

MassXpert User Manual

Polymer chemistry modeling and mass spectrometry simulations

Version: 10.5.12

Filippo Rusconi

Filippo Rusconi, 2026

Version 10.5.12
Published: 02/11/2026

To María Cecilia

To all the admirable people acting in the Free Software Movement *for a better and more ethical computing world*

To all involved in the development of the K Desktop Environment (KDE) *and of the* Plasma *desktop*

To all the readers who helped me with this manual.

Contents

2 Generalities 26

2.1 On Chemical Formulæ and Chemical Reactions 26

2.2 The *MassXpert* Framework Data Format 27

2.3 Chemical Entity Naming Policy 28

2.4 Installation with an AppImage software bundle 28

3 Configuring the application preferences 30

3.1 Setting the application preferences - general options 30

3.2 Setting the application preferences - shortcuts 31

3.3 Setting the application preferences - filesystem hierarchy structure 31

4 Basics in Polymer Chemistry 33

4.1 Polymers? Where? Everywhere! 33

4.2 Various Biopolymer Structures 34

4.2.1 Proteins 34

4.2.2 Nucleic Acids 36

4.2.3 Saccharides 38

4.3 To Sum Up 41

4.4 Polymer Chain Disrupting Chemistry 41

4.4.1 Polymer Cleavage 42

4.4.2 Polymer Fragmentation 45

4.4.2.1 Protein Fragmentation 45

4.4.2.2 Nucleic Acids Fragmentation 48

4.4.2.3 More Complex Patterns Of Fragmentation 51

4.4.2.4 To Sum Up 52

5 XpertDef: Definition of Polymer Chemistries 53

5.1 The isotopes 54

5.2 The polymer chemical entities 55

5.2.1 The monomers 56

5.2.2 The modifications 58

5.2.3 The cross-linkers 60

5.2.4 The cleavage agent definitions 61

5.2.5 The fragmentation pathway definitions 65

5.2.5.1 Simple fragmentation pathway patterns 66

5.2.5.2 More complex fragmentation pathway patterns 67

5.2.5.3 Even more complex fragmentation patterns 70

5.3 Saving the definition 71

6 XpertEdit: a Powerful Editor and Simulation Center 72

6.1 *XpertEdit* Invocation 72

6.2 *XpertEdit* Operation: *In Medias Res* 72

6.3 The Editor Window Menu 77

6.4 Editing Polymer Sequences 79

6.4.1 Multi-Character Monomer Codes 79

6.4.2 Unambiguous Single-/Multi-Character Monomer Codes 80

6.4.3 Erroneous Monomer Codes 81

6.4.4 Simplified Editing 82

6.5 Finding sequence motifs 82

6.6 Importing Sequences 83

6.6.1 Importing From The Clipboard 83

6.6.2 Importing From Raw Text Files 85

6.7 Multi-region Selections 85

6.8 Polymer Sequence Modification 88

6.8.1 Selected Monomer(s) Modification 88

6.8.2 Polymer Modification 95

6.9 Monomer Cross-linking 97

6.10 Sequence Cleavage 98

6.11 Oligomer Fragmentation 101

6.12 Mass Searching 106

6.13 Oligomer Data Filtering 107

6.14 Monomeric And Elemental Composition 109

6.15 pKa, pH, pI and Charges 109

6.15.1 Ionized Group(s) In Monomers 110

6.15.1.1 Some Theory First 110

6.15.2 Ionized Group(s) In Modifications 113

6.15.3 pH, pI and Charge Calculations 113

6.16 General Options 114

7 *XpertCalc*: a Powerful Mass Calculator 115

7.1 *XpertCalc* Invocation 115

7.2 An Easy Operation 115

7.3 The Programmable Calculator 117

7.4 The Log Book Recorder 120

7.5 The *m/z* Ratio Calculator 120

7.6 *m/z* Ratio Calculations 121

8 Isotopic cluster calculations 123

8.1 Generating an isotopic cluster simulation in *MassXpert* 125

8.1.1 Static Standard Element Tables Shipped within the *IsoSpec* Library 128

8.1.2 User-Modified Standard Element Table Shipped within the *IsoSpec* Library 128

8.1.3 User hand-made isotopic configuration of an elemental composition 130

8.1.4 The results of the computation are centroids 134

8.2 Shaping the centroids of the isotopic cluster peaks into well-profiled peaks 136

8.3 Combining multiple shaped isotopic clusters' peaks into a single mass spectrum 142

9 Utilities 147

9.1 Configuration of the *MassXpert* and *MineXpert* intercommunications 147

9.1.1 Dynamic configuration of the *MassXpert* and *MineXpert* intercommunications 148

10 Data Customizations 149

Appendix 157

GNU General Public License version 3 157

Preamble 157

TERMS AND CONDITIONS 158

0. Definitions. 158

1. Source Code. 159

2. Basic Permissions. 160

3. Protecting User's Legal Rights From Anti-Circumvention Law. 160

4. Conveying Verbatim Copies. 161

5. Conveying Modified Source Versions. 161

6. Conveying Non-Source Forms. 162

7. Additional Terms. 163

8. Termination. 165

9. Acceptance Not Required for Having Copies. 165

10. Automatic Licensing of Downstream Recipients. 165

11. Patents. 166

12. No Surrender of Other's Freedom. 167

13. Use with the #acronym(GNU) Affero General Public License. 168

14. Revised Versions of this License. 168

15. Disclaimer of Warranty. 168

16. Limitation of Liability. 169

17. Interpretation of Sections 15 and 16. 169

END OF TERMS AND CONDITIONS 169

How to Apply These Terms to Your New Programs 169

Colophon 171
About the authors 171
Book preparation 171
Index 172

List of Figures

Figure 1 Adresses to send bug reports or questions to 22
Figure 2 Application preferences - General page 30
Figure 3 Application preferences - Shortcuts page 31
Figure 4 Application preferences - Filesystem hierarchy page 32
Figure 5 Peptidic bond formation by condensation 35
Figure 6 End capping chemistry of the protein polymer 36
Figure 7 Phosphodiester bond formation by esterification 37
Figure 8 End capping chemistry of the nucleic acid polymer 38
Figure 9 Osidic bond formation by condensation 40
Figure 10 End capping chemistry of the polysaccharidic polymer 40
Figure 11 Protein cleavage by water and cyanogen bromide 43
Figure 12 Protein fragmentation patterns most widely encountered 46
Figure 13 Peptide product ion mass calculation for the different ion series 47
Figure 14 DNA fragmentation patterns most widely encountered 50
Figure 15 Nucleic acids product ion mass calculation for the different ion series 51
Figure 16 Select one polymer chemistry definition file 53
Figure 17 Polymer chemistry definition window 54
Figure 18 Isotopes definitions 55
Figure 19 Monomer definition 57
Figure 20 Monomer mass differences 58
Figure 21 Modification definition 59
Figure 22 Cross-linker definition 60
Figure 23 Cleavage agent definition 62
Figure 24 Fragmentation definition in the DNA chemical world 66
Figure 25 Basic fragmentation pathway definition 67
Figure 26 More complex fragmentation pathway definition 68
Figure 27 Fragmentation definition with specific rules 69
Figure 28 Selection of a sample polymer sequence 73
Figure 29 Selection of the polymer chemistry definition 74
Figure 30 The *XpertEdit* main window 74
Figure 31 Multi-character code sequence editing in *XpertEdit* 79
Figure 32 Bad code character in *XpertEdit* sequence editor 81

Figure 33 Simplified editing with the monomer list 82
Figure 34 Finding a sequence motif in the polymer sequence 83
Figure 35 Clipboard-imported sequence error-checking 84
Figure 36 Clipboard-imported sequence purification 84
Figure 37 Multi-region selection 86
Figure 38 Multi-region selection 87
Figure 39 Selected monomer modification 89
Figure 40 Modification of all the monomers of the same code 90
Figure 41 Modification of all the monomers selected in the list 90
Figure 42 Modification of all the monomers 91
Figure 43 Modification of all the monomers in a polymer sequence 92
Figure 44 Modification of a monomer in a polymer sequence 93
Figure 45 Rendering of a monomer modification in a polymer sequence 94
Figure 46 Seryl residue modified twice with *Phosphorylation* 94
Figure 47 Modification of the left end of a polymer sequence 96
Figure 48 Cross-linking of monomers 97
Figure 49 Polymer sequence cleavage window 98
Figure 50 Oligomer fragmentation window 102
Figure 51 Starting a fragmentation that accounts the cross-links 103
Figure 52 Complicated cross-linking situation 104
Figure 53 Searching masses in a polymer sequence 107
Figure 54 Oligomer data filtering 108
Figure 55 Determination of the compositions 109
Figure 56 Different pKa values of chemical groups 110
Figure 57 Acido-basic computations: net charges 114
Figure 58 Selecting a polymer chemistry definition for use with *XpertCalc* 115
Figure 59 Interface of the *XpertCalc* module 116
Figure 60 Interface of the chemical pad 118
Figure 61 The *XpertCalc* recorder window 120
Figure 62 Calculation of ranges of m/z ratios 121
Figure 63 Simulation of the isotopic cluster for a protein 124
Figure 64 Isotopic clusters main menu 125
Figure 65 Isotopic cluster generator window 126
Figure 66 User-modified standard element table 129
Figure 67 Hand-made user isotopic configuration of an elemental composition131

Figure 68 Hand-made user isotopic configuration of an elemental composition 131
Figure 69 Typical manual configuration of the isotopic composition of the chemical elements 133
Figure 70 The user manual configuration might be exported to a text file and loaded back 134
Figure 71 Results from the isotopic cluster calculation 135
Figure 72 Setting-up of the centroid mass peak shaping process 137
Figure 73 Shaped peaks mass spectral data 140
Figure 74 Isotopic cluster automatically displayed in *MineXpert* 141
Figure 75 Combination of mass spectra obtained for different isotopic cluster calculations 143
Figure 76 Reusing the isotopic cluster shaper dialog window 144
Figure 77 Ensuring that the bins are "phased" in the various isotopic cluster mass spectra 145
Figure 78 Avoiding the combination artifact due to having "unphased bins" 146
Figure 79 Utilities 147
Figure 80 Setting-up of the centroid mass peak shaping process 147
Figure 81 Configuration of the dynamic communication between *MassXpert* and *MineXpert* 148
Figure 82 The polymer chemistry definition directory 152

Revision history

10.5.6 September 2026 fr

Removed the documentation about the JavaScript module which is no more part of *MassXpert* because it's been too much of a hurdle maintaining this scripting framework starting from Qt6 because the Qt crew have shoven QML down the scripting path making supporting scripting too time consuming and exhausting an endeavour.

10.1.2 November 2025 fr

Full port of *MassXpert* to the new *XpertMass* project's libraries. Add new chapter on JavaScript features now in *MassXpert*.

7.0.0 November 2022 fr

The manual has not any content changed. Only this revision corresponds to a software bump matching a full code port to Qt6.

6.1.0 October 2022 fr

Entirely reworked the user manual to make it more modern from the coding point of view. The users do not see any difference with the previous version.

6.0.3 20210202 fr

Removed all mml elements (MathML) that made rendering the document too difficult. Replaced the formulae with figures. Adapted the text;

5.7.0 20190115 fr

Finished porting of the historical LaTeX-based documentation to the DocBook/DAPS/FOP publishing system. Please, see Colophon for details;

not_set 20150918 fr

Update the manual to reflect the changes in the way ion product masses are computed in massXpert and update the various examples. Now, the fragmentation specification must produce fragments that are uncharged. The software will then ionize the fragments using the ionization rule currently in use in the XpertEdit module;

not_set 20121205 fr

Update the manual to document the various new ways to import data in the m/z list in the XpertMiner module;

not_set 20111022 fr

Update the manual to describe the new fragmentation of oligomer containing cross-linked monomers feature;

not_set 20110829 fr

Update the manual to describe a number of new features;

not_set 20100429 fr

Updated the section about the definition of monomers to document the new feature about computing the mass difference between any two monomers in the definition;

not_set 20100427 fr

Updated the section about the cleavages of polymer sequences (XpertEdit chapter) to document the new feature that allows to only perform a cleavage in the currently selected region of a polymer sequence;

not_set 20090619 fr

Updated the section about the chemical pad (XpertCalc chapter) to reflect improvements in the use of the chemical pad buttons (either immediate evaluation of the formula or mere insertion of the formula in the formula line edit widget);

not_set 20090617 fr

Updated the section about the chemical pad (XpertCalc chapter) to reflect improvements in the graphical display of the chemical pad buttons as programmed in the chem_pad.conf configuration files;

not_set 20090401 fr

Although not visible in the documentation, I wanted to publically extend my warm thanks to Lionel Élie Mamane, who helped me along these last months with the Debian packaging of massXpert. Note that this work also proved useful for other areas in the project.

not_set 20090220 fr

Updated the XpertEdit chapter to show how to configure the options about the number of decimals to be used for display of numerals in the program;

not_set 20090205 fr

Updated the XpertEdit chapter to show the simplified polymer sequence editing feature whereby editing of the sequence might be performed by clicking on monomer items in the list of all the monomers defined in the polymer chemistry definition. Fixed small bug in the documentation of about multi-region selection behaving as oligomers or residual chains.

not_set 20081211 fr

Updated the XpertEdit chapter to show the feature by which it is now possible to force the calculation engine to take into account the left/right end modification(s) when calculating the masses of a sequence region that does not encompass the left/right end of the polymer sequence. This new feature was essential in trying to perform full simulations of the molecular heterogeneity of the telokin protein (Rusconi et al. 1997 Biochemistry). Added a paragraph about max count of chemical modifications of a given monomer at once in the XpertDef chapter and another one in the XpertEdit chapter to explain its working;

not_set 20080911 fr

Finally indexed the whole document. Performed some minor modifications so that the documentation system does not produce HTML files anymore (the HTML production was not really worth it anyways);

not_set 20080805 fr

Updated the user manual with a bunch of updated screen shots;

not_set 20080730 fr

Updated the XpertMiner chapter to illustrate the m/z−z mass list matching feature;

not_set 20080708 fr

Updated the XpertEdit chapter to illustrate the new multi-cleavage feature;

not_set 20080701 fr

Updated the XpertEdit chapter to illustrate the new multi-region and multi-selection features;

not_set 20080529 fr

Changed the install instructions for the Mac OS X system;

not_set 20080527 fr

Added a chapter about XpertMiner to document the new features in that module. Some fixes here and there.

not_set 20080526 fr

Modified the documentation to reflect switch to version 3 of the GNU General Public License;

not_set 20080425 fr

The installation chapter was updated to illustrate the installation of the software in the Mac OS X system;

not_set 20080424 fr

The installation chapter was updated to reflect the changes in the way the package might be installed (the package is now relocatable, provided the user indicates where the directories are located);

not_set 20080402 fr

The XpertDef chapter was updated to detail the new way of defining fragmentation specifications where the side chain is decomposed in the gas-phase. The section about fragmentations is now much better documented;

not_set 20080330 fr

The XpertEdit chapter was updated to include a description of the new fragmentation/mass searching data in-place filtering. A section is now devoted to data filtering;

not_set 20080325 fr

The XpertEdit chapter was updated to include a description of the new sequence cleavage data in-place filtering;

not_set 20080318 fr

The XpertEdit chapter was updated to include a description of the find sequence motif feature;

not_set 20080313 fr

The XpertEdit chapter was updated to include a new paragraph about monomer cross-linking as this is now implemented in the software;

not_set 20080221 fr

The XpertEdit chapter was updated to include a new figure of the polymer modification procedure and to describe the enhanced modification procedure;

not_set 20080215 fr

The XpertEdit chapter was updated to include a new figure of the monomer modification procedure and to describe the enhanced modification procedure;

not_set 20071217 fr

The chapter about installation of massXpert has been rewritten to reflect changes in the building of massXpert and in the installation of Debian GNU/Linux and Fedora core GNU/Linux packages;

not_set 20071216 fr

The chapter about polymer chemistry definitions has been refactored to reflect the rewriting of the corresponding code. Added a small section about m/z ratio calculation, that was missing, although the feature was added a long time ago;

not_set 20070922 fr

The new multi-charged cleavage and fragmentation oligomers have been documented;

not_set 20070819 fr

Switched back to version 2 of the GPL in the Appendices chapters, as massXpert cannot be licensed otherwise: the Qt libraries are licensed using version 2 of the GPL without the "or any later version, at your option" wording;

not_set 20070728 fr

Updated the XpertDef chapter (modifications) to show the new "targets" feature. Updated the XpertEdit chapter to show the new monomer modification dialog.

not_set 20070719 fr

Added explanation on the arbitrary formula-based polymer sequence ends modification.

not_set 20070713 fr

Revision of the whole document for a better printed output;

not_set 20070710 fr

Added a section to the XpertEdit chapter about the data mining mass list lab feature added recently. Mentioned the installation of Debian packages;

not_set 20070630 fr

Switched the project and all accompanying documentation and other data files to GPL version 3;

not_set 2007 mid-june fr

Start of the writing by taking inspiration of the GNU polyxmass manual.

1 Preface

1.1 Software feature offerings and intended audience

This manual is about the *MsXpertSuite* mass spectrometric software suite, a software environment that contains two modules:

- *MassXpert*: a module that allows users to define brand new polymer chemistries. The polymer chemistry definitions are then used to model linear polymer sequences. Once modelled, a polymer sequence can undergo chemical reactions (enzymatic or chemical cleavages, gas-phase fragmentations...). The obtained results are a model of what a mass spectrum would look like if the modelled experiment had actually been carried over up to the mass spectrometry analysis.
- *MineXpert*: a module that allows users to load mass spectrometry data from mass spectrometry data files. The loaded data are then visualized and explored throughout all their MS^n depth. The mass data visualization starts either at the TIC chromatogram level or the MS data table view and deepens to the MS^n mass spectra, the drift spectra, the XIC chromatograms, various color maps relating *m/z* values with either retention times or drift times. A wide array of mass spectral integrations are available to ease the MS^n data finest scrutiny.

This manual is therefore intended for people willing to learn how to use the comprehensive *MsXpertSuite* software package.

Mass spectrometry has gained popularity across the past twenty years or so. Indeed, developments in polymer mass spectrometry have made this technique appropriate to accurately measure masses of polymers as heavy as many hundreds of kDa, and of any chemical type.

There are a number of utilities—sold by mass spectrometer constructors with their machines, usually as a marketing "plus"—that allow predicting/analyzing mass spectrometric data obtained on polymers. These programs are usually different from a constructor to another. Also, there are as many mass spectrometric data prediction/analysis computer programs as there are different polymer types. You will get a program for oligonucleotides, another one for proteins, maybe there is one program for saccharides, and so on. Thus, the biochemist/massist, for example, who happens to work on different biopolymer types will have to learn to use several different software packages. Also, if the software user does not own a mass spectrometer, chances are he will need to buy all these software packages.

The *MsXpertSuite* mass spectrometric software is designed to provide *free* solutions to all these problems by providing the following features:

- *MassXpert*:
  - Model *ex nihilo* polymer chemistry definitions (in the *XpertDef* module that is part of the *MassXpert* program);
  - Peform simple yet powerful mass computations to be made in a mass desktop calculator that is both polymer chemistry definition-aware and fully programmable (that's the *XpertCalc* module also part of the *MassXpert* program);
  - Edit polymer sequences on a polymer chemistry definition-specific basis, along with chemical reaction simulations, finely configured mass spectrometric computations... (all taking place in the *XpertEdit* module that is the main module of the *MassXpert* program);
  - Customize the way each monomer will show up graphically during the program operation (in the *XpertEdit* module);
  - Edit polymer sequences with immediate visualization of the mass changes elicited by the editing activity (in the *XpertEdit* module);
  - Open an unlimited number of polymer sequences at any given time and of any given polymer chemistry definition type (in the *XpertEdit* module);
- *MineXpert*:
  - Load mass spectometry data files in the mzML format, thanks to the excellent *libpwiz* library of ProteoWizard[1] fame; mass data file loading can be performed in *full-memory* mode (all the data read from file are stored in memory) for faster operations or in *streamed* mode when loading files largers than the available memory.
  - Display the data in powerful ways in a unified graphical user interface. The interface was designed to integrate all the most useful characteristics of the various proprietary environments known by the author, thanks to the excellent *libqcustomplot* library;[2]
  - Configure the way mass spectrometry data integrations are performed from a combination standpoint and optionally apply a Savitzky-Golay smo
  - Configure the level of MS data (MS^n) for which the integrations are performed. Optionally filter data by precursor ion *m/z* values or precursor spectrum indices;

[1] proteowizard.sourceforge.net
[2] qcustomplot.com

- Perform data mining by performing data integrations in various ways;
- Ion mobility mass spectrometry data are supported with a int=f(*m/z*,dt) color map plot calculation;
- A specific data integration mode allows easy mass spectral intensity determination for any feature (peak in a TIC chromatogram, a mass spectrum, a drift spectrum, any color map region);

1.2 Feedback to the developers

We are always grateful to any constructive feedback from the users.

The MsXpertSuite software team might be contacted *via* the following addresses:

```
msxpertsuite@msxpertsuite.org = general mailing list about msXpertSuite
bug-reports@msxpertsuite.org  = report bugs found in msXpertSuite software
```

Figure 1: Adresses to send bug reports or questions to

Adresses to send bug reports or questions to.

1.3 Brief history of *MsXpertSuite*.

- *1998–2000*

  The name *MassXpert* comes from a project I started while I was a post-doctoral fellow at the École Polytechnique (Institut Européen de Chimie et Biologie, Université Bordeaux 1, Pessac, France);

  The *MassXpert* program was published in *Bioinformatics* (Rusconi, F. and Belghazi, M. *Desktop prediction/analysis of mass spectrometric data in proteomic projects by using massXpert. Bioinformatics*, 2002, 644–655.

  At that time, *MS-Windows* was at the *Windows NT 4.0* version and the next big release was going to be *you will see what you'll see*: *MS-Windows 2000*.

  When I tried *MassXpert* on that new version (one colleague had gotten it with a new machine), I discovered that my software would not run normally (the editor was broken). The Microsoft technical staff would advise to *buy a new version of the compiler environment and rebuild*. This was a no-go: I did not want to continue paying for using something I had already produced with legitimate software.

- *2001–2006*

  During fall 1999, I decided that I would stop using *Microsoft* products for my development. At the beginning of 2000 I started as a CNRS research staff in a new laboratory

and decided to start fresh: I switched to *GNU/Linux* (I never looked back). After some months of learning, I felt mature to start a new development project that would eventually become an official GNU package: *GNU polyxmass*.

The *GNU polyxmass* software, much more powerful than what the initial *MassXpert* software used to be, was published in *BMC Bioinformatics* in 2006 (Rusconi, F. *GNU polyxmass: a software framework for mass spectrometric simulations of linear (bio-)polymeric analytes).* BMC Bioinformatics, *225–*.

Following that publication I got a lot of feedback (very positive, in a way) along the lines: *—Hey, your software looks very interesting; only it's a pity we cannot use it because it runs on GNU/Linux, and we only use MS-Windows*.

- *2007–2016*

  In december 2006, I decided to make a full rewrite of *GNU polyxmass*. The software of which you are reading the user manual is the result of that rewrite. I decided to "recycle" the *MassXpert* name because this software is written in C++, as was the first *MassXpert* software. Also, because the first *MS-Windows*-based *MassXpert* project is not developped anymore; taking that name was kind of a "revival" which I enjoyed. However, the toolkit I used this time is not the Microsoft Foundation Classes (first *MassXpert* version) but the Trolltech *Qt* framework (see menu **Help⇨About Qt** in *MassXpert*).

  Coding with the *Qt* libraries has one big advantage: it allows the developer to code once and to compile on the three main platforms available today: *GNU/Linux*, *MacOSX*, *MS-Windows*. Another advantage is that the *Qt* libraries are wonderful software, technically and philosophically (Free Software).

  > **Note**
  >
  > I have abandoned supporting macOS since a number of years because the system is too painful to support.

  The rewritten software was published in 2009 (Rusconi, F. *massXpert 2: a cross-platform software environment for polymer chemistry modelling and simulation/analysis of mass spectrometric data.Bioinformatics*, 2009, 2741–2742.

- *2016–2019*

  In 2016, I started a new project about visualization of mass spectrometric data. The project developed pretty quickly, as we needed at the mass spectrometry facility a software that would allow to cope efficiently with ion mobility mass spectrometric experimental data. *MineXpert* was thus started.

To bundle both *MassXpert* and *MineXpert* in a single software suite, I bought the *MsXpertSuite* domain name http://msxpertsuite.org and created that new name.

- *2019–2021*

  The *MineXpert* software was published in 2019. The reviewers acknowledged that it was a useful piece of software but complained about its not handling the MS/MS data. Reference to cite: Filippo Rusconi. mineXpert: Biological Mass Spectrometry Data Visualization and Mining with Full JavaScript Ability. *J. Proteome Res.* 2019, 18, 5, 2254-2259.

In response to the reviewers and for my personal project that I had started a little earlier, I decided to fully rewrite *MineXpert* into *MineXpert*. The new version of the software had to support MS^n data.

- *2021–2022*

  The *MineXpert* software was published in 2019. The main new capability is now the full handling of MS^n data. This new feature required the full rewriting of the software internal data model. See Langella O., Rusconi F. (2021) mineXpert2: Full-Depth Visualization and Exploration of MS^n Mass Spectrometry Data. *J. Am. Soc. Mass Spectrom.*, 4 (32) 1138-1141.

