.090	AH	208.220	208.096	DGG	229.192	229.070
GL	170.211	170.106	PL	210.276	210.137	DN	229.192	229.070
AV	170.211	170.106	NP	211.221	211.096	TQ	229.236	229.106
GGG	171.156	171.064	PGG	211.221	211.096	SAA	229.236	229.106
NG	171.156	171.064	DP	212.205	212.080	TGA	229.236	229.106
DG	172.141	172.048	VL	212.292	212.152	TK	229.279	229.143
TA	172.184	172.085	NV	213.236	213.111	DD	230.177	230.054
SS	174.156	174.064	GGV	213.236	213.111	TE	230.221	230.090
AC	174.224	174.046	AAA	213.236	213.111	VM	230.331	230.109
SP	184.195	184.085	GR	213.239	213.123	SSG	231.208	231.086
AL	184.238	184.121	DV	214.221	214.095	QC	231.276	231.068
QG	185.183	185.080	TL	214.265	214.132	GAC	231.276	231.068
NA	185.183	185.080	SQ	215.209	215.091	CK	231.319	231.104
GGA	185.183	185.080	TGG	215.209	215.091	EC	232.260	232.052
GK	185.226	185.116	NT	215.209	215.091	TM	232.304	232.088
DA	186.167	186.064	SGA	215.209	215.091	SF	234.255	234.100

## C - Tables

AY	234.255	234.100	GVV	255.317	255.158	NPG	268.272	268.117
PH	234.258	234.112	AAL	255.317	255.158	MH	268.340	268.099
CM	234.343	234.050	VR	255.320	255.170	DPG	269.257	269.101
VH	236.274	236.127	QQ	256.261	256.117	TPA	269.301	269.138
TH	238.246	238.107	QGA	256.261	256.117	AVV	269.344	269.174
PAA	239.274	239.127	NAA	256.261	256.117	GVL	269.344	269.174
CH	240.286	240.068	QK	256.305	256.154	LR	269.347	269.185
SPG	241.247	241.106	GAK	256.305	256.154	NGV	270.288	270.133
AAV	241.290	241.143	KK	256.348	256.190	QAA	270.288	270.133
QL	241.290	241.143	EQ	257.246	257.101	GGR	270.291	270.144
GAL	241.290	241.143	EGA	257.246	257.101	NR	270.291	270.144
LK	241.334	241.179	DAA	257.246	257.101	AAK	270.332	270.169
QGG	242.235	242.102	SAV	257.290	257.138	DGV	271.273	271.117
NQ	242.235	242.102	TGV	257.290	257.138	SSP	271.273	271.117
NGA	242.235	242.102	SGL	257.290	257.138	EAA	271.273	271.117
EL	242.275	242.127	EK	257.290	257.138	DR	271.276	271.128
NK	242.278	242.138	AW	257.292	257.116	SAL	271.316	271.153
GGK	242.278	242.138	TR	257.293	257.149	TAV	271.316	271.153
DGA	243.219	243.086	PGC	257.313	257.083	TGL	271.316	271.153
EGG	243.219	243.086	EE	258.231	258.085	PAC	271.340	271.099
NE	243.219	243.086	NSG	258.234	258.096	SQG	272.261	272.112
DQ	243.219	243.086	DSG	259.219	259.080	NSA	272.261	272.112
TAA	243.263	243.122	TSA	259.262	259.117	NTG	272.261	272.112
DK	243.263	243.122	TTG	259.262	259.117	SGK	272.304	272.148
SGV	243.263	243.122	GAM	259.329	259.099	DTG	273.246	273.096
GW	243.265	243.101	QM	259.329	259.099	DSA	273.246	273.096
SR	243.266	243.133	GVC	259.329	259.099	SEG	273.246	273.096
DE	244.204	244.070	CR	259.332	259.110	SSV	273.289	273.132
PF	244.293	244.121	MK	259.373	259.135	TTA	273.289	273.132
ML	244.358	244.125	PY	260.293	260.116	SW	273.291	273.111
SSA	245.235	245.101	EM	260.314	260.083	AVC	273.356	273.115
TSG	245.235	245.101	LF	260.336	260.152	GCL	273.356	273.115
NM	245.302	245.083	SSS	261.235	261.096	AAM	273.356	273.115
GGM	245.302	245.083	NF	261.280	261.111	HH	274.282	274.118
AAC	245.302	245.083	GGF	261.280	261.111	NGC	274.301	274.074
DM	246.287	246.067	SAC	261.302	261.078	TSS	275.261	275.112
VF	246.309	246.137	TGC	261.302	261.078	DGC	275.285	275.058
SGC	247.275	247.063	DF	262.265	262.095	QF	275.307	275.127
TF	248.282	248.116	VY	262.309	262.132	GAF	275.307	275.127
SY	250.254	250.095	MM	262.397	262.081	SGM	275.329	275.094
LH	250.301	250.143	GCC	263.342	263.040	TAC	275.329	275.094
CF	250.321	250.078	TY	264.281	264.111	FK	275.351	275.163
NH	251.245	251.102	QH	265.272	265.117	EF	276.292	276.111
GGH	251.245	251.102	GAH	265.272	265.117	LY	276.335	276.147
PPG	251.285	251.127	PPA	265.312	265.143	NY	277.280	277.106
DH	252.230	252.086	KH	265.315	265.154	GGY	277.280	277.106
PGV	253.301	253.143	EH	266.257	266.102	SSC	277.301	277.073
PR	253.304	253.154	CY	266.321	266.073	ACC	277.368	277.055
SPA	255.274	255.122	PGL	267.328	267.158	DY	278.265	278.090
TPG	255.274	255.122	PAV	267.328	267.158	MF	278.375	278.109

