

Page 1





GPMAW version 5.10

Page 2Letter from the editor
Extracting sequence tags**Page 3**In the works
Databases and BLAST**Page 4**Upgrading
Annotation and Swiss-Prot


GPMAW version 5.10

Welcome to the third issue of *From the Lighthouse* detailing the release of version 5.10.


Protein Explorer: In addition to the usual 'Open sequence' dialog, you can now use a dialog with a tree-like structure (accessed

through the main toolbar ) . The protein explorer is particularly efficient when you have a large amount of sequence data that you have to navigate. The dialog opens with a list of all sequence files in the currently selected 'user directory'. Files that contain a single sequence are indicated by  while multiple sequence files are shown by  . By default the files are 'collapsed'. When clicking on the  icon the file name expands and shows the contained sequences. Selecting a sequence name shows the details in the right-hand panel. You can open a sequence either by selecting the 'Open' button or double-click on a sequence name. The 'Amount' page has a 'picomole calculator (see below), and the 'Directory' page allows you to change the active directory.

Picomole to weight calculator: Although the conversion between molar amounts and weight is trivial, it is still prone to error. The sequence information window and the Protein Explorer both now have a small calculator. In the sequence information window the calculator has been placed on the 'Masses' page. You select the conversion type (e.g. pMol to ng or mg to nMol) followed by entering the actual value. The conversion is carried out at each key press. Values of 1, 2 and 5 are 'pre-calculated'. Conversion type can be changed at any time. In the Protein Explorer the calculator has a separate page in the right-hand sidebar ('Amount'). The calculations are based on the currently selected protein, meaning that you can quickly calculate the same amount of different proteins.

Sequence tag: Extracting a sequence tag from an ms/ms or PSD spectrum can be tricky if you have many peaks. The sequence tag facility of GPMAW has been improved dramatically with this release. It can be accessed through the Utilities | Mass analysis menu or use the  button in the toolbar. For details see page 2.

Carbohydrate editor: A carbohydrate editor is now included that makes it straightforward to add a glycosylation to your sequence. It can be accessed either through the pop-up window of the sequence window (point to the residue to be modified, right-click and select Simple modification | Glycosylation) and when editing

a modification file ( button). In both cases it works as small wizard that takes as first input the kind of glycosylation (N- or O-), adds a core region, and finally lets you modify the final structure, which is then added as a standard modification (e.g. name and elemental composition).

Alpha-helical wheel: You can now make a graphical representation of parts of your sequence as amphiphatic alpha-helices. The function is present on the Graph section of the main menu.

Minor change department:

Settings from a larger number of windows are now saved automatically between runs.

Bruker and MoverZ mass files in XML format can now be loaded directly in mass list input.

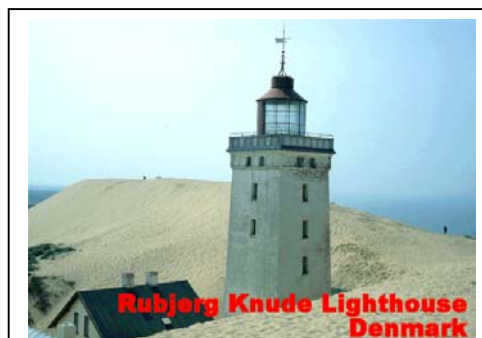
The definition line for defining cleavages has been extended.

The report function for mass search has been improved.

In 'Digest mass search' you can now directly search a FastA formatted database without first generating the index files (which will, however, be generated on the first search).

You can now quickly define 'simple modifications' in the sequence window by pressing shift combined with right-click.

The peptide window now has local menu item enabling you to save the MH+ list to the clipboard.

**Rubjerg Knude Lighthouse.**

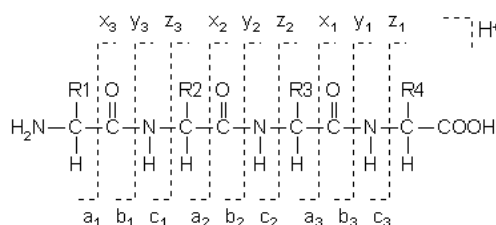
This fascinating lighthouse was built in 1900. Drifting sand created dunes between the lighthouse and the North Sea that got higher and higher. In 1953 the foghorn could no longer be heard at sea and was closed down. Even though attempts were made to dig out the sand, the dunes got higher than the light by the late sixties and the lighthouse was closed in 1968. A museum was made in part of the buildings, but as the sand is still moving, the museum had to be closed as the buildings were slowly covered by sand.