- *2022–2025*

  A major work is now undertaken: the native handling of the timsTOF data from the Bruker instrument vendor. This kind of data contains ion mobility data.

- *2025–2026*

  Rewrite of major portions of the software to ease evolution of the code base, in particular with a strong enhancement of code reuse by putting large portions of code in the *libXpertMassCore* and *libXpertMassGui* libraries, for use by *MineXpert* and *MassXpert* .

- *2026–*

  Following the almost full rewrite, development of new features has resumed, all along regular feature requests by the users of the software. In particular, data processing features have been added to &MineXp;, such as a new automatic deconvolution of low mass analytes with analyte identification (with fully labeled analytes handling), peak labeling management in the plot widgets, new plot management inside a plot widget, network-based communication between *MineXpert* and the sister application *MassXpert* .

1.4 Program and Documentation Availability and License

The programs and all the documentation that are shipped along with the *MsXpertSuite* software suite are available at http://www.msxpertsuite.org. Most of the time, a new version is published as source, and as binary install packages for *MS-Windows* (64-bit systems only). No *GNU/Linux* binary packages are created outside of the autobuilder of the various distributions. As a Debian Developer, the author creates *Debian*[3] packages that are uploaded on the distribution servers. These packages are available using the system's software management infrastructure (like using the *Debian*'s `apt` command, for example, or the graphical application).

The software and all the documentation are all provided under the Free Software license *GNU General Public License, Version 3, or later, at your option*. For an in-depth study of the *Free Software* philosphy, I kindly urge the reader to visit http://www.gnu.org/philosophy.

[3] http://www.debian.org/

2 Generalities

In this chapter, I wish to introduce some general concepts around the *MassXpert* program and the way data elements are named in this manual and in the program.

The *MassXpert* mass spectrometry software suite has been designed to be able to "work" with every linear polymer. Well, in a certain way this is true... A more faithful account of the *MassXpert*'s capabilities would be: "*The* MassXpert *software suite works with whatever polymer chemistry the user cares to define; the more accurate the polymer chemistry definition, the more accurate* MassXpert *will be*".

For the program to be able to cope with a variety of possibly very different polymers, it had to be written using some *abstraction layer* in between the mass calculation engine and the mere description of the polymer sequence. This abstraction layer is implemented with the help of "polymer chemistry definitions", which are files describing precisely how a given polymer type should behave in the program and what its constitutive entities are. The way polymer chemistry definitions are detailed by the user is the subject of a chapter of this book (see menu **XpertDef** of the program). However, in order to give a quick overview, here is a simple situation: a user is working on two polymer sequences, one of chemistry type "protein" and another one of chemistry type "DNA".

The protein sequence reads "ATGC", and the DNA sequence reads "CGTA".

Now imagine that the user wants to compute the mass of these sequences. How will *MassXpert* know what is the formula (hence mass) of each monomer code in each polymer sequence? There must be a way to inform *MassXpert* that one of the sequences is a protein while the other is a DNA oligonucleotide: this is done upon creation of a polymer sequence; the programs asks of what chemistry type the sequence to be created is. Once this "chemical parentage" has been defined for each sequence, *MassXpert* will know how to handle both the graphical display of each sequence and the calculations for each sequence.

2.1 On Chemical Formulæ and Chemical Reactions

Any user of *MassXpert* will inevitably have to perform two kinds of chemical simulations:

- Define the formula of some chemical entity;
- Define a given chemical reaction, like a protein monomer modification, for example.

While the definition of a formula poses no special difficulty, the definition of a chemical reaction is less trivial, as detailed in the following example. The lysyl residue has the

following formula: $C_6H_{12}N_2O$. If that lysyl residue gets acetylated, the acetylation reaction will read this way:

> — *"An acetic acid molecule will condense onto the ε amine of the lysyl side chain."*

This can also read:

> — *"An acetic acid enters the lysyl side chain, loosing an hydroxyl group while a hydrogen atom leaves the lysyl side chain; water is lost in the process."*

.

The representation of that reaction is:

$$R{-}NH_2 + CH_3COOH \leftrightharpoons R{-}NH{-}CO{-}CH_3 + H_2O$$

When the user wants to define that chemical reaction, they can use that representation: "$-H_2O+CH_3COOH$", or even the more brief but chemically equivalent one: "$-H+CH_3CO$".



A formula can represent the mechanism of a reaction

In *MassXpert*, the chemical reaction representation is considered a valid formula and is called an *actionformula*

There is another interesting feature of formulæ in *MassXpert*: the title of a formula. Anywhere a formula might be entered, it can be entered with a title in the following form: "Acetylation"$-H_2O+CH_3COOH$. This is of particular interest in cases where the user wants to store formulæ for later use when that is offered by the *MassXpert* software.

2.2 The *MassXpert* Framework Data Format

All the data dealt with in *MassXpert* are stored on disk as *XML*-formatted files. *XML* is the *eXtensible Markup Language*.

This "language" allows to describe the structure of a document. The structure of the data is first described in a section of the document that is called the *Document Type Definition, DTD* the data then follow in the same file. One of the big advantages of using an *XML* format in *MassXpert* is that it is a text format, and not a binary one. This means that any data in the *MassXpert* package is human-readable (even if the *XML* syntax makes it a bit difficult to read data, it is actually possible). Try to read one of the polymer chemistry definition*XML*-formatted files that are shipped with this software package, and you'll see that these files are pure text files (the same applies for the `*.mxp` *XML*-formatted polymer

sequence files). The advantages of using text file formats, with respect to binary file formats are:

- The data in the files are readable even without the program that created them. Data extraction is possible, even if it costs work;
- Whenever a text document gets corrupted, it remains possible to extract some valid data bits from its uncorrupted parts. With a binary format, data are chained from bit to bit; loosing one bit leads to automatic corruption of all the remaining bits in the file;
- Text data files are searchable with standard console tools (*sed*, *grep*...), which make it possible to search easily text patterns in any text file or thousands of these files in one single command line. This is not possible with binary format, simply because reading them requires the program that knows how to decode the data and the powerful console-based tools would prove useless.

2.3 Chemical Entity Naming Policy

Unless otherwise specified, the user is *strongly* advised *not* to insert any non-alphanumeric-non-ASCII characters (space, %, #, $...) in the strings that identify polymer chemistry definition entities. This means that, for example, users must refrain from using non-alphanumeric-non-ASCII characters for the atom names and symbols, the names, the codes or the formulæ of the monomers or of the modifications, or of the cleavage agent definitions, or of the fragmentation pathway definitionss... Usually, the accepted delimiting characters are - and _. It is important not to cripple these polymer data for two main reasons:

- So that the program performs smoothly (some file-parsing processes rely on specific characters (like # or %, for example) to isolate sub-strings from larger strings);
- So that the results can be easily and clearly displayed when time comes to print all the data.

2.4 Installation with an AppImage software bundle

The *AppImage* software bundle format is a format that allows one to easily run a software program on any GNU/Linux-based distribution. From the appimage.org:

> The key idea of the AppImage format is *"one app = one file"*. Every AppImage contains an app and all the files the app needs to run. In other words, each AppImage has no dependencies other than what is included in the targeted base operating system(s).
>
> — Simon Peter

There are AppImage software bundles for the various *MassXpert3* versions that are available for download. As of writing, the software bundle has been tested on *Fedora* version 39 (release in 2023). This is thus a pretty old distribution and thus *MassXpert3* should also run on more recent versions of these computing platforms. The AppImage bundle of *MassXpert3* was created on a rather current Debian version: the *stable Debian* 13 distribution.

In order to run the *MassXpert3* software AppImage bundle, download the latest version (like `MassXpert3-\<version\>-x86_64.AppImage`. Once the file has been downloaded to the desired directory, change to that directory and change the permissions to make it executable:

```
chmod a+x MassXpert3<version>-x86_64.AppImage RETURN
```

Finally, execute the file that has become a normal program:

```
./MassXpert3-<version>-x86_64.AppImage RETURN
```



Tip

If the program complains about a locale not being found, please, modify the command line to read:

```
LC_ALL=C./MassXpert3-<version>-x86_64.AppImage
RETURN
```

3 Configuring the application preferences

A large application like *MassXpert* has lots of configurable options. The Application preferences window described in this chapter is the central place for these customizations. To enter the application preferences window, select **Preferences⇨Edit preferences**.

3.1 Setting the application preferences - general options

As shown in Figure 2 – Application preferences - General page the general options encompass the following notions:

- **Maximum number of threads:** Set the maximum number of threads when performing resource intensive tasks, like mass spectrum integration.
- **Decimal places for number representation:** Set the numer of decimal digits when representing numbers. Note that not all numbers need the same decimal precision.

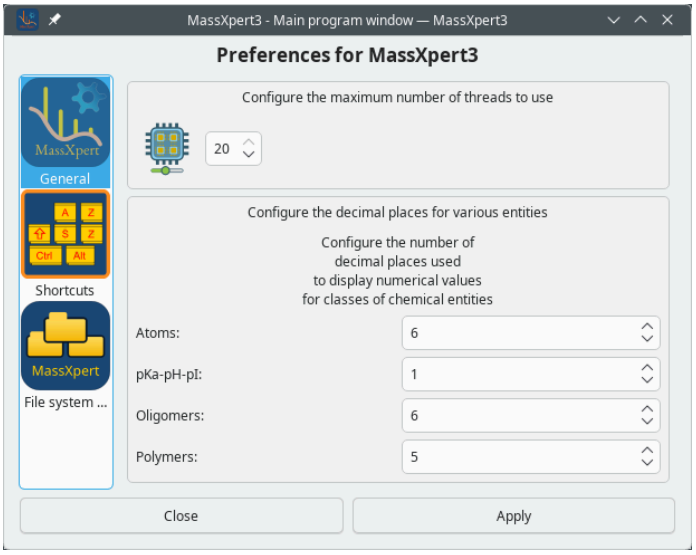


Figure 2: Application preferences - General page

The General page configures the maximum number of threads to use and the number of decimals to be used for a number of numerical entities.

3.2 Setting the application preferences - shortcuts

As shown in Figure 3 – Application preferences - Shortcuts page it is possible to configure the shortcuts for a large number of tasks throughout the application.

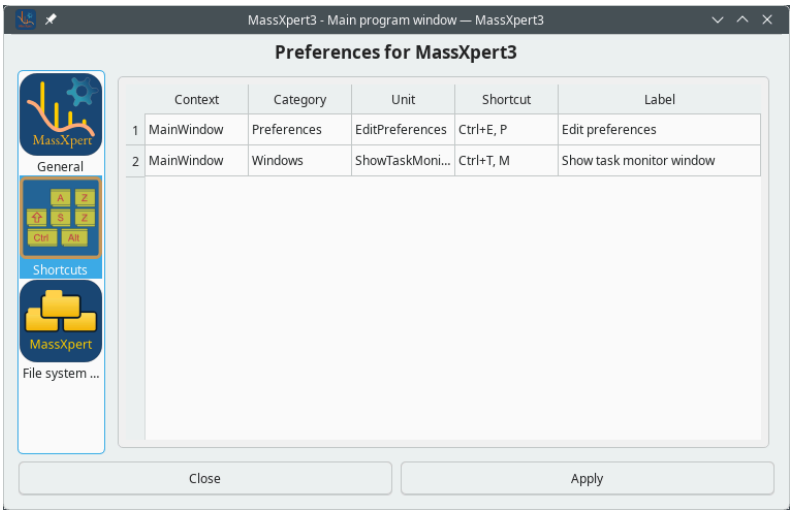


Figure 3: Application preferences - Shortcuts page

The shortcuts are classified according to Context, Category and Unit.

The shortcuts can be configured by double-clicking the item of interest in the cell of the Shortcut column and then entering the desired shortcut. Then simply click somewhere else.

When all the shortcuts have been configured, click the button labeled Apply and restart the program.

3.3 Setting the application preferences - filesystem hierarchy structure

The *MassXpert* program comes with a number of predefined polymer chemistry definitions and polymer sequence files. The locations of these data elements are set during the installation of the software by the system's package manager. It is however possible to configure any directory of the filesystem hierarchy so that the program continues to run even with changed (but documented) locations. In particular it is possible to configure the location of the user's data directories.

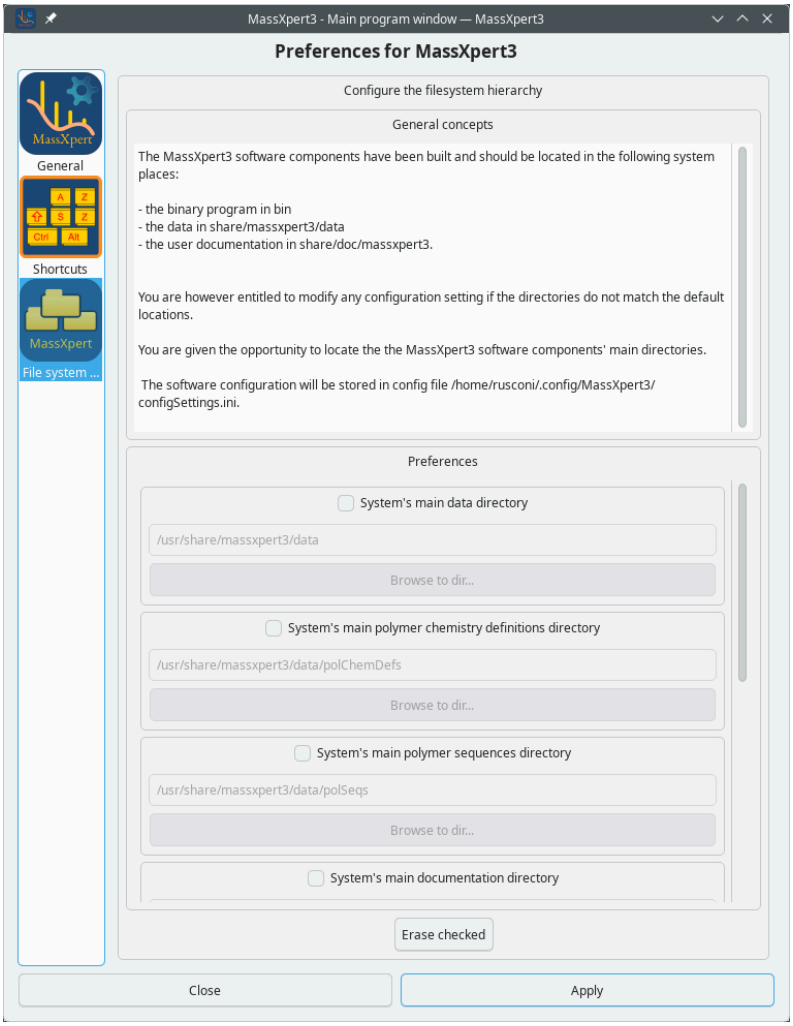


Figure 4: Application preferences - Filesystem hierarchy page

The location of any directory in the filesystem hierarchy can be modified but has to be documented in this window.

4 Basics in Polymer Chemistry

This chapter will introduce the basics of polymer chemistry. The way this topic is going to be covered is admittedly biased towards mass spectrometry and biological polymers. Moreover, the aim of this chapter is to provide the reader with the specialized words that will later be used to describe and explain the (inner) workings of the *MassXpert* program. This manual is not a "crash course" in biochemistry.

4.1 Polymers? Where? Everywhere!

Indeed, polymers are everywhere. If you ask somebody to show you something polymeric, he/she will point you at the first plastic object in the vicinity. Right, plastic materials are made of hydrocarbon polymers. We also have many different polymers in our body. Proteins are polymers, complex sugars are polymers, DNA (the so-called "molecule of heredity" is a *huge* polymer. There are polymers in wine, in wood... Where? Everywhere!

My *Oxford Advanced Learner's Dictionary of Current English* gives for *polymer* the following definition: —*"natural or artificial compound made up of large molecules which are themselves made from combinations of small simple molecules"*.

A polymer is indeed made by covalently linking small simple molecules together. These small simple molecules are called *monomer*s, and it is immediate that a *polymer* is made of a number of monomers. A general term to describe the process that leads to the formation of a polymer is *polymerization*. It should be noted that there are many ways to polymerize monomers together. For example, a polymer might be either linear or branched. A polymer is linear if the monomers that are polymerized can be joined at most two times. The first junction links an *"entering"* monomer to an *"elongating"* polymer (thus making it the new end of the elongating polymer which, by the way, is longer than before by one unit) and the second junction links the new elongating polymer's end to a new monomer. This process goes on until the reaction is stopped, the point at which the polymer reaches its *finished state*. A branched polymer is a polymer in which at least one monomer is able to contract more than two bonds. It is thus clear that a single monomer linked three times to other monomers will yield a "T-structure", which is nothing but a branched structure.

In the following sections we'll describe a number of different kinds of polymers. Each time, they will be described by initially detailing the structure of their constitutive monomers; next the formation of the polymer is described. At each step we shall try to set forth each polymer characteristics in such a manner as to introduce the way *MassXpert* "thinks polymers" and to introduce specialized terminologies. Once the basic chemistries (of the different polymers) have all been described, we will enter a more complex subject that is

of enormous importance to the mass spectrometry specialist: polymer chain disrupting chemistry. We shall see that this terminology actually involves two kinds of chemistries: cleavage, on the one hand, and fragmentation, on the other hand.

While *MassXpert* is basically oriented to linear single-stranded polymer chemistries, it can also be used to simulate highly complex polymer chemistries. Biological polymers are the main focus of this manual, however all the concepts described here may be applied with no modification to synthetic polymer chemistries.

4.2 Various Biopolymer Structures

Biopolymers are amongst the most sophisticated and complex polymers on earth and it certainly is not a mistake to take them as examples of how monomers (be these complex or not) can assemble covalently into life-enabling polymers. In this section we will visit three different polymers encountered in the living world: proteins, nucleic acids and polysaccharides. We shall be concerned with 1) the structure of the monomers, 2) the polymerization reaction and 3) the final end-capping reaction responsible for putting the polymer in its finished state.

4.2.1 Proteins

These biopolymers are made of amino acids. There are twenty major amino acids in nature, and each protein is made of a number of these amino acids. The combinations are infinite, providing enormous diversity of proteins to the living world.

A protein is a polar polymer: it has a left end and a right end, and polymerization actually occurs from left to right (from N-terminus to C-terminus, see below).

Figure 5: Peptidic bond formation by condensation

shows that the chemical reaction at the basis of protein synthesis is a *condensation*. A protein is the result of the condensation of amino acids with each other in an orderly polar fashion. A protein has a left end, called *N-terminus* (or *amino-terminal end*) and a right end, called *C-terminus* or *carboxyl-terminal end*. The left end is an amino group (H_2N–) corresponding to the non-reacted amino group of the amino acid. Upon condensation of a new amino acid onto the first one, the carboxyl group of the first amino acid reacts with the amino group of the second amino acid. A water molecule is released, and the formation of an amide bond between the two amino acids yields a dipeptide. The right end of the dipeptide is a carboxyl group (—COOH) corresponding to the unreacted carboxyl group of the last amino acid to have "polymerized in".

The bond formed by condensation of two amino acids is an amide bond, also called—in protein chemistry— a *peptidic bond*. The elongation of the protein is a simple repetition of the condensation reaction shown in

Figure 5: Peptidic bond formation by condensation

, granted that the elongation *always* proceeds in the described direction (a new monomer arrives to the right end of the elongating polymer, and elongation is done from left to right).

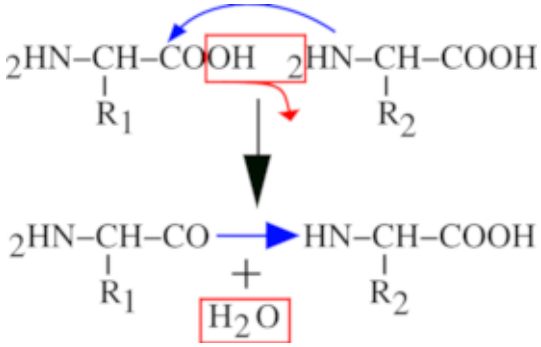


Figure 5: Peptidic bond formation by condensation

The left end monomer R_1 is condensed to the right end monomer R_2 to yield a peptidic bond. A water molecule is lost during the process.

Now we should point at a protein chemistry-specific terminology issue: we have seen that a protein is a polymer made of a number of monomers, called amino acids. In protein chemistry, there is a subtlety: once a monomer is polymerized into a protein it is no more called a monomer, it is called a *residue*. We may say that a residue is an amino acid less a water molecule.

From what we have seen until now, we may define a protein this way:

> — *"A protein is a chain of residues linked together in an orderly polar fashion, with the residues being numbered starting from 1 and ending at n, from the first residue on the left end to the last one on the right end."*

This definition is still partly inexact, however. Indeed, from what is shown in

Figure 6: End capping chemistry of the protein polymer

, there is still a problem with the extremities of the residual chain: what about the amino group on the left end of a protein (the amino group sits right onto the first amino acid of the protein), and what about the carboxyl group of the right end of a protein (the carboxyl group sits right onto the last amino acid of the protein)? Because these groups lie at the extremities of the residual chain, they remained unreacted during the polymerization process. But because we are simulating a residual chain using residues and not amino-acids,

we still need to put the residual chain in its finished state: by *capping* the left end with a proton *cap* (+H, so as to complete the amino group) and the right end with a hydroxyl cap (+OH, so as to complete the carboxyl group). The capping of the residual chain extremities ensures that the polymer is in its finished state, and that it cannot be elongated anymore. The proton is the *left cap* of the protein polymer and the hydroxyl is the *right cap* of the protein polymer.

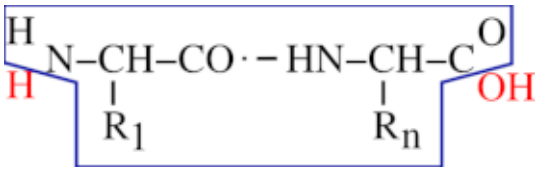


Figure 6: End capping chemistry of the protein polymer

A protein is made of a chain of residues and of two caps. The left cap is the N-terminal proton and the right cap is the C-terminal hydroxyl. Altogether, the residual chain (enclosed here in the blue polygon) and both the H and OH red-colored caps do form a complete protein polymer in its *finished* state.

Now comes the question of unambiguously defining the structure of a protein. It is commonly accepted that the simple ordered sequence of each residue code in the protein, from left to right, constitutes an unambiguous description of the protein's primary structure (that is, of its sequence). Of course, proteins have three-dimensional structures, but this is of no interest to a program like *MassXpert*, which is aimed at calculating masses of polymers. To enunciate unambiguously the sequence of a protein, one would use a symbology like this:

- Using the 3-letter code of the amino acids: Ala Gly Trp Tyr Glu Gly Lys
- Using the 1-letter code of the amino acids: A G W Y E G K

  Alanine is thus the residue 1 and Lysine is the last residue (n = 7)

4.2.2 Nucleic Acids

These biopolymers are more complex than proteins, mainly because they are composed of monomers (*nucleotides*) that have three different chemical parts, and because those parts differ in DNA and RNA. A nucleotide is the nucleic acid's brick: *a nucleotide consists of a nitrogenous base combined with a ribose/deoxyribose sugar and with a phosphate group*. There are two different kinds of nucleic acids: deoxyribonucleic acid (DNA, the sugar is a deoxyribose) and ribonucleic acid (RNA, the sugar is a ribose). DNA is most often found in its double stranded form, while RNA is most often found in single strand form. There are four nitrogenous bases for each: Adenine, Thymine, Guanine, Cytosine for DNA; in RNA only one of these bases changes: Thymine is replaced by Uracile. As for proteins, nucleic acids are polar polymers: the polymerization process is polar, from left to right

(sometimes left is up and right is down in certain vertical representations found mainly in textbooks).

This manual is not to teach biochemistry, which is why the structure of the monomers is not described in atomic detail. However, since it is important to understand how the polymerization occurs,

Figure 7: Phosphodiester bond formation by esterification

represents the polymerization reaction mechanism between a nucleotide and another one, to yield a dinucleotide. That reaction is a *trans-esterification*. A nucleic acid has a left end —*5′ end; often this end is phosphorylated*—and a right end—*3′ end; hydroxyl end*. The trans-esterification reaction is the attack of the phosphorus of the new (deoxy)nucleotide triphosphate by the 3′OH of the right end of the elongating nucleotidic chain. Upon trans-esterification, an *inorganic pyrophosphate* (PP_i) is released, and the formation of a phosphodiester bond between the two nucleotides yields a dinucleotide. The elongation of the nucleic acid polymer is a simple repetition of this esterification reaction so that the chain growth is always in the 5′→3′ direction. This is achieved in the living cells by what is called the *5′→3′ polymerase enzymatic activity*.

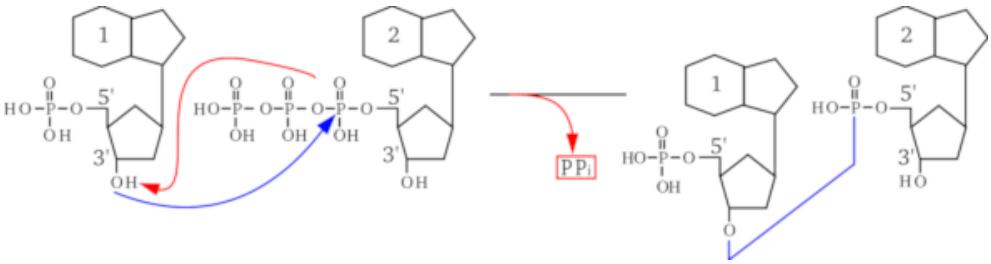


Figure 7: Phosphodiester bond formation by esterification

The “entering” monomer (on the right) has its triphosphate on the 5′ carbon of the sugar esterified by nucleophilic attack of the first phosphorus by the alcohol function beared by the 3′ carbon of the (deoxy)ribose sugar ring of the left monomer. The bond that is formed is a phosphodiester bond, with release of a pyrophosphate group (PP_i). Note that the sugar and nitrogenous bases are schematically represented in this figure.