## C - Tables

AAH	279.299	279.133	NAC	288.327	288.089	DAL	299.327	299.148
SGH	281.271	281.112	GCK	288.371	288.126	EAV	299.327	299.148
SFP	281.312	281.138	DSS	289.245	289.091	TTT	299.327	299.148
PAL	281.355	281.174	TTS	289.288	289.127	NGK	299.330	299.159
QPG	282.299	282.133	EGC	289.312	289.073	SVL	299.370	299.185
NPA	282.299	282.133	DAC	289.312	289.073	TVV	299.370	299.185
PGK	282.343	282.169	AAF	289.334	289.143	LW	299.373	299.163
DPA	283.284	283.117	SAM	289.356	289.110	PAM	299.394	299.130
EPG	283.284	283.117	TGM	289.356	289.110	PVC	299.394	299.130
SPV	283.327	283.153	SVC	289.356	289.110	DNA	300.271	300.107
PW	283.330	283.132	CW	289.358	289.088	DQG	300.271	300.107
GLL	283.371	283.190	GAY	291.307	291.122	NEG	300.271	300.107
AVL	283.371	283.190	QY	291.307	291.122	DGK	300.315	300.143
NGL	284.315	284.148	SGF	291.307	291.122	TQA	300.315	300.143
QGV	284.315	284.148	PGH	291.310	291.133	NSV	300.315	300.143
NAV	284.315	284.148	TSC	291.328	291.089	NW	300.317	300.122
FH	284.318	284.127	YK	291.350	291.158	GGW	300.317	300.122
GAR	284.318	284.160	PPP	291.350	291.158	YH	300.317	300.122
QR	284.318	284.160	GCM	291.395	291.071	SGR	300.318	300.155
GVK	284.359	284.185	EY	292.291	292.106	TAK	300.358	300.180
KR	284.362	284.196	GVH	293.326	293.149	DDA	301.256	301.091
NNG	285.260	285.107	HR	293.329	293.160	DEG	301.256	301.091
DAV	285.300	285.132	PPV	293.366	293.174	DSV	301.299	301.127
DGL	285.300	285.132	SCC	293.368	293.050	TEA	301.299	301.127
EGV	285.300	285.132	FF	294.353	294.137	DW	301.302	301.106
TSP	285.300	285.132	MY	294.375	294.104	TTV	301.343	301.164
ER	285.303	285.144	SAH	295.298	295.128	TSL	301.343	301.164
SVV	285.343	285.169	TGH	295.298	295.128	PGF	301.345	301.143
TAL	285.343	285.169	TPP	295.338	295.153	TPC	301.367	301.110
VW	285.346	285.148	PVV	295.382	295.190	AVM	301.410	301.146
PGM	285.367	285.115	QPA	296.326	296.148	GML	301.410	301.146
DNG	286.244	286.091	PAK	296.370	296.185	VVC	301.410	301.146
SQA	286.288	286.128	EPA	297.311	297.132	SSQ	302.287	302.123
TQG	286.288	286.128	GCH	297.338	297.090	NTS	302.287	302.123
NTA	286.288	286.128	SPL	297.354	297.169	SSK	302.330	302.159
SAK	286.331	286.164	TPV	297.354	297.169	NGM	302.354	302.105
TGK	286.331	286.164	PPC	297.378	297.115	QAC	302.354	302.105
DDG	287.229	287.075	VVV	297.398	297.205	ACK	302.398	302.141
TEG	287.272	287.112	ALL	297.398	297.205	DTS	303.272	303.107
DTA	287.272	287.112	NSP	298.299	298.128	SSE	303.272	303.107
SEA	287.272	287.112	NAL	298.342	298.164	TTT	303.315	303.143
TSV	287.316	287.148	QAV	298.342	298.164	EAC	303.339	303.089
SSL	287.316	287.148	QGL	298.342	298.164	DGM	303.339	303.089
TW	287.318	287.127	AAR	298.345	298.175	GVF	303.361	303.158
SPC	287.340	287.094	GLK	298.385	298.200	FR	303.364	303.170
GVM	287.383	287.130	AVK	298.385	298.200	TVC	303.382	303.125
ACL	287.383	287.130	DSP	299.283	299.112	SCL	303.382	303.125
MR	287.386	287.142	NQG	299.286	299.123	TAM	303.382	303.125
NSS	288.260	288.107	NNA	299.286	299.123	PCC	303.406	303.071
QGC	288.327	288.089	EGL	299.327	299.148			