Extracting sequence tags

A sequence tag is usually understood as a short peptide sequence obtained by ms/ms analysis of a peptide, usually obtained by electrospray MS. An alternative way of acquiring a sequence tag is the use of post source decay (PSD) by MALDI MS analysis.

Interpreting a sequence tag can be quite laborious, as the spectrum usually contains a large number of fragment ions arising from fragmentation from either terminals, side-chain loss, multiple fragmentations etc.

The newly introduced Chemically Assisted Fragmentation (CAF) technique pioneered by T. Keough (Keough et al.) and commercialized by Amersham Biosciences results in a very simplified and easy-to-interpret spectrum (see Hellman & Bhikhabhai). This method has the advantage that the resulting spectra almost exclusively consist of y-ions (see Roepstorff & Fohlmann).



The Roepstorff notation for peptide fragmentation.

Extracting a sequence tag can be greatly simplified by GPMW. The sequence tag routine of GPMW uses a recursive algorithm (i.e. an algorithm that calls itself). First the algorithm looks for mass differences corresponding to amino acid residues as defined by the currently loaded mass file. Then it tries to extend the first residue found by more residues. If more than one possibility exists, it will call itself and thus add another tag to the list. Each time a mass residue is 'used' it will be marked as such, so as not to start out a new tag using this residue.

In this way up to 100 sequence tags can be created and reported to the user. If you have a very complex spectrum or your precision is set too wide, you may exceed this number, and will thus have either to decrease the mass list or increase the precision of the search.

When you look at your sequence tags, you should always compare them to the original mass spectrum, as the algorithm does not (yet) take mass intensity into consideration, which you will have to do manually. Furthermore, the algorithm does not try to guess at whether the tag represents a b-type or y-type ion (i.e. originating from the N-terminus or the C-terminus).

The display of the sequence tag is controlled

through the window toolbar:



Compact/expanded display. Compact mode displays the tag in 1-letter code and only the mass range and average precision is reported. In expanded mode the tag is in 3-letter code and all mass values used in the tag are displayed.



The residue button switches (in expanded mode only) between displaying the mass difference (i.e. the residue mass) and the difference between the calculated residue and the theoretical mass.

The precision can be changed in the toolbar with the tags being recalculated for every change. The **Sequence sync.** will highlight, if possible, any sequences corresponding to the one currently selected in the tag list in the **top-most** sequence window. Check the **N-term sulf + K as hR** to treat the mass list as originating from a Chemically Assisted Fragmentation spectrum. The N-terminus is sulfonated and lysine residues are calculated as homoarginine. Furthermore, in case of 'holes' in the sequence, GPMW will try to fit in dipeptides containing Gly or Pro as these residues gives rise to low peaks.

GPMW does not have a sequence tag search program, for this you will have to use one of the excellent programs available on the net or from various vendors. However, if you have a sequence tag of sufficient length (i.e. at least five residues, some of which are among the more 'rare' residues) you can perform a direct BLAST search (see also next page).

This search is performed by selecting the sequence tag line to use for the search, right-click and select 'Perform BLAST'. This command comes in two versions 'as y-ions' or 'as b-ions'. You choose the first option when looking at C-terminal fragment ions (always the case when using CAF chemistry) and the second option when looking at b-ions. The BLAST search dialog will open with the selected sequence tag in 1-letter code and parameters set for peptide search ('Expect value' is set to 10000).

The BLAST search option is particularly efficient when using the CAF chemistry due to the long sequence tags generated.

From the pop-up menu you can also either copy the whole list (as displayed) to the clipboard or just copy the selected tag (in 1-letter code).