The conventional representation of a nucleic acid involves showing the 5′ end on the left, and the 3′ end on the right, horizontally. Sometimes, to clearly indicate that the left end is phosphorylated, while the right end is not, the ends are indicated as “5′P” and “3′OH”.

Figure 8: End capping chemistry of the nucleic acid polymer

shows a simple way to formalize what a nucleic acid polymer is. The molecule represented on the left is the “monomer” in the sense that the polymer is made of n monomers. On the right side of that figure, the polymer made of n monomers is shown as a residual chain (inside the blue polygon box) that got capped with OH on its left end and H on its right end (red-colored atoms). Thus, in the case of the nucleic acid polymers, the left cap is a

hydroxyl and the right cap is a proton. This anecdotically happens to be the exact converse of what was described earlier for proteins.

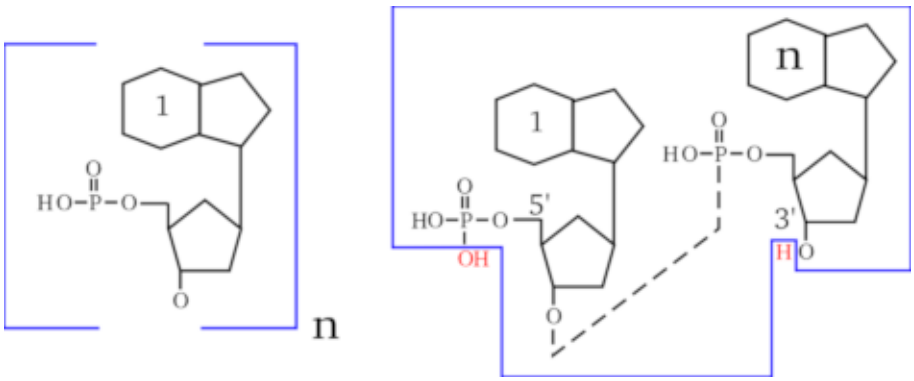


Figure 8: End capping chemistry of the nucleic acid polymer

A nucleic acid is made of a chain of nucleotides (left formula) and of two caps. The left cap is the hydroxyl group that belongs to the terminal phosphate of the 5′ carbon of the sugar. The right cap is the proton that belongs to the hydroxyl group of the 3′ carbon of the sugar ring (right formula). Altogether, a finished nucleic acid polymer is made of the nucleotidic chain (enclosed here in the blue polygon), made of the repetitive elements (one of which is shown on the left), and of the two caps (red-colored OH and H, outside of the polygonal box on the right).

Now comes the question of unambiguously defining the structure of a nucleic acid. It is commonly accepted that the listing of the named nitrogenous bases in the nucleic acid—from left (5′ end) to right (3′ end)—constitutes an unambiguous description of the nucleic acid sequence. To enunciate the sequence of a gene, one would use a symbology like this:

for a DNA, using the 1-letter code of the nitrogenous bases:

A T G C A G T C

for an RNA, using the 1-letter code of the nitrogenous bases:

A U G C A G U C

Adenine is thus the base 1 and Cytosine is the last base (n = 8)

4.2.3 Saccharides

These biopolymers are certainly amongst the most complex ones in the living world. This is mainly due to the fact that saccharides are usually heavily modified in living cells with a huge variety of chemical modifications. Furthermore, the ramifications in the polymer structure are more often the normal situation than not. Interestingly, these molecules are first thought of as the “fuel” for the cell, which is certainly far from being total nonsense, but it is also undoubtful that their structural role is extremely important (often in combination with proteinaceous material). Another interesting aspect of their ability to form complex structures is their use as “key” systems for identification processes: a number of

complex sugars are located on the cell walls and provide "recognition patterns" for the other cells to deal with...

Nonetheless, the general picture is not that complex, if the way monomers are polymerized together is the only concern (which is the case in this manual). As far as we are concerned, in fact, the polymerization mechanism is a simple condensation (much like what has been described for proteins), yielding a sugar bond. Indeed, some people use the same terminology: a monomeric sugar becomes a residue once polymerized in the saccharidic chain. There are two main different kinds of sugars: *pentoses* (in C_5) and *hexoses* (in C_6); it should be noted, however, that there is a variety of other common molecules, like *sialic acids*, *heptoses*...

Like already seen for proteins and nucleic acids, a saccharidic polymer is polar: it has a left end and a right end. The terminology regarding the ends of a saccharidic polymer is rather unexpected at first sight: the left end is said to be the *non-reducing end* while the right end is said to be the *reducing end*. Historically this was observed with monosaccharides (also called *monoses*), which reduced cupric (Cu^{2+}) ions, thus getting oxydized themselves on the carbonyl (when in the open ring aldehydic form).

Figure 9: Osidic bond formation by condensation

shows the polymerization reaction between a sugar and another one (2 glucose monomers, actually), to yield a maltose disaccharide. The polymerization mechanism is a simple condensation. The elongation of the saccharidic polymer is a simple repetition of this condensation reaction so that the chain growth is always in the same orientation, from the non-reducing end to the reducing end. The conventional representation of a polysaccharide involves showing the non-reducing end on the left, and the reducing end on the right, horizontally.

Figure 10: End capping chemistry of the polysaccharidic polymer

shows a simple way to formalize what a saccharidic polymer is. The top formula is the representation of the monomer. The bottom formula represents a polysaccharide, with the repetitive elements boxed (there are n monomers polymerized). The atoms shown in red (outside the boxed repetitive elements) are the saccharidic polymer caps. Thus, we see clearly that in the case of polysaccharides, the left cap is a proton and the right cap is a hydroxyl. This anecdotically happens to be identical to proteins and the exact converse of what we described previously for nucleic acids.

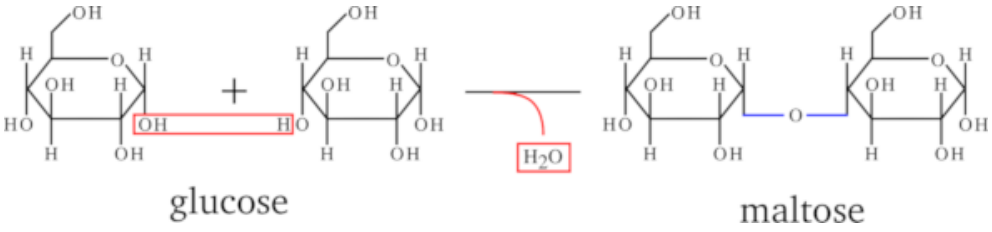


Figure 9: Osidic bond formation by condensation

The two monomers are subject to condensation with loss of one molecule of water.

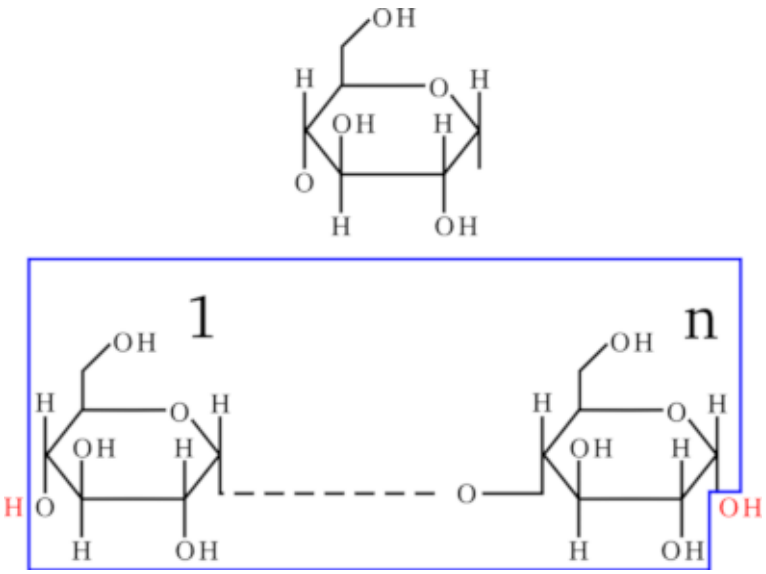


Figure 10: End capping chemistry of the polysaccharidic polymer

A polysaccharide is made of a chain of osidic residues (blue-boxed formula) and of two caps (red-colored atoms). The left cap is the proton group that belongs to the non-reducing end of the polymer. The right cap is the hydroxyl group that belongs to the reducing end of the polymer

Now comes the question of unambiguously defining the structure of a saccharidic polymer. It is commonly accepted that the simple ordered sequence of the named monoses in the saccharidic polymer, from left (non-reducing end) to right (reducing end), constitutes an unambiguous description of the glycan sequence. To enunciate the sequence of a glycan, one would use a symbology like this:

Using a 3-letter code:

Ara Gal Xyl Glc Hep Man Fru

Arabinose is thus the monose 1 and Fructose is the last monose (n = 7)

Incidentally, this is where the ability of *MassXpert* to handle monomer codes of non-limited length comes in handy!

4.3 To Sum Up

We made a rapid overview of the three major polymers in the living world. A great many other polymers exist around us.

Table 1: Quick comparison of three biopolymers with examples of monomers

tries to sum up all the informations gathered so far. Note that the formulæ given for the monomers are the "residual" ones. For example, the formula of the glycyl residue corresponds to the formula of the Glycine monomer less one molecule of water. Many synthetic polymers are much simpler than the ones we have rapidly reviewed, and it should be clear that, if *MassXpert* can deal with the complex biopolymers described so far, it certainly will be very proficient with less complex synthetic polymers. Describing the formation of polymers is one thing, but we also have to describe how to disrupt polymers. This is what we shall do in the next section.

polymer	name	code	formula	left cap	right cap
protein				H	OH
	Glycine	G	$C_2H_3O_1N_1$		
	Alanine	A	$C_3H_5O_1N_1$		
	Tyrosine	T	$C_9H_9O_2N_1$		
nucleic acid				OH	H
	Adenine	A	$C_{10}H_{12}O_5N_5P_1$		
	Cytosine	C	$C_9H_{12}O_6N_3P_1$		
saccharide				OH	H
	Arabinose	Ara	$C_5H_8O_4$		
	Heptose	Hep	$C_7H_{12}O_8$		

4.4 Polymer Chain Disrupting Chemistry

The "polymer chain disrupting chemistry" was mentioned earlier as a complex subject that was of *enormous* importance to the mass spectrometry scientist. This is why that subject will be treated in a pretty thorough manner. First of all it should be noted that a chemical modification of a polymer does not necessarily involve the perturbation of the chain structure of the polymer. Here, however, we are concerned specifically with a number of chemical modifications that yield a polymer chain perturbation; *cleavage* and *fragmentation*:

Cleavages These are chemical processes by which a cleaving agent will act directly on the polymer chain making it fall into at least two separated pieces (the *oligomers*). As a result

of the cleavage reaction, groups originating in the cleavage agent molecule may remain attached to the polymer at the precise cleavage location;

Fragmentations These are chemical processes by which the polymer structure is disrupted into separated pieces (the *fragments*) mainly because of energy-dependent electron doublet rearrangements leading to bond breakage.

4.4.1 Polymer Cleavage

We said above that, upon cleavage of a polymer, the cleaving molecule reacts with it, and by doing so directly or indirectly *"dissolves"* an inter-monomer bond. A polymer cleavage always occurs in such a way as to generate a set of *true* polymers (smaller in size than the parent polymer, evidently, which is why they are called *oligomers*). Indeed, let us take the example shown in

Figure 11: Protein cleavage by water and cyanogen bromide

, where a tripeptide (a very little protein, containing a methionyl residue at position 2) is submitted either to a water-mediated cleavage (hydrolysis, upper panel) or to a cyanogen bromide-mediated cleavage (lower panel). The two cases presented in this figure are similar in some respects and different in others:

- In the first case the molecule that is responsible for the cleavage is water, while in the second case it is cyanogen bromide;
- In both cases the bond that is cleaved is the inter-monomer bond (in protein chemistry this is a peptidic bond);
- In both cases the Oligomer 2 has the same structure;
- The structures of the Oligomer 1 species differ when produced using water or cyanogen bromide as the cleaving molecule.

The difference between hydrolysis and cyanogen bromide cleavage is in the generation of the Oligomer 1 species: the cyanogen bromide cleavage has a side effect of generating a homoserine as the right end monomer of Oligomer 1, while hydrolysis generates a genuine methionine monomer. This is because water reverses in a very symmetrical manner what polymerization did (hydrolysis is the converse of condensation), while cyanogen bromide did some chemical modification onto the generated Oligomer 1 species.

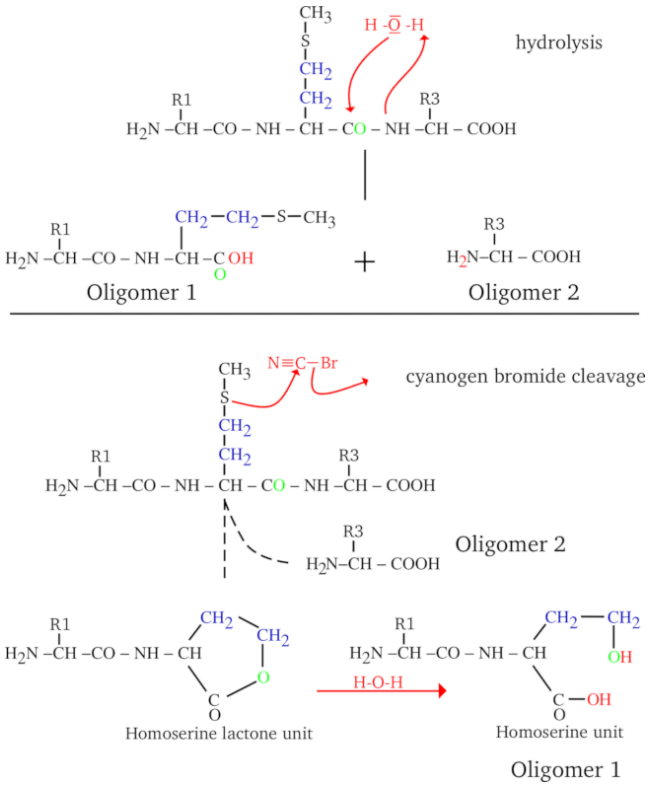


Figure 11: Protein cleavage by water and cyanogen bromide

A tripeptide is cleaved at position 1 either by hydrolysis (top) or by cyanogen bromide (bottom). Cyanogen bromide cleaves specifically on the right of a methionine monomer. Upon cleavage, the methionyl monomer gets converted into homoserine by the cyanogen bromide reagent.

Nonetheless, the reader might have noted that—interestingly—all the four oligomers do effectively have their left cap (a proton) and their right cap (the hydroxyl). This means that in both water- and cyanogen bromide-mediated cleavages, all the generated oligomers are indeed true polymers in the sense that: 1) they are a chain of monomers (modified or not) and 2) they are correctly capped (*i.e.* they are polymers in their finished state). This is important because it is the basis on which we shall make the difference between a cleavage process and a fragmentation process. Thus, the *MassXpert* definition of an oligomer can be formulate like so:

> — *"An oligomer is a polymer (of at least one monomer) in its finished state that was generated upon cleavage of a longer polymer."*

When the polymer cleavage reaction precisely reverses the reaction that was performed for the same polymer's synthesis, there is no special difficulty. But when the cleavage reaction modifies the substrate, then this should be carefully modelled. How? To answer this question we might start by comparing the two different Oligomer 1 species that were yielded upon the water-mediated and the cyanogen bromide-mediated cleavage reactions. The difference might be worded like so:

> — *"The hydrolysis-generated Oligomer 1 is equal to the cyanogen bromide-generated Oligomer 1 +S1 +C1 +H2 −O1."*

Turns out that this is a big difference! The observations we did so far might be worded this way:

> — *"Whenever a protein undergoes a cyanogen bromide-mediated cleavage, the "−C1H2S1+O1" chemical reaction should be applied to the resulting oligomers if and only if they have a methionine monomer at their right end."*

In *MassXpert*'s jargon, this logical condition is called a *cleavage rule* (described later; see Section 5.2.4 – The cleavage agent definitions.

Well, all this sounds reasonable. But what about the "normal" case, when the cleavage is done using water? Nothing special: the mass of the oligomer is calculated by summing the mass of each monomer in the oligomer (since the monomers are not modified, this is easily done) and the masses corresponding to the left and right caps (these are defined in the polymer chemistry definition; in our present case it would be a proton on the left end, and a hydroxyl on the right end). In this way, the oligomer complies with its definition, which states that it is a faithful polymer made of monomers and that it is in its finished state.

Yes, but then how will *MassXpert* manage to calculate the mass of the modified oligomer, like our Oligomer 1 in the case of the cyanogen bromide-mediated cleavage? Simple enough: in a first step it does exactly the same way as for the unmodified oligomer. Next, each oligomer is checked for presence or absence of a methionine residue on its right end. If a methionine is found, the mass corresponding to the "−C1H2S1+O1" chemical reaction is applied. And that's it.

> **Important**
>
> There is one situation where the cleavage rule formula is not applied to an oligomer that has a "M" residue at its right end: if the oligomer was not generated by cleaving at that residue ; that is, if the oligomer corresponds to the C-terminus of the cleaved protein and if the C-terminal residue is "M".

In the previous cyanogen bromide example, the logical condition was involving the identity of the oligomers' right end monomer, but other examples can involve not the right end monomer, but the left end monomer, if some chemical modification was to occur to the monomer sitting right of the cleavage location. In this case the user would have to analyse the situation and provide *MassXpert* with the proper chemical reaction by stating something analog to: "*If and only if they have a Xyz monomer at their left end*". This

introduction to polymer cleavage abstraction should be enough to later delve into the cleavage specification definition as *MassXpert* conceives it and that is thoroughly detailed at Section 5.2.4 – The cleavage agent definitions.

4.4.2 Polymer Fragmentation

In a fragmentation process, the bond that is broken is not necessarily the inter-monomer bond. Indeed, fragmentations are oft-times high energy chemical processes that can affect bonds that belong to the monomers' internal structure. This is one of the reasons why fragmentations do differ from cleavages: they are specific of the polymer type in which they occur. Hydrolyzing a protein and an oligosaccharide is just the same process, from a chemical point of view. But fragmenting a protein or an oligosaccharide are truly different processes because the way that the fragmentation happens in the polymer sequence is so much dependent on the nature of each monomer that makes it.

Another peculiarity of the fragmentations, compared with the cleavages that were described above, is the fact that there is no cleaving molecule starting the process. Instead, a fragmentation process is often initiated by an intra molecular electron doublet rearragement that propagates more or less in the polymer structure to eventually break it. Fragmentations are mainly a gas phase process, not some reaction that happens in solution as a result of putting in contact the polymer and some reagent. It is precisely because no cleaving molecule is involved in the fragmentation process that the fragments are not necessarily capped like a normal polymer should be; and this is another really important difference between cleavage and fragmentation. The following examples should illustrate these concepts using protein and nucleic acid fragmentation pathways.

4.4.2.1 Protein Fragmentation

There is a pretty important number of different kinds of fragments that can be generated upon fragmentation of peptides. We are going to detail the most common ones; the user is invited to use the *MassXpert* fragmentation-specification grammar to add less frequent (or newly discovered) fragmentation types. Note that the fragmentation schemes below apply to positively-charged precursor ions. To compute the product ions' masses obtained in negative mode fragmentation experiments, then, simply remove as many protons as required. For example, to switch from a fragment positively charged once (+H), then remove a first proton to go back to the uncharged state and then remove another proton to yield the deprotonated (thus singly negatively charged) ion product. The requirement to be able to computed masses for the positively- and negatively-charged ion products imposes a specific way to define fragmentation specifications in the *XpertDef* module (to be detailed later in this manual, see Section 5.2.5 – The fragmentation pathway definitions.

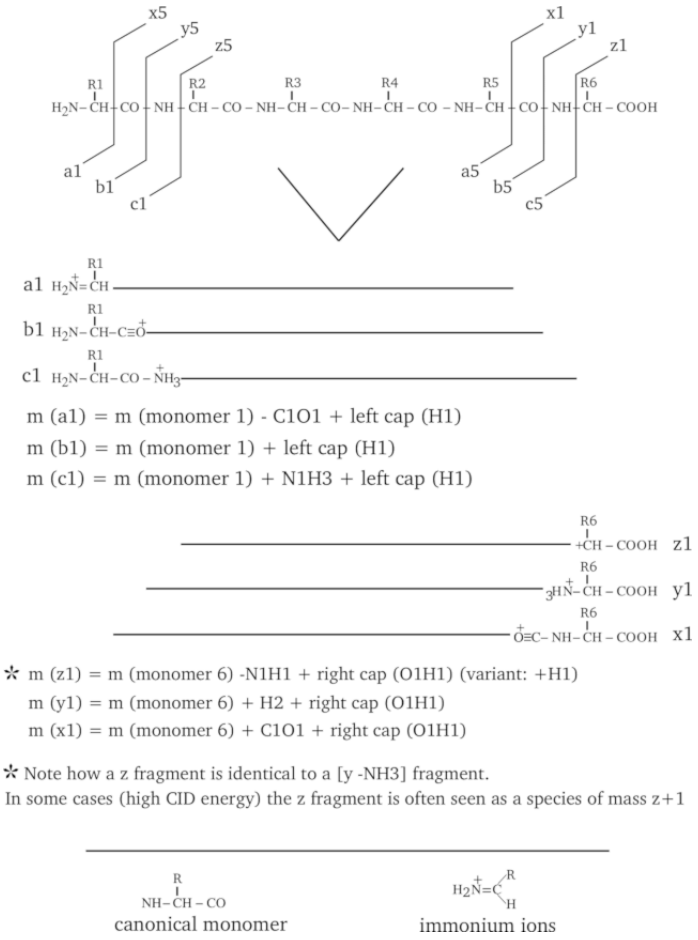


Figure 12: Protein fragmentation patterns most widely encountered

An hexapeptide is fragmented in the seven most widely encountered manners, such as to generate a, b, c, x, y, z and immonium fragment ions. The figure illustrates the position of the cleavage for each kind of fragment (exemplified using the case of the smallest fragment possible) and the mass calculation method is described for each fragment kind; consider that each fragment bears only *one positive* charge.

As can be seen from

Figure 12: Protein fragmentation patterns most widely encountered

, the fragmentations do generate fragments of three categories: the ones that include the left end of the precursor polymer (a, b, c), the ones that include the right end of the precursor polymer (x, y, z), and finally the special case in which the fragment is an *internal fragment*, like the immonium ions. When looking at the fragmentations described in the figure it becomes immediately clear why a fragmentation cannot be mistaken for a cleavage: the ionization of the fragment is not necessarily due to the captation of a proton by the fragment. Furthermore, we can also see that a fragmentation is not a cleavage because the fragment that is generated is *absolutely* not necessarily what we call a polymer, in the sense that the fragment might not be capped the same way as the precursor polymer is (that is, the fragment is not in its finished polymerizaton state).

The two observations above should make clear to the reader that calculating masses for fragments is a more difficult process than what was described earlier for the cleavate-generated oligomers. Indeed, while it was simple to calculate the mass of an oligomer (by simply adding the masses of its constitutive monomer units, plus the left and right caps, plus ionization), here there is no chemical formalism generally applicable to all the fragment types. This is why the specification of the fragmentation is left to the user's responsibility.

By looking at

Figure 12: Protein fragmentation patterns most widely encountered

, the reader should have noticed that the fragment naming scheme takes into consideration the fact that the fragment bears the left or the right end of the precursor polymer (or none, also). Indeed, the numbering of fragments holding the left end of the precursor polymer sequence begins at the left end, and for fragments that hold the right end, at the right end. Thus the third fragment of series *a* (*a3*) would involve monomers [1→] and the third fragment of series *y* (*y3*) would involve monomers [6→] (in the figure, these left-to-right and right-to-left directions are symbolized using arrows). Therefore, it should appear to the reader how important—when specifying a fragmentation—it is to clearly indicate from which end of the precursor polymer the fragment is generated (in *MassXpert*'s jargon this is "LE" for left end, "RE" for right end and "NE" for no end). *MassXpert* knows what action it should take when it encounters one of these three specifications; for example, if a "LE" specification is found for a given fragmentation specification, *MassXpert* adds to the fragment's mass the mass corresponding to the left cap of the precursor polymer.

The mass calculations for the different fragment product ion types is shown in

Figure 13: Peptide product ion mass calculation for the different ion series

$$a_i = LC + \sum_1^i M_i - CO$$

$$b_i = LC + \sum_1^i M_i$$

$$c_i = LC + \sum_1^i M_i + NH_3$$

$$x_i = RC + \sum_1^i M_i + CO$$

$$y_i = RC + \sum_1^i M_i + H_2$$

$$z_i = RC + \sum_1^i M_i - NH$$

$$z_i = RC + \sum_1^i M_i - NH + H \text{ (HCD)}$$

$$imm_i = M_i + H - CO$$

Figure 13: Peptide product ion mass calculation for the different ion series

The way masses are computed from peptide fragments change according to the fragmention pattern that produces product ions of different series. *LC* and *RC* mean *left cap* and *right cap*, respectively.