## Configuring GPMaw.

### Configuring GPMaw startup

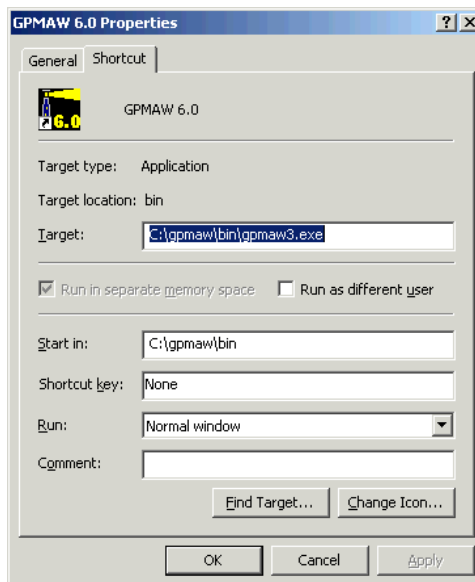
#### Changing the default behavior of GPMaw through in-line parameters

When GPMaw starts it accepts a number of in-line parameters, which enables you to modify the way the program works. This is especially useful in circumstances where multiple users access the same machine or you like to use different setups depending on the type of work you have to perform.

#### In-line parameters

When you start a program from a shortcut (icon placed on the desktop) you activate a command line string containing the full name of the executable file (e.g. C:\GPMaw\BIN\GPMaw3.EXE). You can gain access to the command line and modify it in the following way:

Highlight and right-click on the shortcut on the desktop. From the pop-up menu select the last option '**Properties**'. Then select the tab labeled '**Shortcut**'. The command line is in the line labeled '**Target**' (highlighted below).



You can now add commands to this line.

## D - Configuring GPMaw startup



**Note:** do not delete or change the name of the executable file – only add commands to the end of the line.

When you add in-line parameters always remember to put a space before each parameter.

**/D:** Startup directory. GPMaw usually starts in the directory called C:\GPMaw\USER where it reads the GPMaw3.INI file for initialization parameters. By using the '/D:' option you can change this to another directory, for example '/D:C:\GPMaw\MARTIN'. You have to specify the complete path to the directory, including drive. The directory has to exist beforehand and should include a copy of GPMaw3.INI unless the '/U:' option is used to specify a different INI file.