Refs: Hellman U, Bhikhabhai R. *Rapid Commun. Mass Spectrom.* 2002; **16**, 1851-1859
Keough T, Youngquist RS, Lacey MP. *Proc. Natl. Acas. Aci. USA* 1999; **96**, 7131
Roepstorff P, Fohlmann J. *Biomed Mass Spectrom.* 1984; **11**, 601

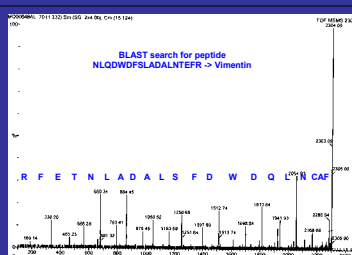
Letter from the editor

This issue of *From the Lighthouse* introduces version 5.10 of GPMW. The jump in version number (from 5.02 to 5.10) is justified by the large number of new features and improvements made in the program. A .1 increase in version number is normally accompanied by a new manual, but the current manual was upgraded recently to reflect most of the changes made up to version 5.02. The next release of the manual will likely be released with version 5.11.

The current release has been a little delayed due to a number of reasons. The main reasons being the summer holiday and a number of small errors that tended to crop up at inconvenient moments. The release of the N-terminal sulfonation chemistry by Amersham Biosciences also contributed as the sequence tag window had to be improved to accommodate this method.

I would like to thank all that have contributed with ideas for this version of GPMW (and later releases). Some of the suggestions have not made it yet. They are not forgotten, but are stored in the 'future improvements' box. So keep the suggestions coming. If anyone would like to contribute to *From the Lighthouse* or have suggestions for themes to cover in the next issue, please contact me by e-mail (php@bmb.sdu.dk).

Peter Højrup



In the works

What can we expect of the next version(s) of GPMaw?

The program is developed in close collaboration with the Protein-Research Group at the University of Southern Denmark, so a lot of the input to the program comes from the research carried out here. However, if you have any suggestions, please contact Lighthouse data. I try to answer all mails, but even if your suggestions do not appear immediately, they are archived and are likely to show up in later releases

Last time I 'promised' an improved ms/ms analysis part. I am sorry to say that I haven't gotten around to that yet. More requests of this kind has arrived, so I will have to .get going. My only excuse is that I got side-tracked into the extraction of sequence tags, as outlined in these pages.

The mass search and the peptide mass search was also promised an improvement, which to some degree has happened.

Next time you can expect to see improvements in the database handling. Currently a database has to be indexed in several ways, each in a different location of the program. I will try and 'collect' all database handling in a single place. I can probably not change the number of indices needed (some, like BLAST, is an independent program called by GPMaw) but centralizing the handling should ease the pain slightly.

I will also try to make the program more interactive with other programs (i.e. data analysis) and put in more graphs to illustrate the mass spectrometry.

Finally, if you work at peptide mass searching, you should have a look at the peak list pre-processor PeakErazor, which will be available soon as a free download from the web site. This program will also be incorporated into GPMaw.

Databases and BLAST

FastA formatted sequence databases are so-called flat-file databases, which means that they are text files where one record just follows the next without any index or size constraint on each record.

On order to use these file efficiently, they have to be indexed. To use them in different ways (e.g. text search and peptide mass search) you have to create different or very complex indices. As for example the BLAST search routine is an external program (supplied by NCBI) it has not been possible to generate a single index, and the multiple index route has been chosen.

Three kinds of indices are created by GPMaw on FastA formatted databases:

- 1) Text indices for performing a text search on the name and species lines. This enables a fast retrieval of sequences from databases.
- 2) Peptide mass search indices. These are actually a conversion of the database into a 'digested' mass database. In parallel with the 'digest', an index-file into the database proper is also generated, so a peptide mass hit is able to find the original sequence when requested.
- 3) BLAST also make a derivative database. The exact structure is not revealed.

To further confuse the matter, the different indices are generated in different locations. Furthermore, a FastA database is not just a FastA database as different formats are present, and different computer systems operate with different file formats.

This means that when you download a database from the Internet, the database has to be pre-processed before GPMaw can use it. For this purpose a small program called Dbindexer has been made and is delivered as standard with GPMaw, and is called from the Utilities menu. The first thing to do with a new database is to perform a conversion: Select 'Convert file', then select the 'Maximum name length' (has to be less than 255, 200 is usually an OK value) and finally press the 'Convert' button. You then select the database file, either enter your own name or just accept the one suggested by GPMaw. This will quickly (a minute or two) generate a processed file for GPMaw (correct long name lines, conversion from Unix format, sequences in a single standard).