If we take the *a* fragment series,

Figure 12: Protein fragmentation patterns most widely encountered

indicates that the fragments include the left end and that their last monomer lacks its carbonyl group (see at the top of the figure that the *a1* arrow goes between the CαH and the CO of monomer 1). So we would say that each fragment of the *a* series should be accounted for with the following chemical treatments: 1) addition of the mass corresponding to the left cap (proton), 2) removal of the mass corresponding to the lacking CO group. This way we have the mass of fragment *a1*. If we were interested in the fragment *a4* we would have summed the masses of monomers 1 to 4, added the mass of the left cap, and finally removed the mass of a CO.

For the *x* series of fragments we do not add the left cap anymore, but replace it with the right cap, since the fragments hold the right end of the precursor polymer. Note also that the numbering of the monomers using the variable *i* in the following mathematical expressions goes from right to left (contrary to what happened for the *a, b, c* fragment series. All the fragments that hold the precursor polymer right end are numbered this way, so this applies to fragments *x, y, z*.

In low energy CID, the *z* fragments are expressed in a distinct way, when compared to high energy CID.

In the *immonium* fragment series, the fragments are internal fragments in the sense that they do not hold neither of the two precursor polymer's ends. *MassXpert* understands that the user is speaking of this kind of fragment when the "from which end" piece of data in the fragmentation specification states "NE" instead of "LE" or "RE" (see Section 4.4.2 – Polymer Fragmentation. The mass calculation for these fragments does not take into account the monomers surrounding the one for which the calculation is done. The mass for an immonium ion—at position *i* in the precursor polymer—will be the mass of the monomer at position *i*, less the mass of a CO, plus the mass of a proton.

4.4.2.2 Nucleic Acids Fragmentation

The fragmentations that can be obtained with nucleic acids are numerous and it is more complicated than with proteins to describe them fully. The main reason for this is that there are a big number of fragmentation combinations because of the loss of nitrogenous bases from the skeleton. The mechanisms by which this loss happens are fairly complex, and I am not going to detail any of them.

Figure 14: DNA fragmentation patterns most widely encountered

shows the most common fragmentations (without taking into consideration the potential loss of bases). An example of fragment is given for each fragment series (pretty the same way as we did before for proteins). Note that the fragment representations are aimed at helping the reader to figure out what the product ion is, not taking into account where

the negative charge lies on the fragment, since this charge can float around at every deprotonatable group. All the fragments shown bear one and one only charge.

Another remark pertaining to the ionization mode of the ion products: the fragmentation schemes below apply to negatively-charged precursor ions (by loss of a proton, typically). To compute the product ions' masses obtained in positive mode fragmentation experiments, then, simply add as many protons (or any other cationic ionization agent) as required. For example, to switch from a fragment negatively charged once (-H), then add a first proton to go back to the uncharged state and then add another proton to yield the monoprotonated (thus singly positively charged) ion product. The requirement to be able to computed masses for the positively- and negatively-charged ion products imposes a specific way to defined fragmentation specifications in the *XpertDef* module (to be detailed later in this manual; see Section 5.2.5 – The fragmentation pathway definitions).

The reader might have noticed at the bottom of

Figure 14: DNA fragmentation patterns most widely encountered

that a provision is made in the case the fragmented molecular species are not 5′ end-phosphorylated but 5′ end-hydroxylated. Indeed, the canonical monomer is such that, upon polymerization and left capping, the 5′ end is phosphorylated. However, oft-times the oligonucleotides are synthesized chemically without the 5′ end phosphate group, thus ending in hydroxyl. This special case should be accounted for by applying to all the fragments that bear the left end of the precursor polymer the following chemical reaction: "$-HPO_3$". This chemical reaction should be applied *in addition* to the chemical reaction that yields the fragment *per se*.

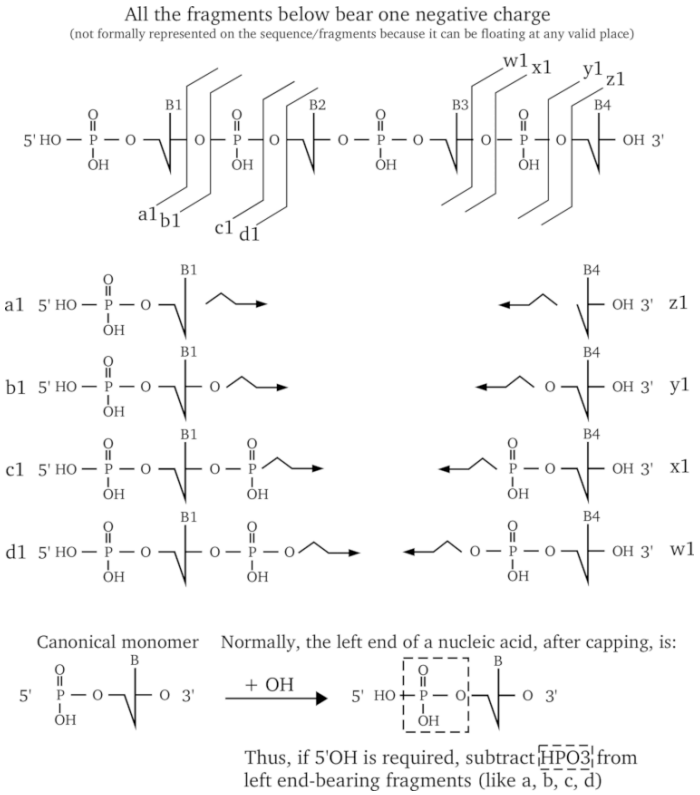


Figure 14: DNA fragmentation patterns most widely encountered

A short DNA sequence is fragmented in the eight most widely encountered manners, such as to generate a, b, c, d, w, x, y, z fragment ions. The figure illustrates the position of the cleavage for each kind of fragment (exemplified using the case of the smallest fragment possible). and the mass calculation method is described for each fragment kind; considering that each fragment is protonated only once (+1).

Exactly as done earlier for the protein fragments, the mathematical expressions used to calculate the mass of different series of nucleic acid fragments are provided in

Figure 15: Nucleic acids product ion mass calculation for the different ion series

In these calculations it is assumed that the left end of the precursor polymer is phosphorylated (5′P) and the reader should bear in mind that this precise phosphate might itself be expelled by the fragmentation. The fragment naming schemed detailed earlier for proteins applies to nucleic acids in the very same manner.

$$a_i = LC + \sum_1^i M_i - O$$

$$b_i = LC + \sum_1^i M_i$$

$$c_i = LC + \sum_1^i M_i - HPO_2$$

$$d_i = LC + \sum_1^i M_i - HPO_3$$

$$w_i = RC + \sum_1^i M_i + O$$

$$x_i = RC + \sum_1^i M_i$$

$$y_i = RC + \sum_1^i M_i - HPO_2$$

$$z_i = RC + \sum_1^i M_i - HPO_3$$

Figure 15: Nucleic acids product ion mass calculation for the different ion series

The way masses are computed from oligonucleotide fragments change according to the fragmention pattern that produces product ions of different series. *LC* and *RC* mean *left cap* and *right cap*, respectively.

In the *a* fragment series, the fragments most often appear with base loss. There are also a variety of fragments for which a base is lost.

4.4.2.3 More Complex Patterns Of Fragmentation

Before finishing with fragmentations, it is necessary to describe a powerful feature of the fragmentation specification grammar available in *MassXpert*. This feature was required for the fragmentation of oligosaccharides and also sometimes for proteins. When the fragmentation (the bond breakage reaction itself) occurs at the level of certain monomers, it might be necessary to be able to specify some particular chemistry that would arise on the monomer in question.

We have seen in the cleavage documentation that, upon cleavage of a protein sequence with cyanogen bromide, for example, a particular chemical reaction had to be applied to the oligomers that were generated with a methionine monomer as their right end monomer. Well, in a fragmentation specification it is possible to apply comparable chemical reactions but in a more thorough manner. Indeed, while in the cleavage it was possible to say something like: — *"Apply a given chemical reaction to the oligomer if the right end monomer is Xyz"*, in the fragmentation the logical condition can be bound not only to the identity of the currently fragmented monomer, but also (optionally) to the identity of the previous and/or next monomer in the precursor polymer sequence. For example:

> — *"Apply a given chemical reaction if fragmentation occurs at the level of 'Xyz' monomer only if it is preceded by a 'Yxz' monomer and followed by a 'Zyx' monomer."*

These logical conditions are called *fragmentation rules*. A *fragmentation specification* can hold as many rules as necessary. All of this is described in great detail at Section 5.2.5.2 – More complex fragmentation pathway patterns.

4.4.2.4 To Sum Up

To sum up all what we have seen so far with polymer chain disrupting chemistries:

- Fragments are never ionized automatically; ionization (gain/loss of a charged group) is necessarily integrated post-fragmentation as a configuration bit to mass calculations.
- A polymer sequence gets fragmented into fragments when a bond breakage occurs, without the help of any exterior molecule, at any level of the polymer structure, with no limitation to the inter-monomer bond; monomer-specific chemical reactions can be modelled into the fragmentation specification using any number of fragrules;
- Oligomers are automatically capped—*on both ends*—using the rules described in the precursor polymer's definition;
- Fragments are capped automatically only—*on the end they hold, if any*—using the rules described in the precursor polymer's definition;
- Oligomers are automatically ionized (if required by the user) using the rules described in the precursor polymer's definition;
- Fragments are never ionized automatically; ionization (gain/loss of a charged group) is necessarily integrated in the fragmentation specification.

5 XpertDef: Definition of Polymer Chemistries

After having completed this chapter the reader will be able to accomplish the very first steps needed to use *MassXpert*'s features at best: the normal workflow, indeed, is to first make a polymer chemistry definition, in order to be able to edit polymer sequences of that specific definition. The *XpertDef* module is made available in *MassXpert* by pulling down the **XpertDef** menu item from the program's menu. It is possible to start a new polymer chemistry definition from scratch, but it is certainly usually easier to first duplicate a polymer chemistry definition shipped with *MassXpert* and then open that copy and edit it. Please, refer to Section 10 – Data Customizations for an explanation of how this is safely done.

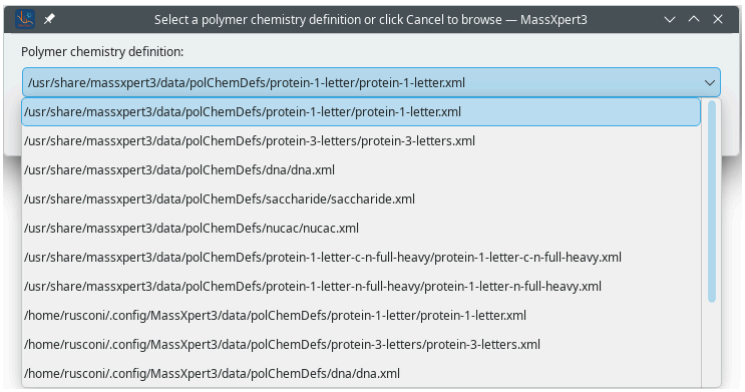


Figure 16: Select one polymer chemistry definition file

It is possible to immediately select a polymer chemistry definition already registered with the system, or open an arbitrary file by browsing the filesystem (click the Cancel button, hidden by the drop down list widget in this figure, if so desired).

To open a polymer chemistry definition, the user may either select one that is already registered with the system, and that appears listed in the drop-down list widget shown in

Figure 16: Select one polymer chemistry definition file

or click the Cancel button so as to open one definition file by browsing the filesystem. In the polymer chemistry definition window that shows up, the user accomplishes different tasks:

- Define the name of the polymer chemistry definition;
- Define "singular" data like the left cap and the right cap of the polymer, the ionization rule governing the default ionization of the polymer sequence;
- Define the isotopes needed to operate the different polymer chemistry entities (these are "plural" data);

- Define all the polymer chemistry entities needed to work on polymer sequences (all these are also “plural data”).

The definition of the isotopes and of all the chemical entities belonging to a given polymer chemistry are collectively called a *polymer chemistry definition*. The polymer chemistry definition window that shows up is shown in figure

Figure 17: Polymer chemistry definition window

.

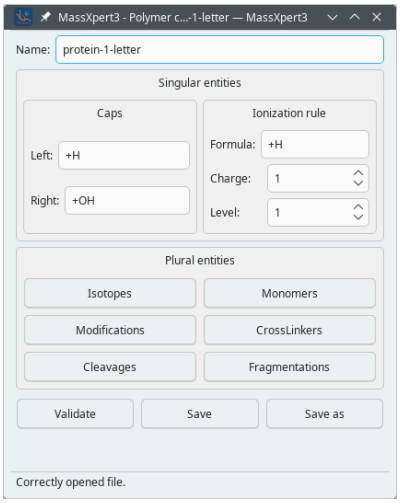


Figure 17: Polymer chemistry definition window

All the polymer chemistry entities are defined in this window. The different buttons dealing with isotopes, monomers, modifications, cross-linkers, cleavage and fragmentation specifications open up specific dialogs (see below).

5.1 The isotopes

The definition of the isotopes is performed through the user interface shown in figure

Figure 18: Isotopes definitions

(Isotopes button in the polymer chemistry definition window). In this dialog window, the user defines all the isotopes that collectively form the chemical elements (at least one isotope per atom, logically). In the table view, each row corresponds to an isotope.

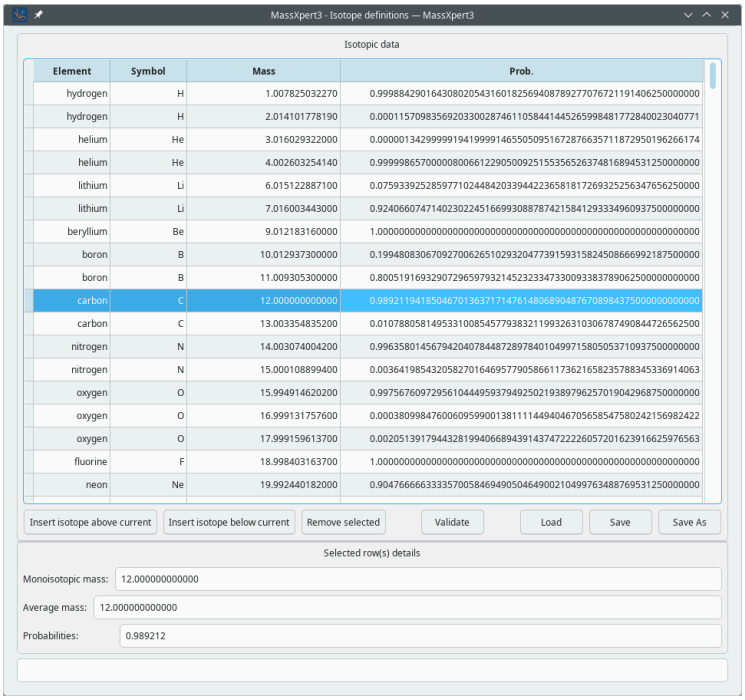


Figure 18: Isotopes definitions

Each row of the table view corresponds to an isotope. For any given element, there must be at least one isotope.

The table view is editable. Each cell, when double-clicked, switches to an edit mode in which the user might enter a value. A series of buttons below the table view trigger various actions. To add a new isotope, either above or below the currently selected one, click one of the two Insert isotope... buttons. To remove isotopes, select them and click the Remove selected button. The three buttons on the right hand side of the window, allow loading, saving and saving with a new file name the isotopes currently displayed in the table view.

When the user selects one or more rows of the table, the program updates some data specific to the selected isotopes. As long as the selected isotopes all have the same chemical element, they are taken into account for the calculation of the average mass (Average mass). The Probabilities line edit widget displays the sum of the abundances of the selected isotopes (only if these are of the same chemical element).

5.2 The polymer chemical entities

Once the isotopes have been properly defined (note that such isotopes are already available in the distributed package), it is possible to start entering data for the other polymer chemical entities (figure

Figure 17: Polymer chemistry definition window

). These are often defined using chemical formulas, which explain why it is necessary to first define the isotopes.

The following are the data that need to be entered so as to obtain a usable polymer chemistry definition:

- The polymer chemistry definition's name: protein-1-letter
- The chemical capping reactions that should happen on the left end and on the right end of the polymer sequence:
  - +H: left capping of the polymer sequence. Proteins are capped at the N-terminal end with a proton;
  - +OH: right capping of the polymer sequence. Proteins are capped at their C-terminal end with a hydroxyl group;
- The Ionization rule describes the manner in which the polymer sequence should be ionized by default, when the mass is calculated. This rule actually holds three elements:
  - Formula: *+H*: chemical reaction that ionizes the polymer sequence.In the example, all the polymer sequences of polymer chemistry definition *protein-1-letter* are protonated by default;
  - Charge: *1*: charge that is brought by the chemical agent ionizing the polymer (the formula above). In the example, a protonation reaction brings a single positive charge.
  - Level: *1*: ionization level, that is, the number of times that the ionization reaction above must be performed by default on any polymer sequence of this chemistry definition. In this example, monoprotonation is set as the default ionization rule.

At this point, time has come to deal with "plural" data. The first chemical entities to deal with are monomers.

5.2.1 The monomers

The monomers are the constitutive blocks of the polymer sequence. In the *MassXpert*'s jargon, "monomer" stands *not* for the molecule that may be used to perform a polymer synthesis; it stands for this molecule *less* the chemical group(s) that were eliminated upon polymerization. In the case of the biological polymers, the creation of chemical link between two monomers invariably leads to the loss of a water molecule (a condensation reaction, in organic chemistry).

Click onto the Monomers button, which triggers the opening of the dialog window shown in

Figure 19: Monomer definition

.

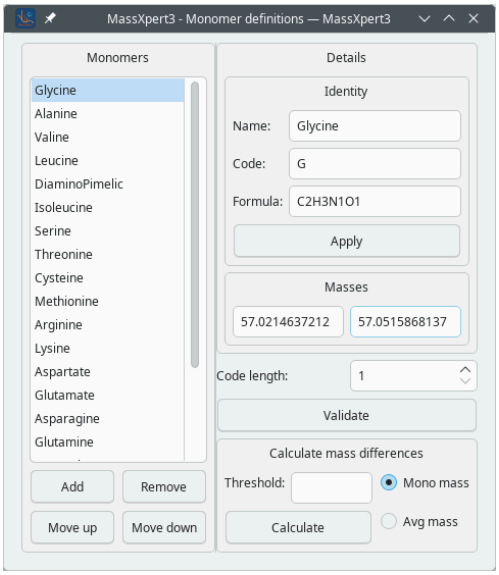


Figure 19: Monomer definition

Each monomer is defined with a name, a code and a chemical formula.

The way this dialog is operated is similar to what was described for the atom definition, unless it is simpler, because monomers are non-deep objects: there are no contained objects. One data element is critical: the number of characters that might be used to define the code of the monomer cannot be greater than the value entered in the `Code length` spinbox widget.[4]

> **Important**
>
> The first character of a monomer code must be uppercase, while the remaining characters (if any) must be lowercase. That means that—if `code length` is 3—"A", "Al", "Ala" would be perfectly fine, while "AlaN", "AL", "a", "aLA" would be wrong.

Each time a formula is either displayed by selecting a new monomer in the list or modified by editing it in its line edit widget, the monoisotopic and average masses are recalculated.

As of version 2.3.5, it is possible to calculate the mass difference between any two monomers in the definition. This is useful, for example, to grasp the resolution and mass accuracy requirements for a given polymer definition. The user sets a threshold to filter the results (in the example, that `Mono mass` threshold was set to *0.1*. The results of such a calculation are displayed in figure

[4] Allowing more than one letter to craft monomer codes might seem trivial at first. But that design decision triggered the requirement for non-trivial algorithms throughout all the code of the program. This is easily understandable at least in the polymer sequence editor: how are monomer codes keyed-in if "a" and "ala" are valid monomer codes in a polymer chemistry definition? The magic is described in the chapter about *XpertEdit* (Section 6.4 – Editing Polymer Sequences
).

Figure 20: Monomer mass differences

.

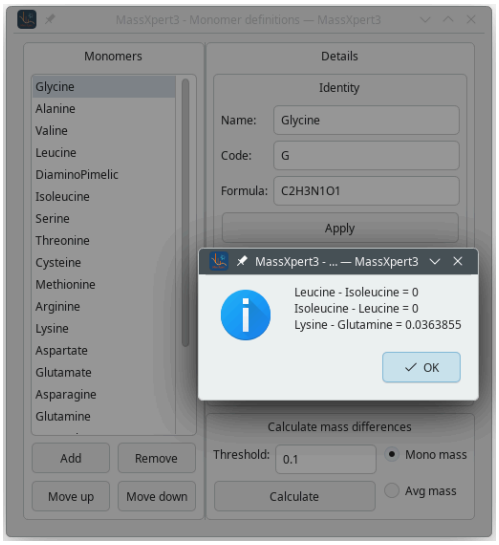


Figure 20: Monomer mass differences

The mass difference between any two monomers in the definition is computed and displayed only if it is less or equal to a threshold.

Once all the monomers have been added to the polymer chemistry definition, it is always a good idea to validate them by clicking Validate.

5.2.2 The modifications

Polymers are often either chemically or biochemically modified. In nature, biopolymers are modified more often than not. Some of the more common modifications in the protein realm are phosphorylation, acetylation and methylation, for example. Nucleic acids are modified with a sheer number of chemical modifications, saccharides also. The *MassXpert* software provides entire freedom to define any number of intelligent modifications, that is, modifications with any chemical formula but also that are knowledgeable of what monomers they can modify. Indeed, it would make no sense to phosphorylate a glycyl residue in a protein, for example.

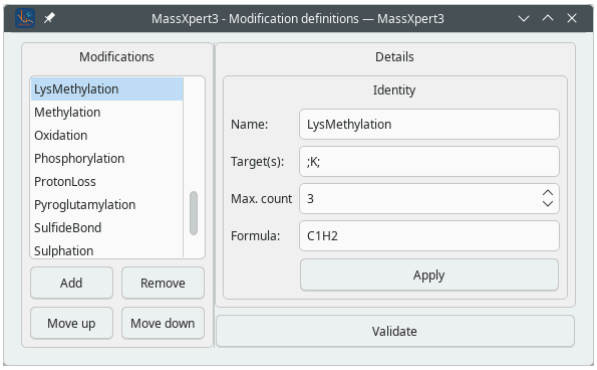


Figure 21: Modification definition

Each modification is defined with a Name, Target(s), a Max. count number and a chemical Formula.

Click the Modifications button to open the dialog window shown in figure

Figure 21: Modification definition

. In the example shown, the *LysMethylation* modification is being defined. A modification is defined by:

- Name: the name of the modification;
- Targets: a ';'-separated list of codes of the monomers that might be modified by this modification:;
- Max. count: the maximum number of times a given monomer might be modified with this modification. This feature is essential when working on methylation of proteins, for example, with arginyl and lysyl residues being multi-methylated;
- Formula: the formula that defines the modification chemical reaction, as explained in Section 2.1 – On Chemical Formulæ and Chemical Reactions.

Note that in the example of the figure, for *LysMethylation*, the *C1H2* formula is a net formula. That formula could be more explicit by entering "−H+CH3". The net formula is thus the one visible on the figure. The *LysMethylation* reaction can thus be read like this:

> — *"The polymer looses a proton and gains a methyl group."*

The *LysMethylation* is being defined as having K targets only, that means that when the user will try to modify non-lysyl monomers, the program will complain that these monomers are not targets of *LysMethylation*. There is, however, and for maximum flexibility, the possibility to override these target-limiting data when modifying monomers (See overriding the maximum target count limitation).

When a monomer is modified with this modification, its masses will change according to the net mass of this *LysMethylation* modification.

5.2.3 The cross-linkers

Polymers are often either chemically or biochemically modified by interconnecting monomers from the same polymer sequence. In the protein reign, one classical example of intra-sequence cross-linking is the formation of disulfide bonds. Another wonderful example is the formation of the fluorophore in the fluorescent proteins: there is a chemical reaction involving the side chains of three consecutive residues going on, resulting in the formation of a complex intra-sequence cross-link. Each side chain of the three monomers involved undergoes a chemical modification.

Cross-linkers are defined in the dialog window shown in

Figure 22: Cross-linker definition

. This dialog window is opened by clicking the CrossLinkers button.

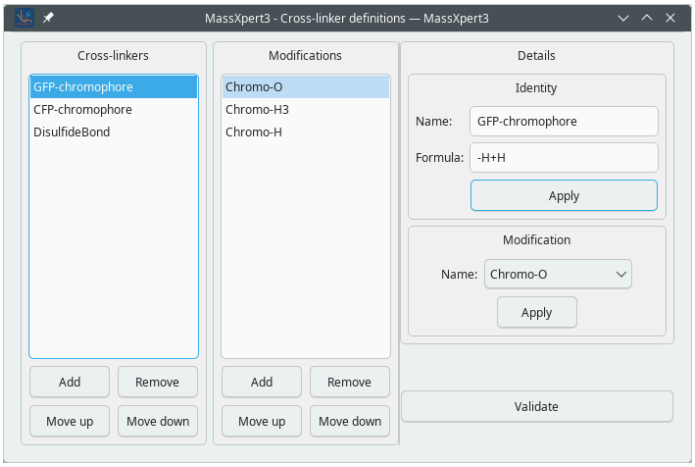


Figure 22: Cross-linker definition

Each cross-linker is defined using a Name, a Formula and either no or as many Modifications as there are monomers involved in the formation of the cross-link.