**/U:** User configuration (.INI) file. This option specifies that GPMaw should read a different initialization file than the normal GPMaw3.INI. This in-line parameter has to specify an existing INI file present in the gpmaw\bin directory. The option must not include the path or extension '.INI', for example '/U:MARTIN' if your INI file is called MARTIN.INI. You make a different INI file by copying the already existing GPMaw3.INI (using File Manager or Explorer), after which you specify the '/U:' option in a new shortcut.

**/CLOSE** This command will close the splash screen immediately (default will be a lighthouse displaying version information). The splash screen may flicker on your screen, but will close as soon as the program is ready.

See also Chapter 5.7 for setting up user configurations from inside GPMaw.

**Filename:** If you enter a full filename (e.g. C:\GPMaw\USER\BLOOD.SEQ) as an in-line parameter GPMaw will open this file upon startup. The file should be either in GPMaw or FastA format (see Appendix A). If the file contains multiple sequences you will be greeted by the 'Select sequence' dialog box (Chapter 2.1), if the file contains only a single sequence it will be opened directly. Using the filename in-line parameter you can thus have several shortcut icons on your desktop (or Program Manager), each customized to opening a specific sequence file. You can obtain a similar effect from the Explorer if you have created a link between the .seq extension and GPMaw, either when you installed the program or through the Explorer (menu 'Tools | Folder options | File types' – see below).

The different in-line parameters can be combined. However, the /D: option has to precede the /U: option.

### Auto-starting GPMaw

If you register the .SEQ extension with Windows you can auto-start GPMaw by double clicking a file (or icon) with a .SEQ extension in File Manager (Windows 3.1x) or Explorer (Windows '95/'98/NT).

#### Associating an extension with GPMaw:

**Windows:** Open 'Explorer' and select **View|Options...** or **Tools|Folder options...** Select the '**File types**' page and press the '**New type**' button. In the '**Add new file type**' dialog box enter a description in the '**Description**' field (e.g. 'GPMaw sequence file'). In the '**Associated**



## **D - Configuring GPMW startup**

**extension'** you enter '.SEQ'. Press **'New...'** and in the **'New action'** enter **'open'** in the **'Action'** field. In the **'Application'** field you either enter the complete file name of GPMW or use the **'Browse..'** button to browse and select the GPMW.EXE file. Select **'OK'**, **'Close'** and **'Close'** again to accept your selections.



---

## License

### License agreement

The enclosed program is furnished subject to the terms and conditions of this end user license agreement (EULA).

1) *License*. The license is a single user license and may only be used by a single person at a time. You are allowed to make backup copies of the disks and, furthermore, you are allowed to install the program on several computers (e.g. your home computer or portable) as long as only a single copy of the program is active at any one time.

2) *Restriction on use*. You may not modify or translate all or any part of the program. You agree not to distribute all or any part of the program to others except as expressly provided by section 4. You agree not to disassemble or reverse engineer the program, in whole or in part.

3) *Limitation of liability*. Lighthouse data will not in any event be liable to you for any damages, including any lost profits, lost savings or other incidental or consequential damages arising out of the use or inability to use the program, even if Lighthouse data has been advised of the possibility of such damages, or for any claim by any other party

The liability of Lighthouse data, its affiliates, subsidiaries and the licensors for damages or losses for any cause whatsoever, and regardless of the basis of the claim or action will be limited to the amount you actually paid for the specific software that caused the damages or losses.

Some states do not allow the limitation or exclusion of liability for incidental or consequential damages and to such extent the above limitation or exclusion may not apply to you.

4) *Transfer*. You are only allowed to transfer the program to another party if all original diskettes, manual etc. are transferred and all installed and backup copies are destroyed.

5) *General*. If any provisions of this agreement are legally unenforceable or illegal, such provision shall be severed from this agreement and the balance of this agreement shall continue in full force and effect. This agreement shall be governed by Danish law.