Generating the indices is then performed in various places:

- 1) Text indices are performed on the 'Index database' page of Dbindexer. Here you first press the 'Load database' button at the bottom of the window and select the newly converted database. Then you click on the 'Index database' button. You also have the

option of combining several databases before indexing and further handling. The text indices generation takes quite a while for large databases (5-20 minutes for the NCBI nr database).

- 2) Peptide mass search indices are generated by a wizard activated through the 'Setup' | 'Make digest database' option in the GPMaw main menu. The wizard asks you for: Mass file (i.e. chemistry – usually derivatization of Cys), file locations, enzyme used for cleavage and finally you can enter a comment. The file generation is quite fast (1-2 minutes). You can also perform a peptide mass search without first generating a derived database, in this case the generation takes place the first time and is saved for later searches. **Note:** You have to generate a 'derived' database for each chemistry (i.e. for each enzyme and cys-derivatization you use).
- 3) BLAST also needs its own conversion. This takes place in the 'System setup' dialog on the BLAST page. Press the 'Format' button, select the FastA database to format and press 'OK'. **Note:** This operation is performed by a so-called 'console' application that works in a DOS box. This will be running minimized (you can see it appear in the task bar) and GPMaw will monitor its progress. Do NOT close the Setup dialog while the 'Generating database' text flashes. When the derivatization of the database is complete, a dialog asks you to add the database to the list of available databases. – in general you should answer 'Yes'.

NOTE: Whenever you download a new database to replace an old one, you have to delete all indices previously generated. For this reason it is advantageous to place each database in its own directory (conveniently below the \gpmaw\database\ directory). You can then delete all files in the directory prior to download of the new database.

Note on BLAST.

The BLAST function as implemented by GPMaw is a call to an external program called 'blastall.exe' (forms part of the NCBI suite of BLAST programs). This program should be located in the \gpmaw\bin\ directory (along with 'gpmaw3.exe'. If it located in a different location, you will have to tell GPMaw the location by pressing the 'Install BLAST' button on the Setup page. navigate to the correct location and select the 'blastall.exe' file.

If you have ready-derived blast database files, you can add them to the list through the 'Add database' button (likewise entries can be deleted by pressing the 'Delete entry' button.

Upgrading

Included in a license of GPMW is the right to upgrade your program to the latest version within one year of purchase. Current releases of the program are coded to accept licenses that are up to 18 month old. The reason for this is that OEM versions of the program may be several month underway before reaching the end-user.

You can check whether your copy of GPMW can be upgraded by opening the 'About' box (Help | About). In the middle of the window you read 'License date:' followed by the month and year of your license. If the current release is within 18 month of this date, you should be able to upgrade.

The upgrade is easily performed if you have access to the Internet. Point your web browser at <http://welcome.to/gpmaw>, go for the 'Update' button and locate the update to most recent version of the program. Click on the name of the download, and when asked whether to download answer 'Yes' and specify the download location.

The upgrade is an executable file that you just double-click from 'Explorer'. The install program searches your disk drive for the present location of GPMW, and if found you can just accept the default for upgrading.

If the program does not find your copy of GPMW you will have to specify a location where the program will be located. From here you have to move the two files "gpmaw3.exe" and "gpmaw3.hlp" to replace the files with the same name. The default location of GPMW is C:\gpmaw\bin\.

If you do not have access to the Internet, you will have to contact Lighthouse data to obtain an upgrade on CD-ROM. Remember to specify your license number.

If you want to upgrade and your license is too old, you can upgrade to the latest version for US\$ 120.-. This represents 50% saving compared to the price of a full version of the program. If you need additional copies you may buy them with a saving of 25%. Prices includes postage and handling.

We now accept credit cards from MasterCard, EuroCard and VISA (not available in Denmark). Please contact Lighthouse data for information.