The formation of a cross-link between one or more monomers often involves chemical reactions to occur at the level of the engaged monomers. Cross-linkers defined in *MassXpert* should refer to these modifications as modification objects already available in the polymer chemistry definition. Note that, in some cases, it is not necessary to define modifications to occur at the level of the cross-linked monomers.

> **Note**
>
> When a cross-link does not involve any specific modification, as defined in the polymer chemistry definition, then a chemical formula must be entered in the Formula edit box widget, otherwise the cross-link definition will have no effect. In the figure example, the

> GFP-chromophore cross-linker has *-H+H*, which is nothing, as the value of the Formula field. This means that is there is no chemical reaction defined for the cross-linker *per se*.

The example described in figure

Figure 22: Cross-linker definition

, corresponds to the cross-linking reaction involved in the formation of the chromophore of the green fluorescent protein. That reaction involves the three following residues: threonyl–tyrosinyl–glycyl. Each monomer undergoes a distinct chemical modification: "–O", "–H_3" and "–H", respectively. Three modifications were thus defined: chromo-0, chromo-h3 and chromo-h, in that specific order, as these modifications are going to be sequentially applied to their corresponding residue in the cross-linking reaction.

> **Important**
>
> When multiple modifications are used, the number of these modifications must match the number of monomers involved, and their order must match the order with which the monomers are cross-linked. If no modification is defined, then, the chemical reaction that occurs upon cross-linking might be defined in the formula of the cross-linker.

5.2.4 The cleavage agent definitions

It is common practice—in biopolymer chemistry, at least—to cut a polymer into pieces using molecular scissors like the following:

- Proteases, for proteins;
- Nucleases, for nucleic acids;
- Glycosidases, for saccharides...

> **Note**
>
> Not only biological scissors can be defined, but also chemical ones, like cyanogen bromide, for example, that cleaves at a methionyl residue. *MassXpert* allows the user to define such kind of chemical scissor.

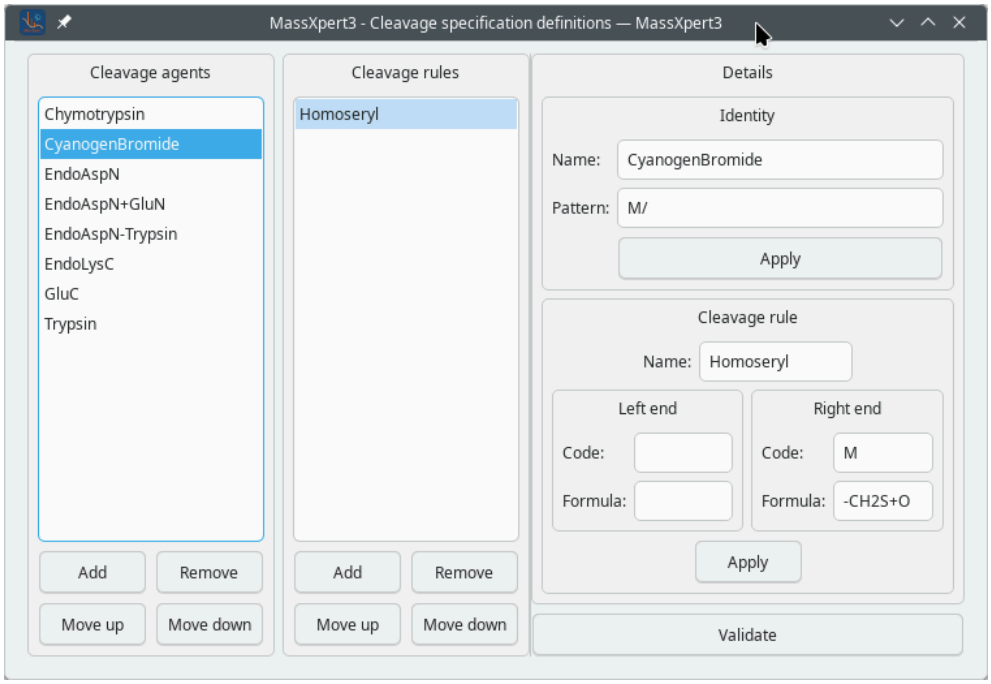


Figure 23: Cleavage agent definition

Each cleavage specification is defined using a name, a cleavage pattern and any number of cleavage rules.

For each different polymer type, the molecular scissors are specific. Indeed, a protease will not cleave a polysaccharide. This is why cleavage specifications belong to polymer chemistry definitions. In the example of

Figure 23: Cleavage agent definition

, the definition of the *CyanogenBromide* cleavage agent (this organic reagent cleaves right of methionyl residues) is detailed as so:

- Name: *CyanogenBromide*: the name of the cleavage agent;
- Pattern: *M/*: the sequence pattern recognized by the cleaving agent. In this case, the cleavage agent cleaves the protein right after *M* residues (mind that the '/' character is right of 'M');
- The Cleavage rules group box allows the user to define the cleavage rules that might be added to the cleavage agent. When a new cleavage rule is added, the corresponding data must be entered in the Cleavage rule group box, specifically with a Name and specifications for the chemical reactions that might occur at the left or right end of the oligomers resulting from the polymer cleavage (Left end and Right end). The chemical reactions are described by inserting the Code at which the reaction should occur and the Formula.

The notion of "Left end" or "Right end" are to be construed in the sense that when the oligomer is generated upon cleavage, the "Left end" rule applies to the left end of the

generated peptides and the “Right end” rule applies to the right end of the generated peptides.

The case of the cyanogen bromide reagent is interesting in this regard. The cleavage occurs right of methionyl residues. The generated peptides thus systematically end with a *M* methionyl residue. The cleavage reaction by cyanogen bromide involves a chemical modification of the methionyl residue at which the cleavage has occurred. This means that each generated peptide will have a modified methionyl residue on its right end. This is exactly what is described in

Figure 23: Cleavage agent definition

.

- the Left end chemical reaction data comprise no specific reaction data because once the reagent has cleaved right of a methionyl residue, the generated peptide has no modification on its left end.
- the Right end chemical reaction data comprise value *M* for Code and *-CH2S+O3* for Formula, respectively. Same explanation as above. This cleavage rule stipulates that upon cleavage of a protein using cyanogen bromide, the methionyl residue that gets effectively cleaved must be converted to a homoseryl residue. See below for a detailed explanation.

Here are some examples of more complex cleavage patterns:

- Pattern: *K/;R/;-K/P*, which reads like

> — *“Trypsin cuts right of a ‘K’ and right of a ‘R’. But it does not cut right of a ‘K’ if this ‘K’ is immediately followed by a ‘P’.”*

- Pattern: */D*, which reads like

> — *“EndoAspN cuts left of a ‘D’.”*

- Pattern: *Hypothetical = T/YS; PGT/HYT; /MNOP; -K/MNOP*:

> — *“Hypothetical cuts after ‘T’ if it is followed by ‘YS’ and also cuts after ‘T’ if preceded by ‘GP’ and followed by ‘HYT’. Also, Hypothetical cuts prior to ‘M’ if ‘M’ is followed by ‘NOP’ and if ‘M’ is not preceded by ‘K’.”*

.

> **Note**
>
> Please, *do* note that the letters in the examples above correspond to monomer codes and *not* to monomer names. If, for example, we were defining a "trypsin" cleavage specification pattern—in a protein polymer chemistry definition with the standard 3-character monomer codes—we would have defined it this way: "Trypsin = Lys/;Arg/;-Lys/Pro".

Now comes the time to explain in more detail what the Left end and Right end specifications mean, along with their Code and Formula fields, are for. For this, we shall consider that we have the following polymer sequence (1-character monomer codes): THISMWILLMBECUTMANDTHATMALSOM. If that sequence had been cleaved using cyanogen bromide and if the cleavage had been total,[5] that would have generated the following oligomers:

THISM WILLM BECUTM ANDTHATM ALSOM.

But if there had been partial cleavages, one or more of the following oligomers would have been generated:

THISMWILLM BECUTMANDTHATM ALSOM WILLMBECUTM ANDTHATMALSOM and so on...

Now, the biochemist knows that when a protein is cleaved with cyanogen bromide, the cleavage occurs effectively right of monomer "M" (this we knew already) *and* the "M" monomer that underwent the cleavage is changed from a methionyl residue to an homoseryl residue (this chemical change involves this formula: "−CH_2S+O"). Amongst all the oligomers generated above, there are two oligomers that should not undergo the "−CH_2S+O" cleavage rule: ALSOM and ANDTHATMALSOM. Indeed, these two oligomers were generated by the *CyanogenBromide* cleavage, but were not actually cleaved at the right side of a methionyl residue, because they correspond the the right end terminal part of the protein sequence.

> **Note**
>
> Imagine that the protein did have a "M" C-terminal residue: that residue is by essence never cleaved and as such would never undergo the methionyl-to-homoseryl reaction. This notion is of course taken into account by *MassXpert* when handling the C-terminal peptide of the protein.

[5] Cleavage occurs at every possible position, right of each "M" monomer.

This example should clarify why the definition clearly stipulates—in the cleavage specification for *CyanogenBromide*—that the oligomers resulting from this cleavage should undergo the "−CH2S+O" reaction only if they have a M as their right end monomer code. These *cleavage rules* need to be defined in a very careful way: imagine that—in some experiments involving cyanogen bromide —that reagent would cleave right of "C" (cysteine) residues, but with no chemical modification of the "C" monomer.[6] In this case, it would be suitable to put the flexibility of *MassXpert* at work by specifying that the generated oligomers should "undergo the "−CH2S+O" formula" only if they have a "M" as their right end monomer, so that "C"-terminated oligomers are not chemically modified. Thus the cleavage pattern might be safely defined: "M/;C/"...

5.2.5 The fragmentation pathway definitions

The specification of fragmentation pathways in a polymer chemistry definition is not a trivial task. In this section three different cases will be described, from simplest to most complex. One major rule is that the fragmentation pathway should be defined in such a manner that the resulting fragment is neutral. The ionization of the fragments will be then automatically performed by *MassXpert* upon calculation of the ion products according to the current ionization rule. This is a major improvement over previous versions, that forced the user to define fragmentation specifications by assuming a product ion of a given ionization status ($[M+H]^+$ for proteins or $[M-H]^-$ for nucleic acids, for example).

Will use the "dna" and "saccharide" polymer chemistry definitions to describe the notions involved in the definition of fragmentation pathways.

In the DNA chemical world, the fragmentation pathways involve the cleavage of bonds at the level of the phosphodiester link between nucleosides. There are also fragmentation pathways that also involve the decomposition of the nucleic base. This is shown in

Figure 24: Fragmentation definition in the DNA chemical world

.

[6] This is a purely hypothetical situation that I never observed personally.

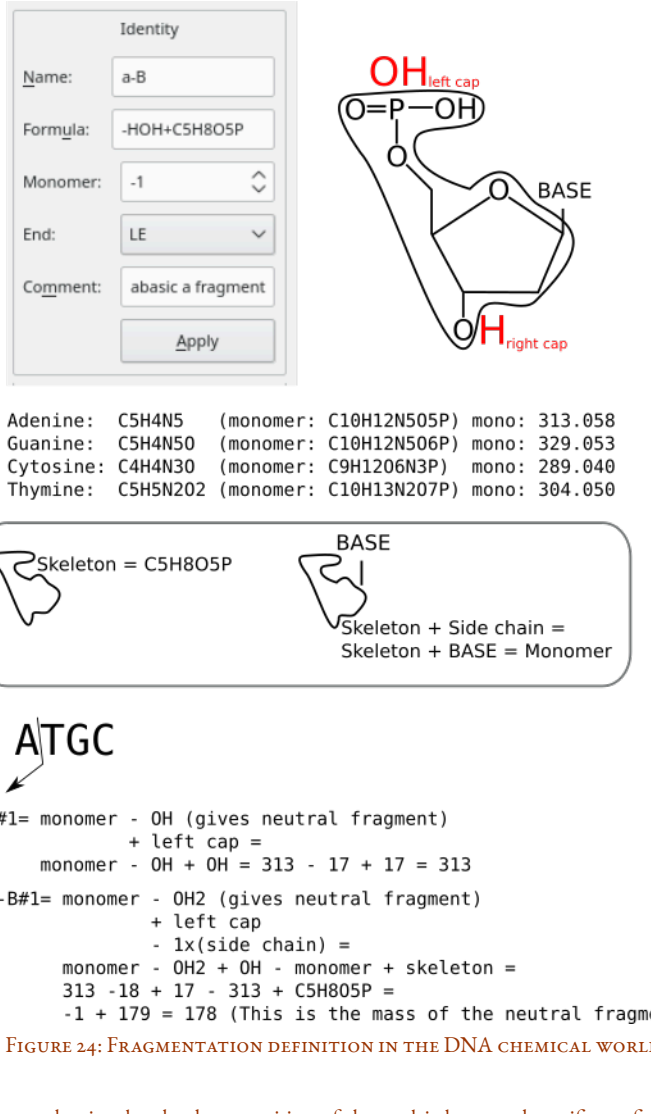


Figure 24: Fragmentation definition in the DNA chemical world

Fragmentation patterns that involve the decomposition of the nucleic base need specific configuration adjustments. Here, the removal of the nucleic base is done by first removing the whole monomer and then readding the skeleteon.

5.2.5.1 Simple fragmentation pathway patterns

One simple example of a polymer chain fragmentation pathway is the formation of *a* fragments within a nucleic acid molecule (DNA, in this example). Bond cleavage occurs right before the sugar-carbon-linked oxygen of the phosphodiester bond linking one deoxyribonucleoside to the next. Thus, the molecular weight of the fragment corresponds to the sum of the monomer masses from the left end of the polymer up to and including the monomer being decomposed *less* one oxygen. Note that this specification yields a nucleic acid ion product that is protonated. We thus need to remove a proton to change its charge to 0, thus the formula of the *a* fragmentation pattern is "−OH". This is illustrated in figure

Figure 25: Basic fragmentation pathway definition

: the name of the specification for fragmentation pattern “a” is `a` ; the formula associated to this fragmentation pattern is `−OH` ; the fragment encompasses the left end of the polymer chain (`End` is set to *LE*); the `Monomer` value is set to *0*, which will be explained later.

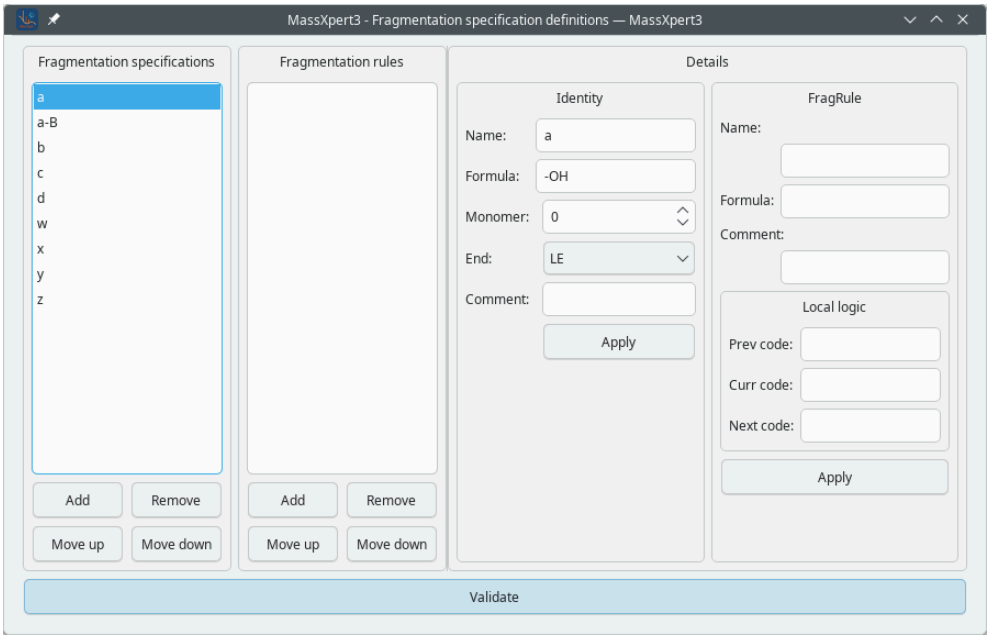


Figure 25: Basic fragmentation pathway definition

Each fragmentation pathway definition involves setting a name, a formula and a local logic, that is, a set of logical conditions which must be verified for the fragmentation specification to be applied to the fragment.

5.2.5.2 More complex fragmentation pathway patterns

In nucleic acids gas-phase chemistry, it often happens that not only the fragmentation occurs at the level of the phospho-ribose skeleton, but also at the level of the nucleic base.

These fragmentation pathways are said to be “*abasic*”. The decomposition of the base occurs at the monomer position where the fragmentation occurs. For example, if a “ATGC” oligonucleotide is fragmented according to pattern *a* but with nucleic base decomposition, and that fragmentation occurs at position 1, then the computation of the mass should occur like represented in

Figure 24: Fragmentation definition in the DNA chemical world

. The figure illustrates a number of things, amongst which some known basics. The top left panel shows what the configuration would be in the fragmentation definition window for this kind of fragmentation. The top right panel shows the basic constituents of the DNA polymer chemistry definition: the caps are “OH” on the left end and “H” on the right end; the circled formula is the skeleton (also called backbone) and the base attached to the deoxyribose ring identifies the nucleotide. That base might be adenine, guanine, cytosine, thymine. In the “dna” polymer chemistry definition, the monomers are made of

the skeleton (formula $C_5H_8O_5P$) plus the formula of the base. The following paragraphs detail two ways to configure a base-loss (*abasic*) fragmentation pattern.

Using a monomer-generic specification In this manner of defining a base decomposition for a fragmentation pathway in the "dna" polymer chemistry definition, the base decomposition is performed in two steps (see

Figure 26: More complex fragmentation pathway definition

):

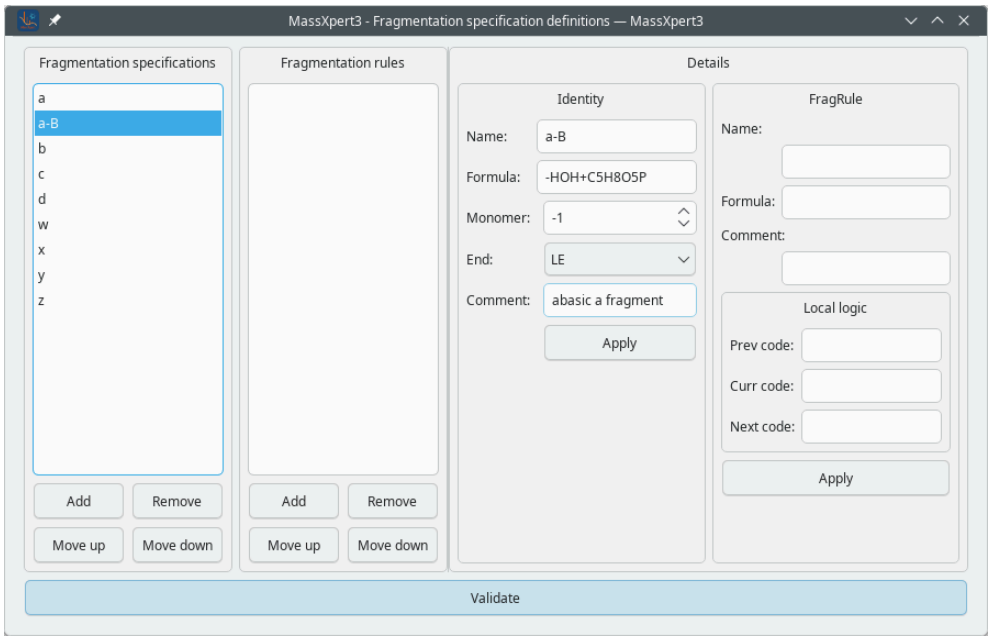


Figure 26: More complex fragmentation pathway definition

In this example, the definition involves accounting for the decomposition of the nucleic base using Monomer and a Formula component.

- Apply an adapted specification for *a* fragments: removal of the H due to the insaturation on the sugar and also removal of the OH related to the formation of an *a* fragment: the −HOH component of the formula;
- Remove one *full monomer* with Monomer set to *−1* (this equals to the removal of both the skeleton and the side chain—the adenine, here);
- Add back the skeleton: the +C5H8O5P component of the formula;
- Add the left end cap, since *a* fragments start at the left end of the fragmented polymer sequence (End is set to *LE*). That *LE* bit of information will be transformed into a +OH formula, since this is the formula of the left end cap for nucleic acids.

The advantage of working this way is that we need not specify a fragmentation rule for each different monomer in the sequence (see below, for how this might be done). Indeed, by specifying Monomer to be *-1*, we tell the mass calculation engine that once the fragmen-

tation has occurred in the polymer chain, the mass of the monomer that got fragmented —without knowing its identity—should be subtracted from the fragment mass. But since removing the whole monomer is more than what is actually decomposed, we add back the skeleton, so that we actually only decompose the nucleic base (adenine, in our example). This is why we ask in the fragmentation specification formula that the skeleton be added back (the *+C5H8O5P* component of the Formula). Because the skeleton does not change along the polymer chain, even if the base itself changes, this computation method is generalizable, and because of this the polymer chemistry definition works.

It might happen that the base decomposition only occurs at specific nucleotide sequence patterns. In this case it might be necessary to define a specific base decomposition rule that accounts for a specific nucleic base. This is described below.

Using a monomer-specific specification Another way of achieving what was described above is by using fragmentation rules, whereby the fragment's mass computation is made conditional to one or more conditions that should be verified as shown in

Figure 27: Fragmentation definition with specific rules

for the *a-b* fragmentation pattern.

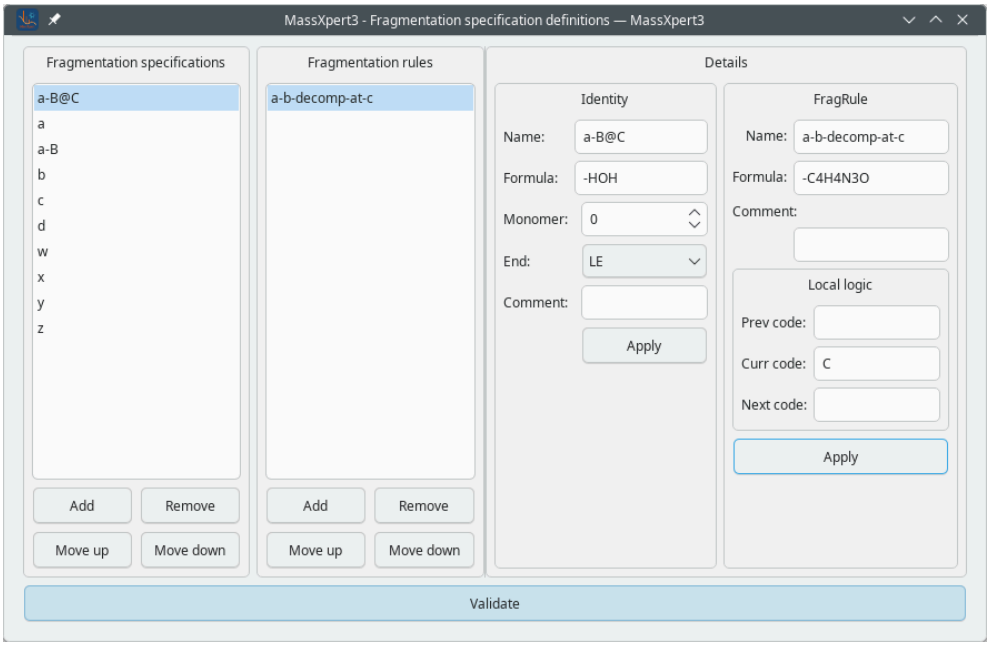


Figure 27: Fragmentation definition with specific rules

The fragmentation for pattern a with decomposition of the nucleic base at the location of the nucleotide undergoing decomposition is defined using a name, a formula and a *"local logic"*, that is, a set of logical conditions which must be verified for the fragmentation rule(s) to be applied to the fragment.

In this example, the *a-b@C* fragmentation pathway definition involves one specific rule (*decompose-C*) that instructs the mass calculation engine to remove specifically the formula of the "C" nucleic base whenever the fragmentation occurs exactly at this nucleotide

(Curr. code in the Local logic group box has value *C*). Here is the detail of this fragmentation pathway definition:

- For the fragmentation specification part, everything is like for fragments of type *a*, that is, the formula is merely −HOH and the End is set to *LE* (see above, for explanations);
- The Monomer field has value *0* because we are not going to remove the monomer's mass;
- One rule named *a-b-decomp-at-c* adds some Local logic for the fragmentation specification: the formula *−C4H4N3O* should be applied upon calculation of the fragment's masses if the monomer at which the fragmentation actually occurs is a cytosine (Curr code is *C*). The Formula of the fragmentation rule is set to *−C4H4N3O*, which is the formula of cytosine (the nucleic base, *not* the monomer).
- There should be other fragmentation rules specified for the remaining three bases A, T, G, but adapting the formula of the fragmentation rule to the actual formula of each base. In this way, the fragmentation rule-based definition of fragmentation pattern *a-b* described here is equivalent to the more generalizable method described earlier in

  Figure 26: More complex fragmentation pathway definition

  .

5.2.5.3 Even more complex fragmentation patterns

In saccharide chemistry the fragmentation patterns are extremely complex, and often totally depend on the nature of the monomers local to the fragmentation site. For example, the fragmentation behaviour at position *E* in a sequence *DEAR* might be different than in a sequence *DERA*.