## INDEX

- About ..... 20
- Accession number ..... 34, 76, 167
- Acquire a sequence ..... 26
- Alpha-helical wheel..... 247, 249
- Alternate display ..... 203
- Amino acid residues..... 298
- Amphiphatic ..... 247
- Analyzing N-linked carbohydrates151
- Annotation..... 25, 70
- Atom mass table ..... 17
- Atomic masses..... 82, 297
- Autoload..... 12, 103
- Automatic digest ..... 193
- Auto-starting GPMW..... 308
- Average mass ..... 297
- Beta versions ..... 21
- BIN..... 98
- BLAST ..... 104, 143
- Bull & Breese index..... 94
- Calibration..... 114
- Caps ..... 77
- Carbohydrate editor ..... 86
- Carbohydrates ..... 300
- CD-ROM ..... 27, 32, 294
- Charge at pH..... 94
- Charge vs pH..... 204
- Charge vs. pH graph..... 217
- Cleanup ..... 77
- Clear coloring..... 56
- Cleavage..... 10, 193
- Cleavage analysis..... 197
- Clipboard ..... 32, 41, 94
- ClustalW ..... 148
- Cluster ions..... 47
- CNBr..... 193, 197
- Collapsed mode..... 223
- Color residues..... 56
- Color single residue ..... 63
- Colors ..... 96
- Combine databases ..... 268
- Combine digest mass search..... 174
- Composition..... 66, 141
- Composition calculator..... 258
- Composition editor ..... 84
- Composition formula..... 88
- Convert Swiss-Prot ..... 270
- Converting ..... 269
- Copy..... 51, 94
- Coverage analysis ..... 276
- Coverage map..... 220
- Cross-linked peptides ..... 121
- Cross-linking..... 121
- Cross-linking reagents..... 123
- Cross-links..... 25, 58
- Cys ..... 16, 55, 58, 59, 102, 214, 215
- Database ..... 33, 161, 293
- Database indexer ..... 263
- Database information ..... 162
- Database mass search..... 157
- DBindex..... 33, 263
- DBIndex..... 35
- Delete sequence ..... 29
- de-novo ..... 225, 259
- Difference table ..... 120
- Digest mass database..... 158
- Digest mass files ..... 288
- Digest mass search11, 157, 161, 165
- Digest mass search parameters.. 100
- DigestAlyzer ..... 250
- Directories ..... 98, 158, 283
- Directory structure ..... 283
- Display defaults ..... 102
- Display page..... 102
- Disulfide cross-links..... 214
- Dot-plot..... 243
- dta file format..... 290
- Edit cross-links ..... 78
- Edit mass files ..... 79
- Edit new sequence ..... 78
- Edit sequence..... 75
- Elemental composition ..... 78
- EMBL..... 33
- Enable/disable masses..... 110
- Entrez..... 36
- Enzyme ..... 109
- Enzyme cleavage definitions..... 195
- Expect value..... 145
- Export..... 115
- Export sequence ..... 40
- FastA 23, 32, 34, 38, 40, 41, 76, 139, 158, 171, 264, 293, 294, 295, 308
- Feature table ..... 71
- File formats..... 283, 295
- Files used by GPMW ..... 284

## Index

Find in FastA database.....	127	Mass difference.....	119, 253
Fit mass to sequence.....	120	Mass file.....	16, 17, 80, 286
Fragment analyzer.....	259	Mass list.....	108, 163, 253
Fragment window.....	64	Mass list input files.....	289
Fragmentation.....	225	Mass list matching.....	131
Frames.....	113	Mass search.....	11, 93, 107
Gapped search.....	145	Mass spectrum.....	204, 216
Garnier.....	239	Mass type.....	92, 297
GenPept.....	33, 38	Mass vs. precision.....	113
Glycosylation editor.....	61	mgf.....	223
gpm file format.....	291	mgf file filtering.....	134
Graph commands.....	235	mgf file format.....	290
Graphical fragment mapper.....	231	Missed cleavages.....	101, 194
Graphs.....	235	Mixed mode.....	109
Heat mode.....	223	Modification.....	85
Help.....	20	Modification file . 18, 62, 83, 258, 289	
Highlight profiles.....	56, 288	Modifications.....	60, 109
Highlight residues.....	55	Modified residues.....	299
Highlight sequence.....	46, 53	Modula 5.....	103
Homology search.....	143	Monoisotopic.....	164, 231, 297
HPLC chromatogram.....	204, 212	Motifs.....	55, 129
HPLC index.....	94	MS difference.....	256
Hydrogen exchange.....	194	MS peak analysis.....	253
Hydrophobicity.....	239	Ms/ms.....	204
Import ASCII.....	29	MS/MS fragmentation.....	225
Import from clipboard.....	32	MS/MS search.....	176
inclusion list.....	206	MS/MS setup.....	229
Indexing.....	266	Multicharged.....	109
In-line parameters.....	307	Multiple alignment.....	148
Installation.....	2	Multiple digest mass search.....	173
Integer mass.....	297	Multiple highlights.....	53
Internet.....	27, 35, 40, 294	Multiply charged ions.....	47, 67
IPI.....	293	N- and C-terminal.....	48, 81
Isobaric residues.....	56	N- and C-terminal modification 18, 25	
Isoelectric focusing.....	218	NCBI.....	33, 38
Isotopic distribution.....	210	N-glycan delta search.....	154
Ladder sequencing.....	230	N-glycan find peptide.....	155
License agreement.....	311	N-glycan known base.....	153
License number.....	20	N-glycan predict.....	152
Lighthouse data.....	22	N-glycosylation.....	56, 205, 216, 301
Low mass filter.....	204	Nucleotide.....	38
Main Toolbar.....	47	Nucleotide sequence.....	27, 31
Main window.....	12	Offset number.....	64
Make digest database.....	158	Open sequence.....	23
Make fragment window.....	54	Other digest.....	197
Manual cleavage.....	197	Overlap.....	101, 164
Manual digest.....	196	Overlaps.....	115
Marked residues.....	46, 57	OWL.....	33
Mass conversion.....	67, 68	Partial.....	194