Annotation and Swiss-Prot

One way to use the **annotation** page is as a free text field where you can write any kind of text that will be saved along with the sequence. However, if the annotation contains a Swiss-Prot record like this:

```
ID CONG_BOVIN STANDARD; PRT; 3
AC P23805;
DT 01-NOV-1991 (Rel. 20, Created)
DT 01-FEB-1994 (Rel. 28, Last sequence up
DT 01-OCT-1996 (Rel. 34, Last annotation
DE CONGLUTININ PRECURSOR.
...
FT DISULFID 275 369 BY SIMILA
FT DISULFID 347 361 BY SIMILA
SQ SEQUENCE 371 AA; 37994 MW; 867BB41
MLLLPLSVLL LLTPQWRSLG AEMTTFSQKI ...
GEKGDGSPG PAGRAGRPGW VGPIGPKGDN ...
GKQSGMPPG TPGPKGETGP KGVGAPGIQ ...
GAIGPQGPSP ARGPPGLKGD RGDGPGETGAK ...
QYKKAFLFPD QQAVGEKIFK TAGAVKSYSD ...
QEKNAFLSMN DISTEGRFTY PTGEILVYSN ...
CSKQLLVICE F
```

//

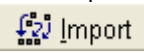
(note the central part has been omitted) it will be recognized by GPMW and given special treatment.

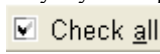
If the record contains a feature table (recognized by lines beginning with FT followed by three spaces) it will be dissected and entries that can be transferred to the sequence by GPMW will be transferred to the separate 'Feature table' page.

Annotation	Feature table		
From	to	Type	Description
<input type="checkbox"/>	1-	20	SIGNAL
<input type="checkbox"/>	21-	371	CHAIN CONGLUTININ.
<input type="checkbox"/>	46-	216	DOMAIN COLLAGEN-LIKE (G-X-Y).
<input type="checkbox"/>	273-	371	DOMAIN C-TYPE LECTIN (SHORT FORM).
<input type="checkbox"/>	63-	63	MOD_RES HYDROXYLATION.
<input type="checkbox"/>	87-	87	MOD_RES HYDROXYLATION.
<input type="checkbox"/>	99-	99	MOD_RES HYDROXYLATION.
<input type="checkbox"/>	135-	135	MOD_RES HYDROXYLATION.
<input type="checkbox"/>	141-	141	MOD_RES HYDROXYLATION.
<input type="checkbox"/>	159-	159	MOD_RES HYDROXYLATION.
<input type="checkbox"/>	162-	162	MOD_RES HYDROXYLATION.

Unrecognized features are shown in gray and cannot be selected.

From this page you can select to have features transferred to the sequence, either individually or all at the same time. Transferring individual modification is done by first marking the relevant checkboxes and then press the **Import** button

 **Import**. You can check all features simultaneously by first pressing the **Check**

all button  **Check all**. You may revert to the unmodified protein by pressing the **Reset**

button  **Reset**.

Note: GPMW will import features based on the residue positions listed in the 'From-to' column. This means that if you already have imported a feature that has changed the length of the sequence (signal or propeptide), features will not imported into the correct position. GPMW will warn you about this

by writing 'Warning: Sequence length differs from annotation' in red above the feature table. In this case you will have to revert to the full-length sequence before importing the new features.

Currently the recognized features are
SIGNAL (removed)
PROPEP (removed)
DISULFID (create cross-link)
ACETYLYATION (modification)
AMIDATION (modification)
FORMYLATION (modification)
HYDROXYLATION (modification)
PHOSPHORYLATION (modification)
SULFATATION (modification)

You may now add your own modifications to the list by following these rules:

1) The feature entries have to go into the annotation just before the sequence (e.g. the line starting with SQ).

2) Each feature line starts with 'FT' followed by three spaces. Then follows the name of the modification spelled as listed in the above table. After some space you enter the start position of the feature, some space and the end of the feature and finally some space and either a comment or the actual modification in the fifth column.

3) If the modification is just in a single position, the start and end will be the same. If the feature is a modification, the second column has to be MOD_RES and the fifth column holds the name of the modification according to the above table.

Swiss-Prot record demands that each column starts in a given position, but apart from the first two characters and three spaces, GPMW does not care as to how many spaces are present.

Entering a phosphorylation for residue number 214 thus has to read:

```
FT MOD_RES 214 214 PHOSPORYLATION
```

If your sequence does not have a Swiss-Prot annotation, but you would like to be able to quickly turn modifications on and off, you can modify the annotation by having the first line start with ID + three spaces. This will make GPMW believe that there is a Swiss-Prot record connected to the sequence. Then on following lines you put the feature lines formed as described above. Then you add SQ SEQUENCE (remember the three spaces), then the sequence on lines starting with five spaces and finally '/' on a line by itself.

Note: You always have to save the sequence after making changes to the annotation page. GPMW will not do this automatically!