MassXpert had to be able to model these complex situations, and this is done using fragmentation rules where the local logic involves defining the Prev code and/or the Next code for a given Curr code at which the fragmentation occurs. For example, one specific fragmentation pattern for fragmentation at *E* in sequence *DEAR* might be defined this way:

- Prev code: D
- Curr code: E
- Next code: A.

Instead of that fragmentation rule, one would have for fragmentation at *E* in sequence *DERA* the following rule:

- Prev code: D
- Curr code: E
- Next code: R.

Note the change for Next code, from *A* to *R*.

Also, be aware that the "prev", "curr" and "next" notions are polar, that is, they depend on the value of End (that is, *LE* or *RE*). For example, if we wanted to model the fragmentation pattern at *E* for a fragment of End type *RE*, similar to what was done above with sequences *DEAR* and *DERA*, we would have set the local logic like this:

For sequence "DEAR":

- Prev code: *A*
- Curr code: *E*
- Next code: *D*.

For sequence "DERA":

- Prev code: *R*
- Curr code: *E*
- Next code: *D*.

This highly flexible fragmentation specification allows for highly complex fragmentation pathway defintions matching the complex gas phase fragmentation mechanisms of biopolymers.

5.3 Saving the definition

Once the polymer chemistry definition is completed, the user can save it to an *XML* file. Prior to actually writing to the file, the program checks the validity of all the chemical entities in the definition. This check can be triggered manually by clicking the Validate button. If an error is found, it is reported so that the user may identify the problem and fix it.

The location where the file should be saved, and the manner that it may be made available to *MassXpert* is to be described in a later chapter. It is, in fact, very important that *MassXpert* knows where to find newly defined polymer chemistries so as to be able to use them when sequences of that polymer chemistry are created or used.

6 XpertEdit: a Powerful Editor and Simulation Center

After having completed this chapter you will be able to perform sophisticated polymer chemistry simulations on polymer sequences—that can be edited in place—along with automatic mass recalculations.

6.1 *XpertEdit* Invocation

The *XpertEdit* module is easily called by pulling down the **XpertEdit** menu item from the *MassXpert* program's menu. The user may start the *XpertEdit* module by:

- Opening a sample polymer sequence;
- Creating a new polymer sequence;
- Loading a polymer sequence from disk.

6.2 *XpertEdit* Operation: *In Medias Res*

The first manner to start an *XpertEdit* session is by opening a sample sequence out of the list of sequences that were shipped along with *MassXpert*. The **XpertEdit⇨Open Sample Sequence** menu opens the dialog box shown in Figure 28 – Selection of a sample polymer sequence

The drop-down widget in this dialog window lists all the polymer sequence files that were shipped along with *MassXpert*. Simply select one item and click OK. To select another polymer sequence file, click Cancel, which will trigger the system's file selection dialog for you to browse to the location where the polymer sequence file is stored. The process is identical to the normal polymer sequence file opening (see below).

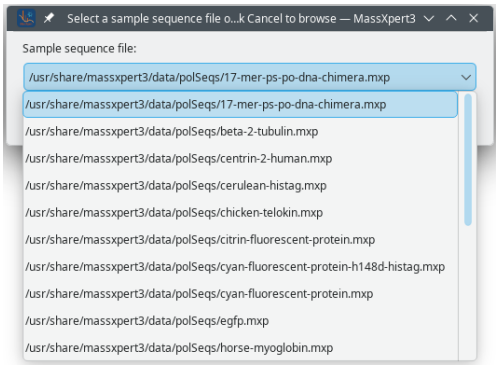


Figure 28: Selection of a sample polymer sequence

MassXpert ships with a number of sample polymer sequences which are designed to easily demonstrate the *XpertEdit* features. This selection dialog lists all the polymer sequence files that were shipped along with *MassXpert*.

The second way to start an *XpertEdit* session is by creating a new polymer sequence using the **XpertEdit⇨New sequence** menu. The program immediately asks to select a polymer chemistry definition, as shown in Figure 29 – Selection of the polymer chemistry definition The drop-down widget lists all the polymer chemistry definitions currently registered on the system. If the polymer chemistry definition is not listed, clicking onto Cancel will let the user browse the disk in search for a polymer chemistry definition file.[7] Once the polymer chemistry definition has been selected and successfully parsed by the program, the user is presented with an empty sequence editor.

The third way to start an *XpertEdit* session is by opening an existing polymer sequence file. Once the sequence file has been opened, the user is presented with a sequence editor as represented in Figure 30 – The *XpertEdit* main window At this point, when the user starts editing a sequence, the characters entered at the keyboard, or pasted from the clipboard, will be interpreted using the polymer chemistry definition that was selected in the initialization window described above.

[7] Once the sequence is saved, the polymer chemistry definition file *must* be registered or the sequence file will not be loadable. This is described in a later chapter.

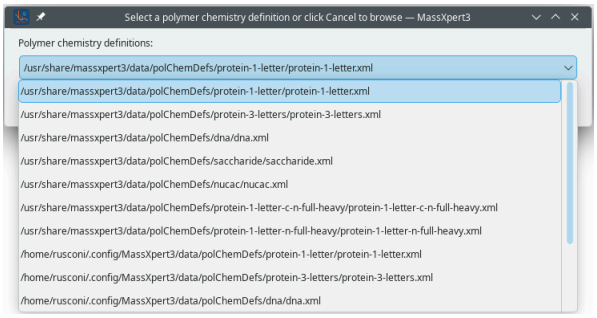


Figure 29: Selection of the polymer chemistry definition

When creating a new polymer sequence, it is necessary to first indicate of what polymer chemistry definition the polymer sequence will be. This window lists all the polymer chemistry definitions currently available on the system.

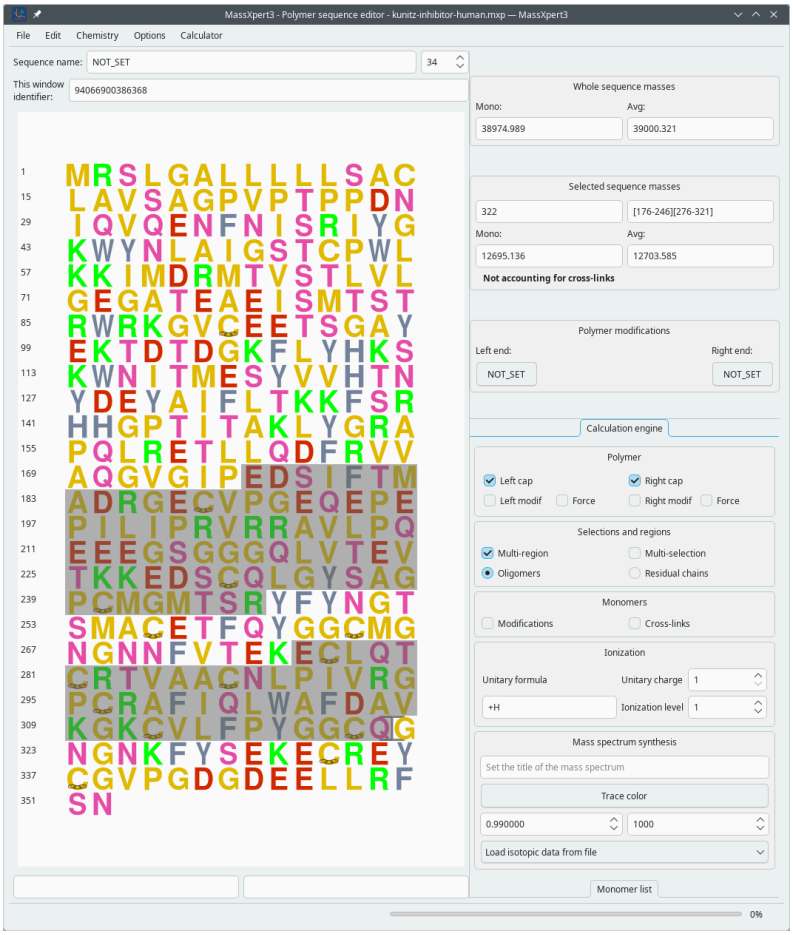


Figure 30: The *XpertEdit* main window

This figure shows a polymer sequence displayed in an *XpertEdit* editor window.

Now, of course, editing a polymer sequence is not enough for a mass spectrometric-oriented software suite; what we want is *to compute masses!* The mass calculation process is immediately visible on the right hand side of the sequence editor shown in Figure 30 – The

XpertEdit main window Two group box widgets are located at the top of the right hand side of the window:

- Whole sequence masses: A group box widget displaying the Mono and Avg masses of the *whole* polymer sequence, irrespective of the current selection;
- Selected sequence masses: A group box widget displaying the Mono and Avg masses of the *currently selected region* of the polymer sequence.

> **Mass or *m/z*?**
>
> Although the labels read masses, the values are actually either masses if the Ionization level is set to *0* or *m/z* values if the ionization level is set to non-*0*.

The user may change the mass calculation engine configuration at any point in time using the widgets in the Calculation Engine tool box that contains the following configurable parameters:

- Polymer:
  - If Left Cap is checked, the left cap of the polymer sequence will be taken into account;
  - If Right Cap is checked, the right cap of the polymer sequence will be taken into account;
  - If Left Modif is checked, the modification of the polymer sequence's left end will be taken into account. Note that if Force is checked also, then the modification is taken into account even when selecting a region of the sequence that does not encompass the monomer located at that end;
  - The behaviour of Right Modif is the same as above, but for the right end modification;
- Selections and regions:

  See Section 6.7 – Multi-region Selections for a detailed explanation of multi-region selection and multi-selection regions.
  - If Multi-region is checked, the sequence editor allows more than one region to be selected at any given time (there is no limitation whatsoever on the number of selected regions);

> **Tip**
>
> In order to select multiple regions, perform the first selection, then maintain the Ctlr key pressed and perform the next selection.

- If Multi-selection is checked, the sequence editor allows not only the selection of multiple regions at any given time, but also the selection of totally or partially overlapping regions.
- When multiple regions are selected, each selected region behaves like an oligomer if Oligomers is selected; that is, it gets its left and right end caps added (if the corresponding calculation engine configuration item is activated);
- When multiple regions are selected, the different regions behave like residual chains if Residual chains is selected; that is, the left and end caps are added only once (if the corresponding calculation engine configuration item is activated).

- Monomers:
- If Modifications is checked, the monomer modifications will be taken into account;
- If Cross-links is checked, the cross-links in the polymer sequence will be taken into account.

> **Important**
>
> Only cross-links fully encompassed by the selected sequence region(s) will be taken into account for the Selected sequence mass calculations. If any number of cross-links are not fully encompassed by the currently selected sequence region, then that number is displayed along with the following label visible in the Selected sequence group box : Incomplete cross-links:.

- Ionization:
- The +H formula represents the ionization agent formula *(that is, a protonation)*;
- The Unitary charge value is set to 1 because, by definition, a protonation brings a single positive charge;
- The Ionization level is set to 1, that is, in the example, the polymer must undergo a single protonation.

> **Tip**
>
> If the ionization agent were Ca^{2+}, the unitary charge would have been *2*. If performing an ionization with level *2*, then the charge of the ion would have been *4*.

When any parameter listed above is changed, the recalculation of the masses—for both the `Whole sequence` and the `Selected sequence`—is triggered and the new masses are updated in their respective line edit widgets described earlier. The fact that the user can specify ionization rules should make it clear that the values that are displayed are actually *m/z* ratios (as long as one ionization is set).

6.3 The Editor Window Menu

The menu bar in the polymer sequence editor displays a number of menu items, reviewed below:

- **File**:
  - **File⇨Close**: close the sequence;
  - **File⇨Save**: save the sequence under a new file name. If the sequence has no filename yet, the user is invited to select a filename;
  - **File⇨Save As**: save the sequence in a new file;
  - **File⇨Import⇨Import Raw**: try to import the sequence. If invalid monomer code characters are found, the user is given a chance to revise the imported sequence;
  - **File⇨Export to Clipboard**: copy the sequence and all the data (masses and calculation options) to the clipboard, in the form of simple text;
  - **File⇨Export to File**: write the sequence to file and all the data (masses and calculation options), in the form of simple text (if a file name was

  already selected, otherwise the user is invited to select a file into which the data are to be written);
  - **File⇨Select export file**: select a file into which the data are to be written.
- **Edit**:
  - **Edit⇨Copy**: Copies the current selected region(s) (if any) to the clipboard. If there are more than one region currently selection, then the user is informed that the copied sequence will correspond to these two sequences joined together.

> **Important**
>
> Be aware that the order in which the region sequences are joined is the order in which the regions were selected and not the order in which the sequences appears in the whole polymer sequence.

- **Edit⇨Cut**: copy the current selection (if any) to the clipboard and removes it from the sequence. Note that it is not yet possible to cut more than one selected region in one single operation;
- **Edit⇨Paste**: paste the sequence from the clipboard into the sequence at the current cursor location. If the pasted sequence is found to contain characters not valid for the current polymer chemistry definition, the user is given a chance to revise the pasted sequence. If one sequence region was selected, it is replaced with the pasted sequence. If more than one sequence region was selected, the operation cannot be performed and the user is informed;
- **Edit⇨Find Sequence**: find a sequence motif in the polymer sequence.

- **Chemistry**:
  - **Chemistry⇨Modify Monomer(s)**: modify (or unmodify) one or more monomers in the polymer sequence;
  - **Chemistry⇨Modify Polymer**: set (or unset) the left (or right, or both) modification(s) of the polymer sequence;
  - **Chemistry⇨Cross-link Monomers**: cross-links monomers;
  - **Chemistry⇨Cleave**: perform a chemical/enzymatical cleavage of the polymer sequence;
  - **Chemistry⇨Fragment**: perform the gas phase fragmentation of the currently selected oligomer;
  - **Chemistry⇨Mass Search**: search in an arbitrary manner for any sequence having a mass matching the searched mass;
  - **Chemistry⇨Compute *m/z* Ratios**: calculate a range of *m/z* ratios with a given ionization agent starting from a given *m/z* ratio and a given ionization status;
  - **Chemistry⇨Determine Compositions**: calculate the monomeric/elemental composition of the whole polymer sequence or of the current selection;

- **Chemistry⇨pKa pH pI**: perform acidity, pH and isoelectric point calculations on the whole sequence or on the current selection.

- **Options**:
  - **Options⇨Decimal places**: set the number of decimal places to be used to display the numerical values.

6.4 Editing Polymer Sequences

As described earlier, in the chapter about the *XpertDef* module, a polymer chemistry definition may allow more than one character to qualify the codes of the monomers (see Section 5 – XpertDef: Definition of Polymer Chemistries). It was noted also that it is not because the number of allowed characters is 3, for example, that all the monomer codes of the polymer chemistry definition must be defined using three characters: 3 is the *maximum* number of characters that may be used.

6.4.1 Multi-Character Monomer Codes

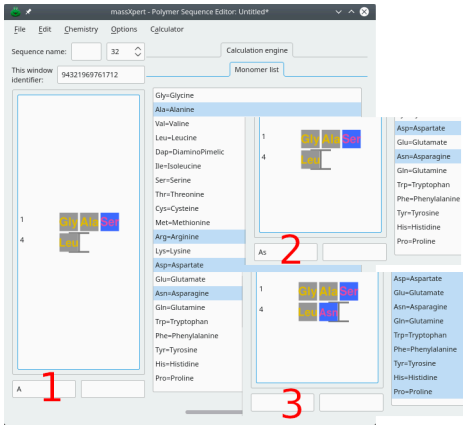


Figure 31: Multi-character code sequence editing in *XpertEdit*

This figure shows the process by which a multi-letter monomer code is entered in the polymer sequence editor.

This section deals with the editing of a polymer sequence for which monomer codes can be made of more than one character. Figure 31 – Multi-character code sequence editing in *XpertEdit* shows the case of a polymer sequence for which the polymer chemistry definition allows three characters to define monomer codes. The example is based on the following real-world situation: the user wants to edit the sequence by insertion—at the end of the sequence (right of the Leu residue). The new monomer to be appended to the sequence is "Asn".

After keying-in `A` (panel 1), no sequence modification is visible in the sequence editor. Instead, an "A" character is now displayed in the left line edit widget under the sequence.

The reason of this apparently odd behaviour is that the polymer chemistry definition allows up to 3 characters to describe a monomer code. If no monomer vignette is displayed in the polymer sequence, that means that more than one monomer code start with an "A" character: *XpertEdit* cannot figure out which monomer code was actually meant by the user when keying-in `A`.

There is a way, called *code completion*, to know which monomer code(s)—in the current polymer chemistry definition—do start with the keyed-in character(s) (currently, "A"). The user can at any moment always enter the *"code completion mode"* by hitting the `Enter` key. This is what is shown in the panel 1st, right hand side Monomer List listview widget (click on that Monomer
List label to show that list if it is not already visible). We see that, in the current polymer chemistry definition, four monomer codes start with an "A" character, and these are "Ala", "Arg", "Asp" and "Asn" (as highlighted in the code completion monomer list).

Because we now know that the code we are to key-in is "Asp", we key-in a `s`. The result is shown in the small panel 2. What we see here is that, this time also, nothing changed in the polymer sequence. What changed is that the character string in the left line edit widget below the sequence is now "As". Let's key-in once more the `Enter` key. This time, only two items are highlighted: "Asp" and "Asn" in the code completion monomer list (panel 2nd). This is easily understood: there are only two monomer codes that start with the two letters "A" and "s" ("As") that we have keyed-in so far. At this time, we key-in a last character: `n`. At this point, the monomer is effectively inserted in the polymer sequence, as the "Asn" monomer left of the cursor, as shown in panel 3. Note how the bottom edit widget is now cleared: there are no more letters in the buffer awaiting to be completed to form a full monomer code. Also, by entering `Enter`, all the monomer codes are selected in the list of monomer codes available for editing the sequence: since a new monomer code might be entered all possibilities are open.

6.4.2 Unambiguous Single-/Multi-Character Monomer Codes

Let's imagine that we have a polymer chemistry definition that allows up to 3 characters for the definition of monomer codes, but that we have one of these monomer codes (let's say the one for the "Glutamate" monomer) that is one-letter-long: "E". This monomer code "E" is the only one in the polymer chemistry definition to start with an "E" character. In this case, when we key-in `E`, we'll observe that the monomer code is immediately validated and that its corresponding monomer vignette is also immediately inserted in the

polymer sequence. This is because, *if there is no ambiguity*, *XpertEdit* will immediately validate the code being edited.

The mechanism described above means that the user is absolutely free to define *only single-character monomer codes* in a polymer chemistry definition that allows multi-character monomer codes; the behaviour of the program is thus to behave exactly as if the multi-character code feature were inexistent in the program: each time a new uppercase letter is keyed-in, it is automatically validated and the corresponding monomer is created in the sequence.

6.4.3 Erroneous Monomer Codes

The typing error detection system triggers immediate alerts whenever the code being keyed-in is incorrect. This is described in

Figure 32: Bad code character in *XpertEdit* sequence editor

. If the user enters an uppercase character not matching any monomer code currently defined in the polymer chemistry definition, or a lowercase character as the first character of a monomer code, the program immediately complains in the right line edit widget below the sequence. In this case, the monomer code is *not* put into the left text widget, which means it is simply ignored.

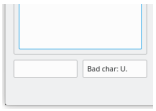


Figure 32: Bad code character in *XpertEdit* sequence editor

This figure shows the feedback that the user is provided by the code editing engine, when a bad character code is keyed-in.

If you start keying-in valid monomer character codes, like for example we did earlier with “As”, and you want to erase these characters because you changed your mind, hit the `Escape` key to remove any character entered previously. These characters will disappear sequentially, at each `Escape` key press, from the line edit widget below the sequence. For example, let’s say you have already keyed-in `A` and `s`. In this case the left line edit widget displays these two characters: “As”. Now, if you change your mind, not willing to enter “Asp” monomer code anymore, but “Gly” instead, all you have to do is to key-in `Escape` once for the “s” character (which disappears) and once more to remove the remaining “A” character. At this point it is possible to start fresh with the “Gly” monomer code by keying-in sequentially `G`, `l` and finally `y`.

> **Warning**
>
> Do not use the `Backspace` key to erase erroneously entered characters, you would delete monomers from the sequence!

6.4.4 Simplified Editing

When the monomer codes of a given polymer chemistry definition are too numerous or too long to remember, one simplified editing strategy is by using the list of available monomers located on the right side of the sequence editor (widget labelled Monomer list ; Figure 33 – Simplified editing with the monomer list). The items in the list are active: if double-clicked, an item will see its corresponding monomer code inserted in the sequence at the current cursor location. This list thus makes it easy to "visually" edit the polymer sequence without having to remember all the codes in the polymer chemistry definition.

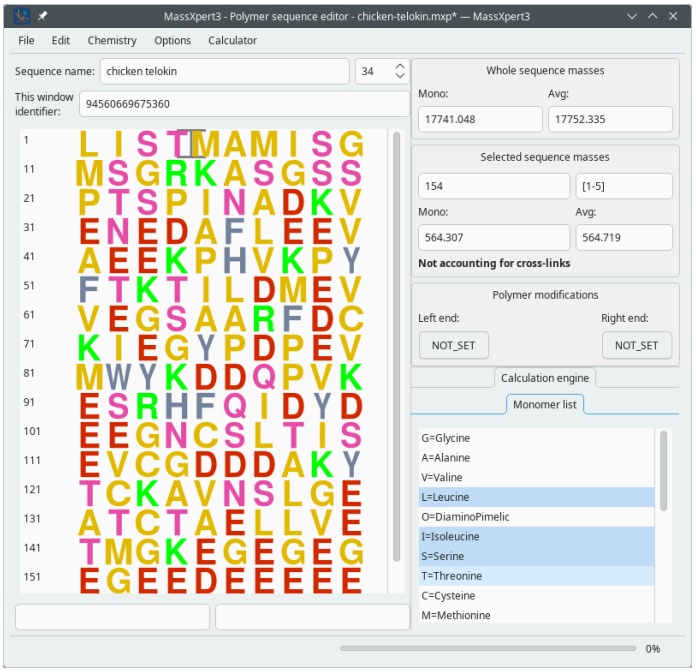


Figure 33: Simplified editing with the monomer list

This figure shows that the sequence has been edited to insert *LIST* by sequentially clicking the corresponding monomer list items.

6.5 Finding sequence motifs

Finding sequence motifs in the polymer sequence is performed by selecting the **Edit⇨ Find Sequence** menu item. The dialog window is shown in Figure 34 – Finding a sequence motif in the polymer sequence. When performing the first search in a polymer sequence,

the Find button should be used. This will trigger a search starting at the beginning of the polymer sequence. For each successive search, the Next button should be used.

Each searched sequence motif will be stored in a history list that is made available by dropping down the combo box widget where the sequence motif is entered. The Clear history button will erase all the searched sequence motifs from the history, thus resetting it.

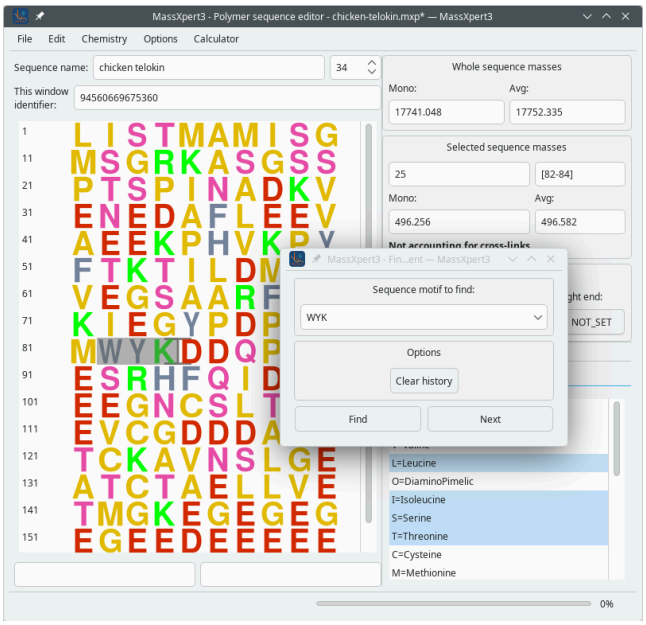


Figure 34: Finding a sequence motif in the polymer sequence

The first iteration should be performed by clicking onto the Find button, and each following iteration should be performed using the Next button.

6.6 Importing Sequences

Very often, the user will make a sequence search on the web and be provided with a polymer sequence that is crippled with non-code characters. That web output might either be saved in a text file for future reference or copied to the clipboard for immediate use in *MassXpert*. The two cases are reviewed below.

6.6.1 Importing From The Clipboard

XpertEdit provides a convenient way to spot non-valid characters in a text and to let the user "purify" the imported sequence. A clipboard-imported sequence is systematically parsed. When invalid characters are found, the window depicted in Figure 35 – Clipboard-imported sequence error-checking is presented to the user for them to make appropriate adjustments (in this example we tried to copy from clipboard the following sequence into a polymer sequence of the "dna" polymer chemistry definition: "!100 ATGCATGC ATGCATGC ATGCATGC ATGCAUGC anotherSilly-Text;").