## Index

Peak files .....	287	Search for mass .....	107
Peptide info .....	209	Second pass search .....	170
Peptide list .....	132, 219, 287	Secondary structure prediction .....	239
Peptide mass fingerprint .....	157	Sequence files .....	285
Peptide parameters .....	93	Sequence information .....	64
Peptide residue mass .....	303	Sequence management .....	9
Peptide window .....	201	Sequence offset .....	78
Persistent .....	47	Sequence tag .....	158, 253, 257
Phosphorylations .....	273	Sequence window .....	45
pl 65, 92, 102, 162, 167, 169, 171, 217, 272		Setup .....	91
pl strip .....	218	Setup system parameters .....	91
Picomole calculator .....	44	Signal sequence .....	140
Picomole to mass converter .....	68	Simple modification .....	61
PIR .....	32, 294	Simulated 2-D gel .....	272
PIR/NBRF .....	158, 295	SS button .....	16, 58
pKa values .....	302	SS cross-links .....	58, 205
pkl file format .....	289	SS profile .....	59
Pop-up menu .....	50, 114	Startup .....	307
Post-translational modifications ..	299	Strict residue check .....	61
Precision .....	92, 93, 168	Swiss-Prot .....	32, 34, 158, 264, 293
Predict SS cross-links .....	205, 214	Synchronize windows .....	205
Pre-screen mass list .....	110	SYSTEM .....	98
Primer multiplicity .....	241	System colors .....	96
Print .....	16, 51, 117, 236	Tab delimited .....	94
Process FastA .....	139	Tables .....	17, 297
Protein Explorer .....	42	Technical support .....	22
Protein mass search .....	174	Terminals .....	195
proxy .....	i, 35, 39, 106	Toolbar ...	13, 47, 112, 169, 172, 203, 231, 235
Proxy .....	39	TREMBL .....	33
Quick conversion .....	263	Trypsin .....	196
QuickColor .....	56	Underline residues .....	57
QuickDigest .....	195	Underline sequence .....	46
Reformat sequences .....	140	Unimod .....	85
Remove partials .....	207	Un-installation .....	6
Replace residue .....	62	Upgrades .....	22
Report .....	115	Upgrading .....	6, 21
Retrieve sequence list .....	36	USER .....	98
Roepstorff notation .....	225	User configuration .....	106, 308
Save .....	28	User graph .....	242
Save as .....	28	Users .....	106
Save sequence .....	28	Utilities .....	253
Score table .....	166	Various searches .....	127
Scoring digest mass .....	157	VMS to DOS conversion .....	270
Scoring parameters .....	101	What if? .....	205
Search all windows .....	117	Wizard .....	159
Search FastA .....	129	World Wide Web .....	35
Search for composition .....	141	X/Y table .....	256