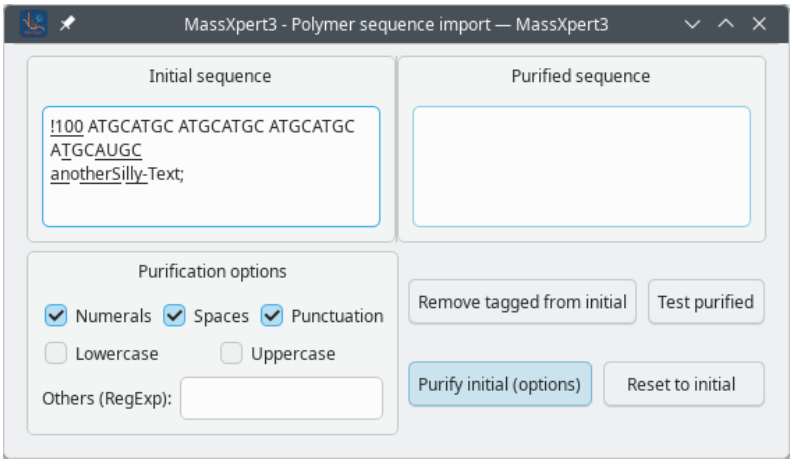


Figure 35: Clipboard-imported sequence error-checking

If a sequence that is imported through the clipboard to the *XpertEdit* sequence editor contains invalid characters, the user is provided with a feature to "purify" the sequence. Often, clicking the Remove tagged from initial button is sufficient to remove all spurious characters.

As soon as a character does not correspond to any valid monomer code, it is tagged, and the sequence is presented to the user in a text edit widget (Initial Sequence) with the all the invalid characters tagged by underlining. At that point, if the user clicks the Remove Tagged From Initial button, all the tagged characters will be automatically removed and the purified sequence will show up in the Purified Sequence text edit widget.

Also, the user is provided with automatic "purification" procedures whereby it is possible to remove one or more classes of characters from the imported sequence (Purification Options frame widget). Checking one or more of the Numerals or Spaces or Punctuation or LowerCase or Uppercase checkbuttons, or even entering other user-specified regular expressions in the Other (RegExp) line edit widget, will elicit their removal from the imported sequence after the user clicks the Purify Initial (Options) button.

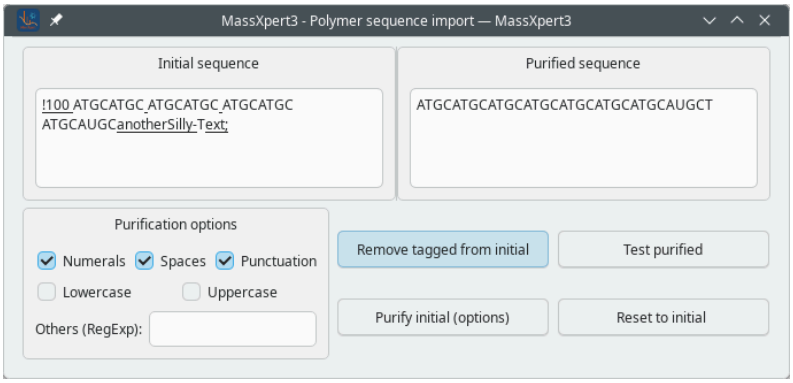


Figure 36: Clipboard-imported sequence purification

There are a number of ways to purify a sequence. Here the Remove Tagged From Initial button was clicked. The purified sequence shows up in the Purified Sequence text edit widget.

← *XpertEdit: a Powerful Editor and Simulation Center*

XpertCalc: a Powerful Mass Calculator →

When you are confident that almost all the erroneous characters have been removed (Figure 36 – Clipboard-imported sequence purification), you can click the Test Purified button, which will trigger a "re-reading" of the sequence in the Purified Sequence text edit widget. If erroneous characters are still found, they are tagged.

> **Note**
>
> For maximum flexibility, you are allowed an immediate and direct editing of the purified sequence in the Purified Sequence text edit widget (that is, that text edit widget is *not* read-only).

Once the sequence is finally depured from all the invalid characters, you can select it in the text edit widget and paste it in the *XpertEdit* sequence editor. This time, the paste operation will be error-free.

6.6.2 Importing From Raw Text Files

It might be of interest to be able to import a sequence from a raw file. To this end, the user is provided the menu **File⇨Import⇨Import Raw** that opens up a file selection window from which to choose the file to import. The program then iterates in the lines of that file and checks their contents for validity. If errors are found, then the same process as described earlier for clipboard-imported sequences is started (Section 6.6.1 – Importing From The Clipboard). The user can then purify the sequence imported from the file and finally integrate that sequence in the polymer sequence currently edited. Note that if any sequence portion is currently selected, it will be replaced by the one that is being imported.

6.7 Multi-region Selections

MassXpert implements a sophisticated multi-region selection model. Two selection modes are available:

- *Multi-region selection mode:*: In this mode, it is possible to select more than one region in the polymer sequence. In all cases below, make sure that the Multi-region checkbutton is checked in Selections and regions group box (see Selections and regions). This is how these selections are performed:
  - *With the mouse*: Left-click and drag to make the first selection. Go with the mouse cursor at the beginning of desired new selection, hold the `Ctrl` key down while left-clicking and dragging to perform the second region selection. Continue as may times as necessary;

- *With the keyboard*: Position the cursor at the beginning of the first region to be selected, hold the two `Ctrl` and `Shift` keys pressed while moving the cursor with the direction keys (the arrow keys of the keyboard). Hold the `Ctrl` key down and use the direction keys to go to the beginning of the new region selection, press the `Shift` key and hold it down while moving the cursor with the direction keys to actually perform the region selection.

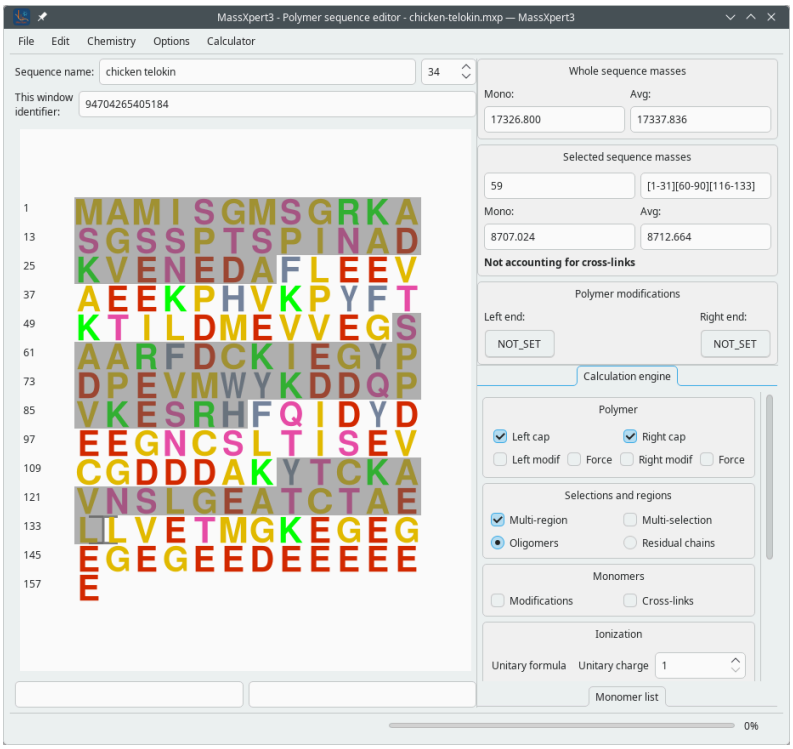


Figure 37: Multi-region selection

The distinct regions of the sequence have been selected.

- *Multi-selection region mode*: In this mode (which requires the multi-region selection mode to be enabled), it is possible to perform selections that overlap. For example, one could select the sequence “MAMISGM” and then select the sequence “SGMSGRKAS”. The overlapping sequence is thus “SGM”.

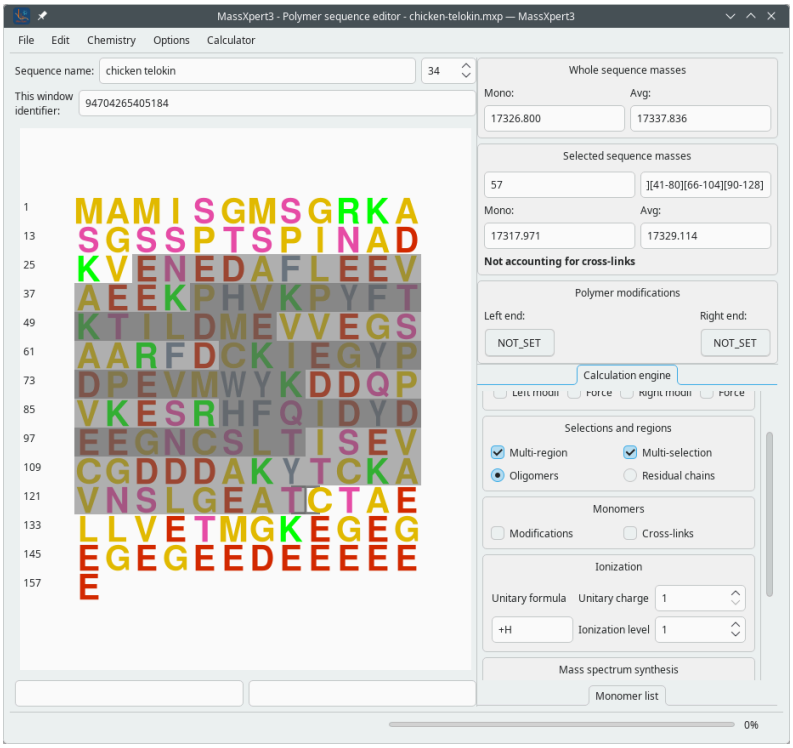


Figure 38: Multi-region selection

The distinct regions of the sequence are overlapping.

Being able to select multiple regions and/or to select multiple times the same region involves some configurations, as far as calculating relevant masses is concerned. Indeed, whatever the selection mode that is enabled, each time one selection (overlapping with another or not) is added or removed, masses are recalculated for the current selection.[8] The way the multi-region selections and the multi-selection regions are handled, from the mass calculation standpoint, is configured as follows:

- *Regions are oligomers*: In this configuration, each selection behaves as an oligomer, and is thus capped on both its left and right ends. This is typically the situation when the user wants to simulate the formation of a cross-linked species arising from the cross-linking of two oligomers: each oligomer is capped on both its ends;

- *Regions are residual chains*: In this configuration, each selection behaves as a residual chain, and thus the oligomer resulting from the multi-region selections is capped on its left and right ends only once. This situation is typically encountered when simulating partial cleavages by first selecting an oligomer, checking its mass and then continuing selection to simulate a longer oligomer resulting from a partial cleavage. Also, the situation might be encountered when there are multiple repeated sequence motifs in a polymer sequence and mass data are difficult to analyze.

[8] "Selection", here, is thus used to collectively represent all multi-region selections and multi-selection regions at any given time in the polymer sequence editor.

6.8 Polymer Sequence Modification

It very much often happens that (bio) chemists use chemical reactions to modify the polymer sequence they are working on. Mass spectrometry is then often used to check if the reaction proceeded properly or not. Further, in nature, chemical modifications of biopolymer sequences are very often encountered. For example, protein sequences get often modified as a means to regulate their function (phosphorylations, for example, or acetylations, methylations...). Nucleic acid sequences are very often and extensively modified with modifications such as methylations...

It is thus crucial that *MassXpert* be able to model with high precision and flexibility the various chemical reactions that can be either made in the chemistry lab or found in nature. The *MassXpert* program provides two different chemical modification processes:

- A process by which monomers belonging to the polymer sequence can be individually modified;
- A process by which the whole polymer sequence can be modified, either on its left end or on its right end or even on both ends.

6.8.1 Selected Monomer(s) Modification

There are a number of manners in which monomers can be modified in a polymer sequence. Figure 39 – Selected monomer modification shows the simplest manner: the user first selects the monomer vignette to be modified and calls the **Chemistry⇨Modify Monomer(s)** menu. A window shows up where all the modifications currently available in the polymer chemistry definition are listed. Because a monomer vignette was initially selected in the editor window, the Current selection target radiobutton is on by default.

> **Note**
>
> Note that if a sequence was selected when the monomer modification task was started, then, selecting Current selection would be required to modify all the monomers in the selection. Alternatively, if this is not what is required, re-selecting the right monomer in the sequence and selecting Current selection will ensure the modification applies only on the currently selected monomer.

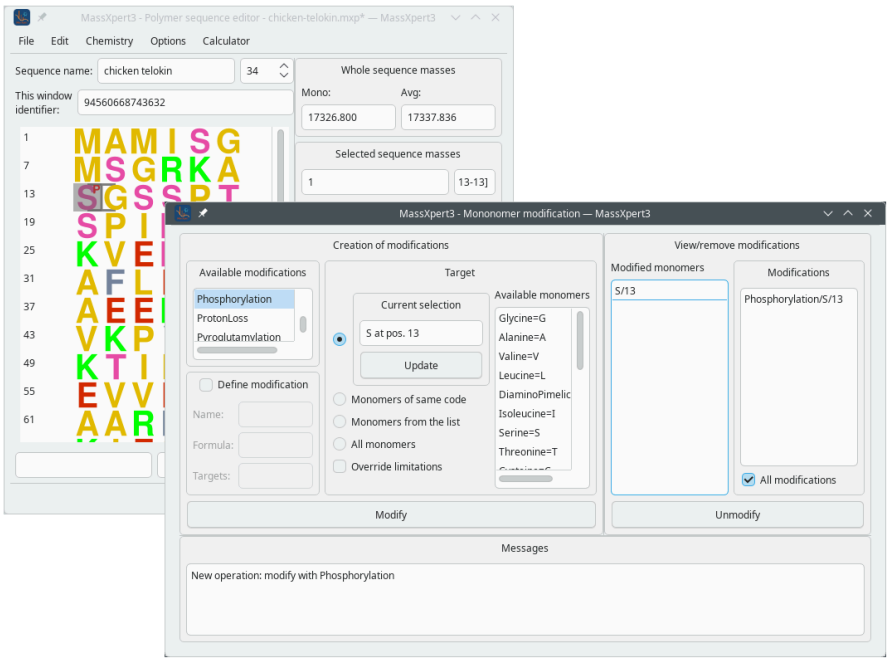


Figure 39: Selected monomer modification

The target monomer was selected to be *Ser13* and the modification was selected to be *Phosphorylation*

It is then simply a matter of choosing the right modification from the Available modifications list and clicking onto the Modify button. The target(s) of a given modification (as selected in the Target frame widget) can be identified according to:

- The Current selection group box will show if a single monomer vignette was selected at the time the monomer modification action was invoked (exactly as in Figure 39 – Selected monomer modification).

> **Warning**
>
> Only the monomer of which the code and the position are displayed will be modified (even if it is no more selected or if the sequence has changed and the monomer at the displayed position is not the same anymore).

- If the Monomers Of Same Code radiobutton widget is

selected, all the monomers in the sequence that have their code identical to the one shown in the Current selection line edit widget are modified, as shown below;

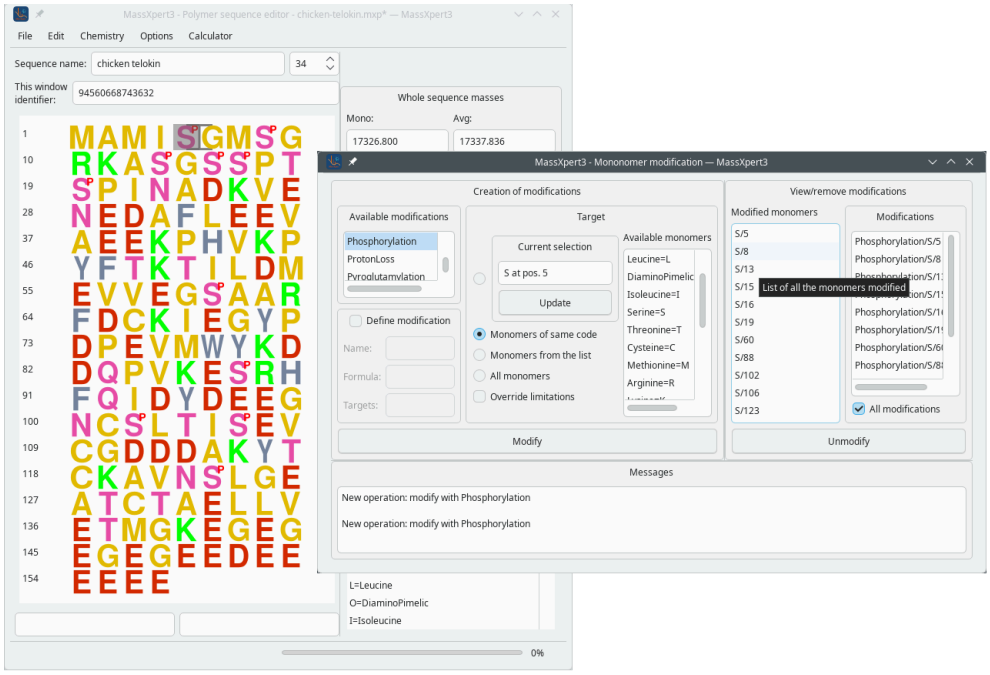


Figure 40: Modification of all the monomers of the same code

The currently selected monomer was used to a global modification step.

- If the Monomers from the list radiobutton widget is selected, all the monomers in the polymer sequence having a code corresponding to any code selected in the Available Monomers list are modified;

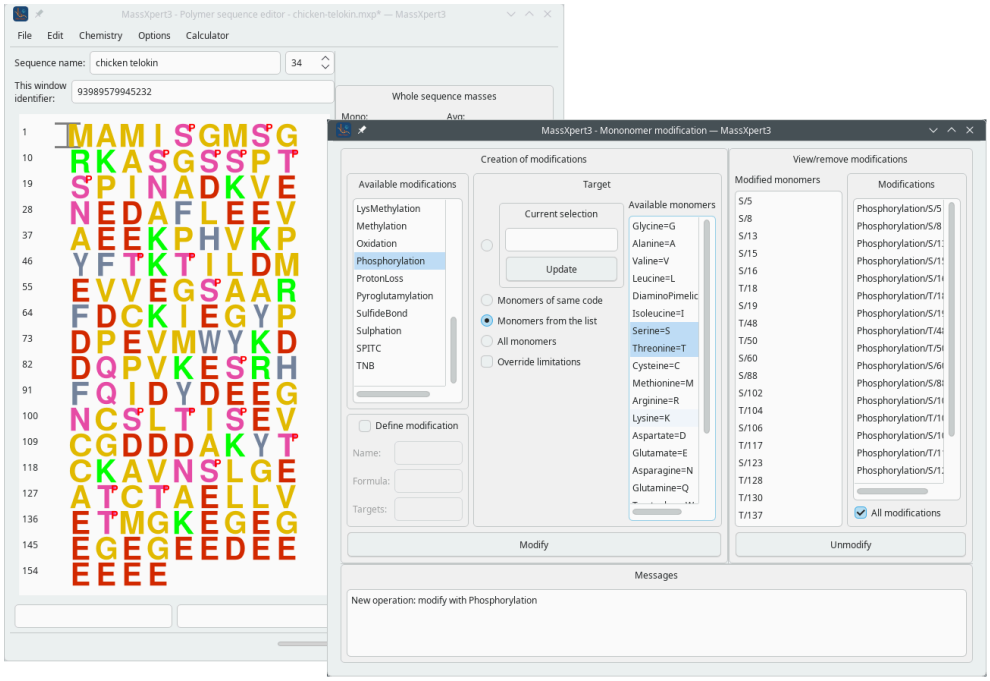


Figure 41: Modification of all the monomers selected in the list

All the *S* and *T* monomers have been phosphorylated.

- If the All Monomers radiobutton widget is selected, all the monomers of the polymer sequence are modified;

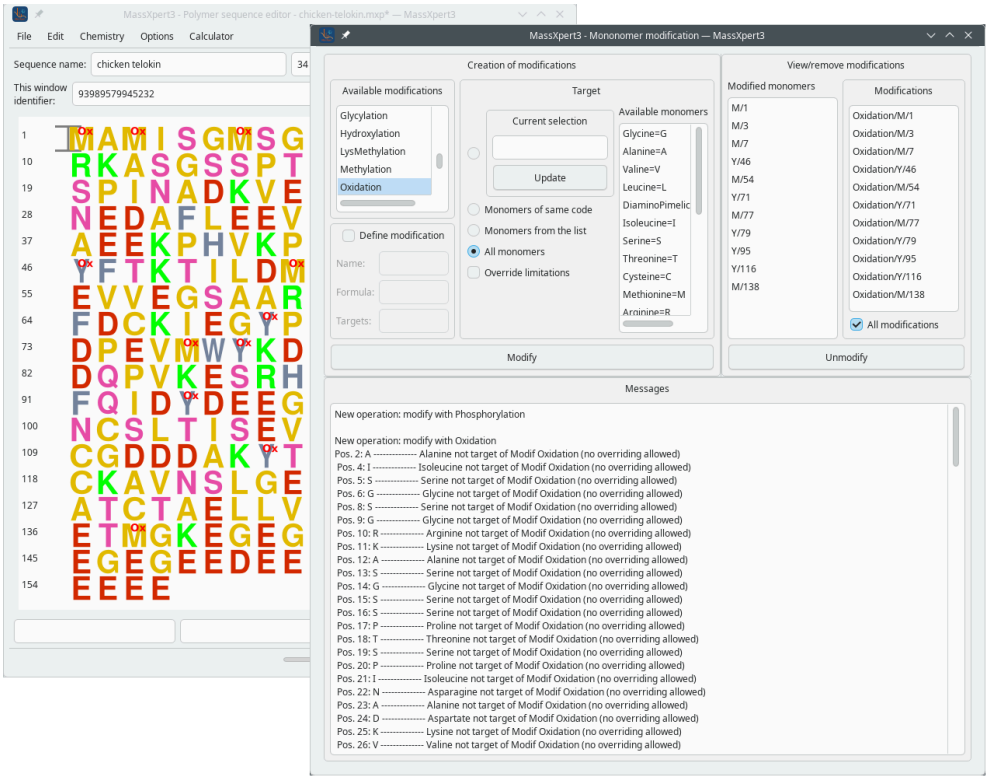


Figure 42: Modification of all the monomers

All the monomers of the sequence should have undergone modification by *Oxidation*. However, since this modification has been defined in the polymer chemistry definition to have *;M;Y;* as targets, only these residues did actually get modified. The other residues are listed in the lower panel of the window as failing for the modification.

Note that there is one checkbox widget (Override limitations) that requires an explanation. In the chapter about the definition of polymer chemistries (Section 5 – XpertDef: Definition of Polymer Chemistries) the definition of modifications was detailed, and the "target" notion was explicited.

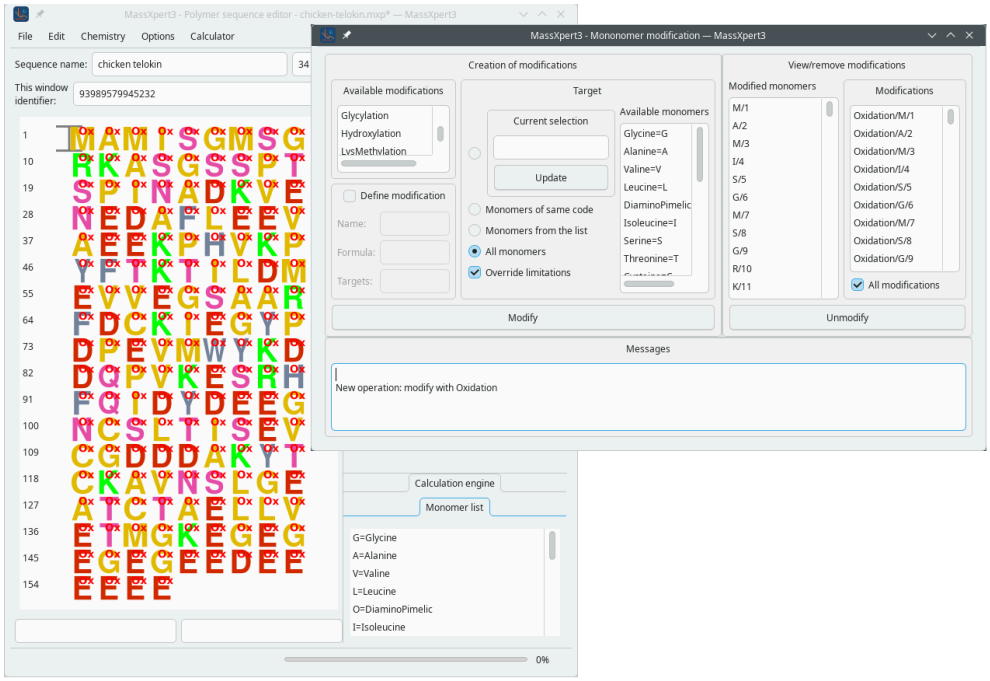


Figure 43: Modification of all the monomers in a polymer sequence

This figure shows how removing the target limitations permits to effectively modify by *Oxidation all* the monomers in the sequence.

If, during a monomer modification, *MassXpert* detects that the user is trying to modify a monomer that is not a target of the modification at hand, it will complain, as shown in the Messages text edit widget of Figure 42 – Modification of all the monomers.

Another situation where target limitations might show up, is when trying to modify a monomer more times than authorized by the Max. count number of times that the monomer might be modified at once with that modification. For example, when working on methylation of proteins, it might happen that lysyl residues get methylated more than once at a time (tri-methylation occurs often in histones). If the chemical modification was defined in *XpertDef* with a maximum modification count of 2 and a third chemical modification is asked on a given target monomer, then the program refuses to perform the modification. To override this limitation, check the Override target limitations checkbox widget.

The general concept about this is: the Override limitations checkbox widget is unchecked by default so that the user does not do mistakes without knowing. However, flexibility is desirable, and that checkbox widget can be checked if required.



Figure 44: Modification of a monomer in a polymer sequence

This figure shows how the chemical modification of monomer(s) is rendered in the sequence editor.

As a result of the monomer modification, the monomer vignette gets modified. Figure 44 – Modification of a monomer in a polymer sequence shows one phosphorylated Seryl residue at position 13: a transparent graphics object (a red "P") was overlaid onto the corresponding seryl monomer vignette. If the user modifies a monomer with a modification that has no corresponding *SVG*-formatted file defined for its graphical rendering in file `modification_dictionary`, then a default modification rendering is used.

The user is responsible for correctly reading the messages that might be published in the `Messages` text edit widget. It is important to understand that, when a monomer is modified, its previous modification (if any) is overwritten with the new one. The user is invited to experiment a bit with the monomer modification process, so as to be confident of the results that she is going to obtain when real polymer chemistry work is to be modelled in *MassXpert*.



Tip

If the modification to be applied is not readily available in the list of modifications defined in the polymer chemistry definition, then it is possible, by checking the `Define modification` check button widget to manually define a modification. This procedure leads to the modification of the target monomer(s) exactly as if the modification had been selected from the list of available modifications. But, because the modification has a name not known to the polymer chemistry definition, the editor cannot modify the monomer vignette with a predefined transparent raster image. Thus, as seen on Figure 45 – Rendering of a monomer modification in a polymer sequence, the modified residue gets visually modified using the default transparent raster image (4 interrogation marks, one at each corner of the monomer vignette square).



Figure 45: Rendering of a monomer modification in a polymer sequence

This figure shows how the chemical modification of monomer(s) is graphically rendered. The “S” residue is modified using an “Phosphorylation” modification. The “I” residue is modified with a modification that has no associated graphical vignette. The default vignette is thus used.

It is perfectly feasible to modify a single monomer more than once (with the same modification or not; for example a tri-methylation with a methylation modification). This is why when the window depicted in Figure 46 – Seryl residue modified twice with *Phosphorylation* shows up, the two lists at the right hand side show the monomers currently modified and the modification(s) that are currently set to these modified monomers. Selecting one item from the Modified monomers list will show only the modifications set to that monomer in the Modifications list. If all the modifications in the polymer sequence are to be displayed then, checking the All modifications check box widget will trigger the display of all the modifications set to any monomer in the whole polymer sequence.

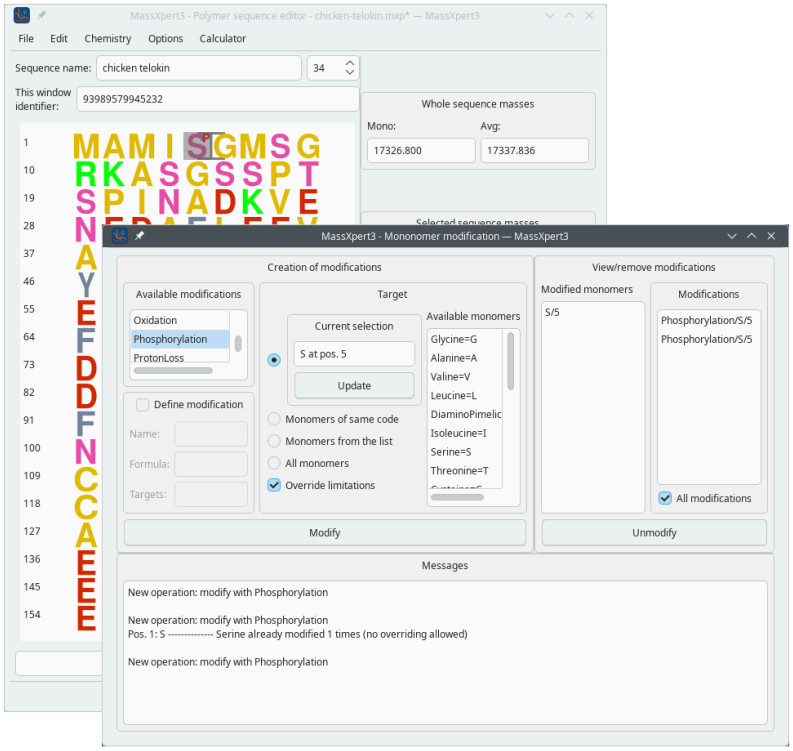


Figure 46: Seryl residue modified twice with *Phosphorylation*

The seryl residue was modified twice with *Phosphorylation*, which required checking the Override limitations check box. The two modifications of the same seryl residue are visible on the right hand side of the window. The Message widget shows the three steps that led to the bi-phosphorylation of the seryl residue.

Unmodification of monomers is easily performed by selecting any number of items from the Modifications list and clicking the Unmodify button.



Warning

It should be noted that once a monomer modification dialog window has been opened, the polymer sequence should not be edited. This is because the modification/unmodification process takes for granted that the polymer sequence is still identical to what it was when the monomer modification dialog was opened. Mecanisms are there to ensure that the irreparable does not happen, but this warning is in order.

6.8.2 Polymer Modification

As described above, it is possible to modify any monomer in the polymer sequence; when any modified monomer is removed, the modification associated to it goes away with it. The modifications that we describe here are not of this kind. They can be applied either to the left end of the polymer sequence or to its right end (or both). But these modifications do belong to the polymer sequence *per se* and are not removed from it—even if the polymer sequence is edited by removing the left end monomer or the right end monomer. This is why these modifications are *polymer modifications* and not monomer modifications.

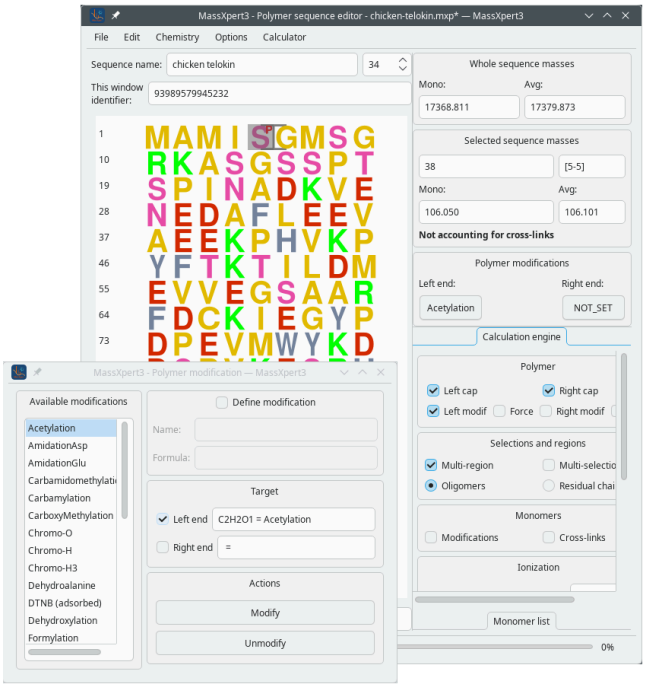


Figure 47: Modification of the left end of a polymer sequence

This figure shows how simple it is to permanently modify a polymer sequence on either or both its left/right ends. The polymer was acetylated on its left end and the calculation engine was configured to account for the Left modif.

The way in which a polymer sequence is modified using *polymer modifications* is much easier than the previous *monomer modifications* case. The modification window is opened by choosing the **Chemistry⇨Modify Polymer** menu as shown in Figure 47 – Modification of the left end of a polymer sequence.

If a formally defined modification is not available, you can define manually the modification that is needed (check the Define modification checkbox to that effect) and then apply to the polymer end of interst. This is illustrated in Figure 47 – Modification of the left end of a polymer sequence for the C-terminal end (the right end) of the protein. The modification object used is created on-the-fly by the program and gets saved in the file as if the user had selected a modification out of the list of available modifications. In the example, the polymer sequence was modified on its left end using the *Acetylation* available in the polymer chemistry definition. The polymer sequence editor window displays the left end and right end modifications as labels of buttons located in the Polymer modifications groupbox widget right above the Calculation engine configuration widget.

6.9 Monomer Cross-linking

A cross-link is a covalent bond that links a monomer with one other monomer. A monomer might be cross-linked more than once. The dialog window in which the user might define cross-links is shown in Figure 48 – Cross-linking of monomers.

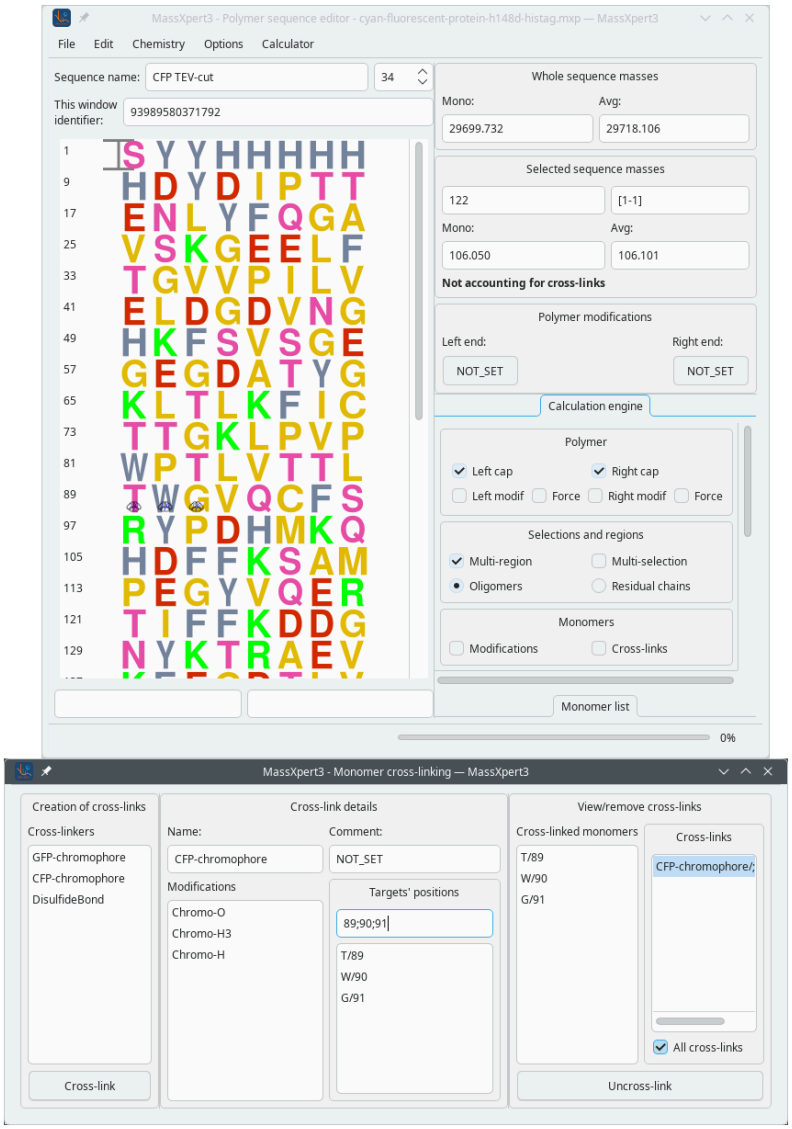


Figure 48: Cross-linking of monomers

This figure shows the window in which monomers can be cross-linked together. A cross-link (as defined in the current polymer chemistry definition) is selected and the targets are specified in the Targets' positions text line edit widget in the form of monomer positions separated by ";" semicolumns.

Cross-linkers were defined in the *XpertDef* chapter (Section 5.2.3 – The cross-linkers). A cross-linker might either define no modification to be applied to the cross-linked monomers or the same number of modifications as there are monomers cross-linked. For example, fluorescent proteins have a chromophore that is made by reaction of three

residues (Threonyl [or Seryl]–Tryptophanyl [or Tyrosinyl or Phenylalanyl]–Glycyl), as shown in Figure 48 – Cross-linking of monomers. When cross-linking with the fluorescent protein cross-linker, there must be three monomers involved as these are three modifications defined in the cross-linker.

When any monomer involved in a cross-link is edited off a polymer sequence, the cross-link(s) it was involved in are automatically dissolved and destroyed. Destruction of a cross-link might be performed by selecting the cross-link in the Cross-links list widget at the right hand side of the dialog window depicted in Figure 48 – Cross-linking of monomers and by clicking the Uncross-link button.

6.10 Sequence Cleavage

It happens very often that polymer sequences get cleaved in a sequence-specific manner. These specific cleavages do occur very often in nature, and are made by enzymes that do cleave biopolymer sequences, like the glycosidases (cleaving saccharides), the proteases (cleaving proteins) or the nucleases (cleaving nucleic acids). But the scientist also uses purified enzymes or chemicals to perform such cleavages in the test tube. *MassXpert* must be able to perform those cleavages *in silico*.

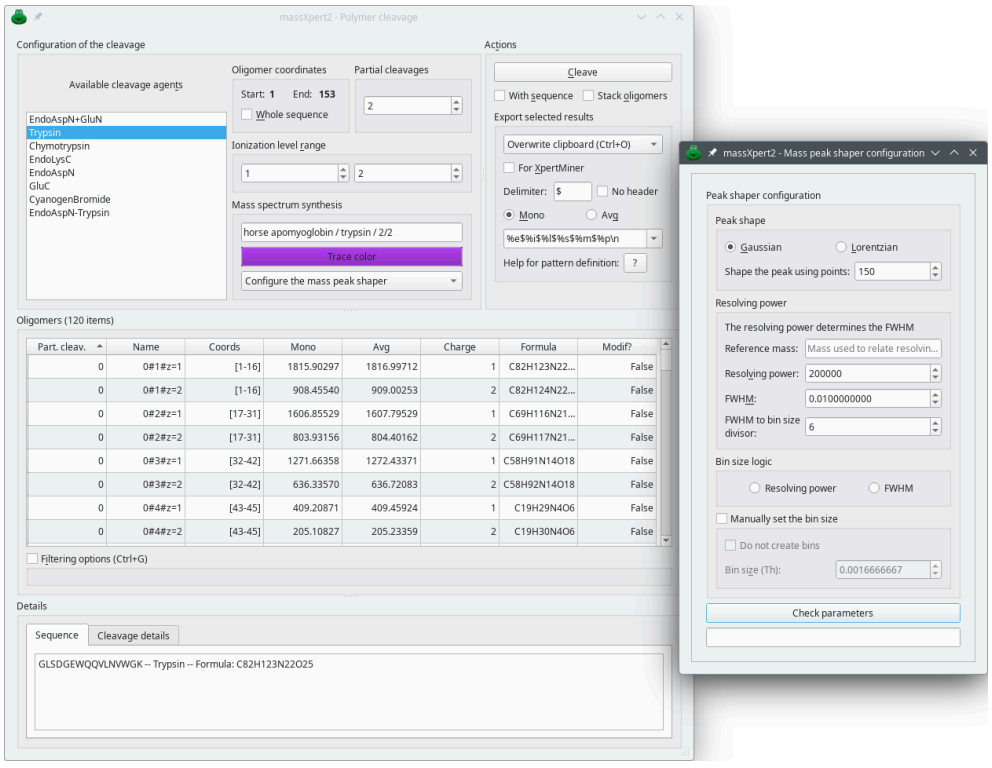


Figure 49: Polymer sequence cleavage window

This figure shows the window in which polymer sequence cleavages are performed. One cleavage agent name is selected, the number of allowed partial cleavages and the ionization level range is set. The results are displayed in the same window. The cleavage might be performed on the currently selected polymer sequence region or the whole sequence.

Starting a polymer sequence cleavage is a matter of having a polymer sequence opened in an editor window and selecting the **Chemistry⇨Cleave** menu. The user is provided with a window where a number of cleavage agent names are listed (Figure 49 – Polymer sequence cleavage window), along with options that permit the customization of the production of oligomers. The parameters that can be set are the following:

- Available cleavage agents: list of the cleavage agents available in the polymer chemistry definition (see Section 5.2.4 – The cleavage agent definitions).;
- Oligomer coordinates: when the window is opened, this group box widget lists the coordinates of the currently selected region of the polymer sequence. Either leave the values as they are shown or check Whole sequence. If Whole sequence is checked, the cleavage will be performed over all the sequence length. Otherwise it will be performed over the sequence in the Start–End range. This feature, which was introduced in version 2.3.0, is useful so as to simulate a first cleavage of a polymer sequence and then a second cleavage of a selected oligomer using a different cleavage agent. In protein chemistry, that would be useful to explore possibilities of double sequential cleavages of a protein, first with EndoAspN, for example, and then with Trypsin.
- Setting the Partial cleavages number defines if the cleavage must be total (value of 0) or if missed cleavages are allowed.
- Setting the Ionization level range defines what charge state the generated oligomer will have.

A number of other features might be configured, either on the way oligomers should be exposed or on the way they should be exported to a number of destinations:

- Checking With sequence tells the cleavage engine to store their sequence in the generated oligomers; see Cleavage oligomer details.
- In normal operation, when multiple cleavages are performed by clicking Cleave, the new oligomer set replaces the one obtained previously. However, it might be useful to generate in the same unique list all the oligomers generated for different cleavages. To stack oligomers from different cleavages, check Stack oligomers.
- It is possible to export the oligomers selected in the tableview widget in a number of ways. Select the destination in the drop down menu widget. A specific export format might be defined and the mass type to be exported might be selected. When exporting for the *XpertMiner* module, check For XpertMiner.

> **Note**
>
> If the list of monoisotopic or average masses is desired in the form of a text list, right-clicking onto the tableview widget will allow copying to the clipboard either the monoisotopic or the average masses. Also, it is possible to either export the data to the clipboard or to a file or even to drag the displayed oligomer items in a text editor. Only the selected items in the treeview widget will be exported.

- Mass spectrum synthesis: This group box is about the configuration of the way a mass spectrum might be calculated and displayed in another program. The Trace color button allows one to select a color with which to plot the mass spectrum's trace once exported to a mass data visualizing program (*MineXpert*). The line edit widget above the button allows one to define a title for the mass spectral trace plotted. Finally, the drop down menu holds three menus, as described below:

  - Configure the mass peak shaper: Click this menu item to display the Mass peak shaper configuration dialog window as shown on Figure 49 – Polymer sequence cleavage window. The configuration of the mass peak shaper works in this dialog window exactly as described in Figure 72 – Setting-up of the centroid mass peak shaping process.

  - Synthesize the mass spectra: This button actually triggers the computation of the mass spectra (the isotopic clusters) of all the selected fragmentation oligomers currently selected in the table view. The calculation is only performed if the previous configuration step has yielded a valid isotopic cluster shaper configuration.

  - Mass spectrum to clipboard: This menu item copies the mass spectrum to the clipboard. For this menu to be available, a mass spectrum synthesis must have been performed successfully, otherwise the menu reads Mass spectrum not yet available.

    > **Note**
    >
    > Upon successful completion of a mass spectrum synthesis, the obtained mass spectral trace is automatically made available to any requesting program *if and only if* as server is started on the *MassXpert* side and *if* a client is started on the *MineXpert* side. See Section 9.1 – Configuration of the *MassXpert* and *MineXpert* intercommunications for details. When the mass spectral trace is successfully displayed in *MineXpert* via the

network, the trace color and title are honoured. When the trace is copied to the clipboard and then loaded in *MineXpert*, the color and title are not transferred along with the mass spectral data.

The Details frame widget at the bottom of the window displays a number of informative data. In particular, the Sequence tab widget displays the sequence of the oligomer currently selected in the Oligomers table view along with the name of the cleavage agent which it arose from. The Cleavage Details tab widget displays the mass calculation engine configuration at the time the *last* cleavage was performed (checked items mean that the corresponding feature was on, unchecked items mean that the related feature was off). In the example (Figure 49 – Polymer sequence cleavage window, the mass calculation for the oligomers did not account for the monomer modifications nor for the left/right end modifications of the polymer, nor for the cross-links.

When the user triggers a cleavage, the mass calculation engine configuration currently set in the sequence editor is used for the calculation of the mass of the oligomers obtained *per* the cleavage. This process allows an easy change in the mass calculation engine configuration between one cleavage and another so as to allow comparison of masses obtained for the same cleavage but with different mass calculation engine configurations.

For oligomer data filtering, please refer to Section 6.13 – Oligomer Data Filtering.



Tip

It is possible to display any given fragment in the polymer sequence editor simply by double-clicking its corresponding row in the table view widget.

6.11 Oligomer Fragmentation

It happens very often that polymer sequences need to be fragmented in the gas phase (in the mass spectrometer) so that structure characterizations may be performed. For protein chemistry, this happens very often in order to get sequence information for a given peptide ion selected in the gas phase. *MassXpert* must be able to perform those fragmentations *in silico*. Let's see how an oligomer can be fragmented using *MassXpert*.

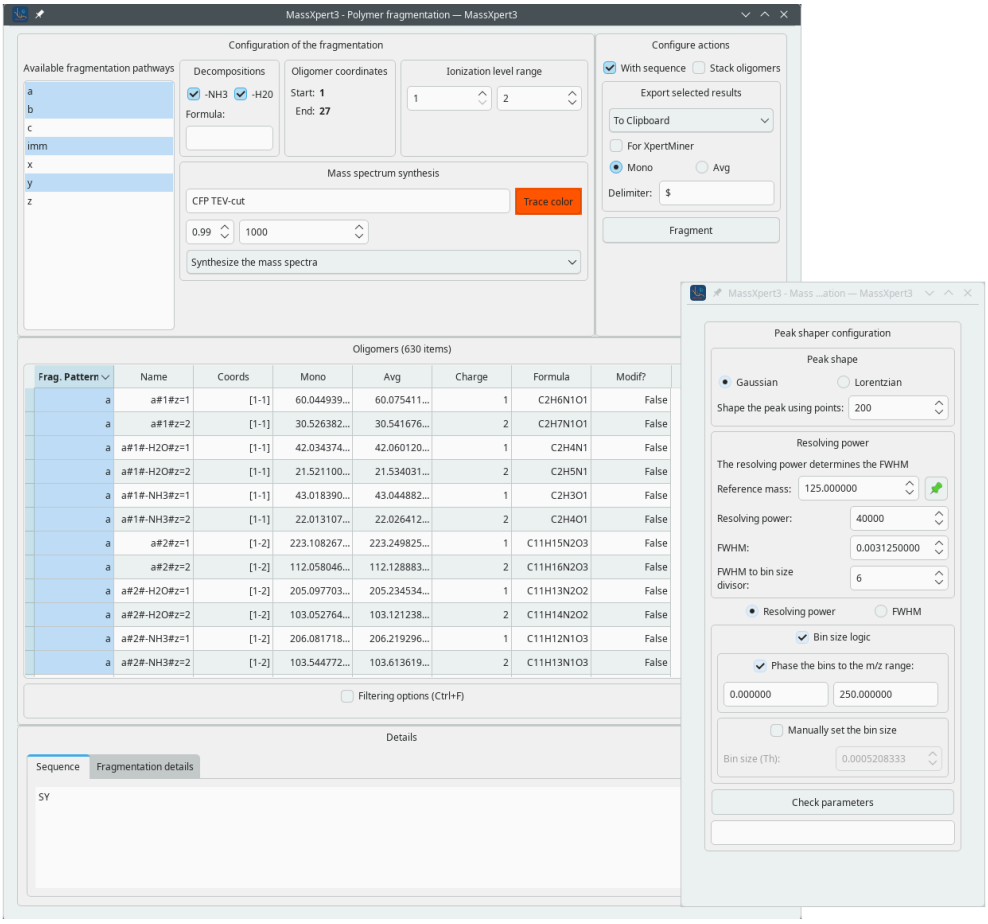


Figure 50: Oligomer fragmentation window

This figure shows the window in which oligomer fragmentations are performed. One or more fragmentation patterns might be selected in one fragmentation step.

Starting an oligomer fragmentation is a matter of having a polymer sequence opened in an editor window, selecting the sequence region to be fragmented and finally selecting the **Chemistry⇨Fragment** menu. The user is provided with a window where a number of fragmentation pathway names are listed (Figure 50 – Oligomer fragmentation window). As detailed for the cleavage of polymers, these fragmentation pathways are listed by looking into the polymer chemistry definition corresponding to the polymer sequence of which an oligomer is to be fragmented.

Select the fragmentation pathways(s) of interest, set the ionization range required for the generated fragment oligomers (the same as for polymer sequence cleavage) and click Fragment. Upon successful termination of the fragmentation reaction, the generated fragments are displayed in the Oligomers tableview widget.

As detailed for the cleavage of polymer sequences, the Details frame widget displays data about the fragments generated and the way masses were calculated for them.

It is possible to take into account cross-links that are beared by monomers contained in the oligomer.

← *XpertEdit: a Powerful Editor and Simulation Center*

XpertCalc: a Powerful Mass Calculator →

> **Warning**
>
> Only cross-links that are fully contained in the oligomer are taken into account. Partial cross-links, that is, cross-links involving at least one monomer outside of the oligomer, are ignored.

The cross-links are accounted for during the fragmentation of an oligomer if Cross-links is checked in the calculation engine configuration panel of the sequence editor window (Figure Figure 51 – Starting a fragmentation that accounts the cross-links). This is a partial view of the cyan fluorescent protein, with the "TWG" chromophore tripeptide.

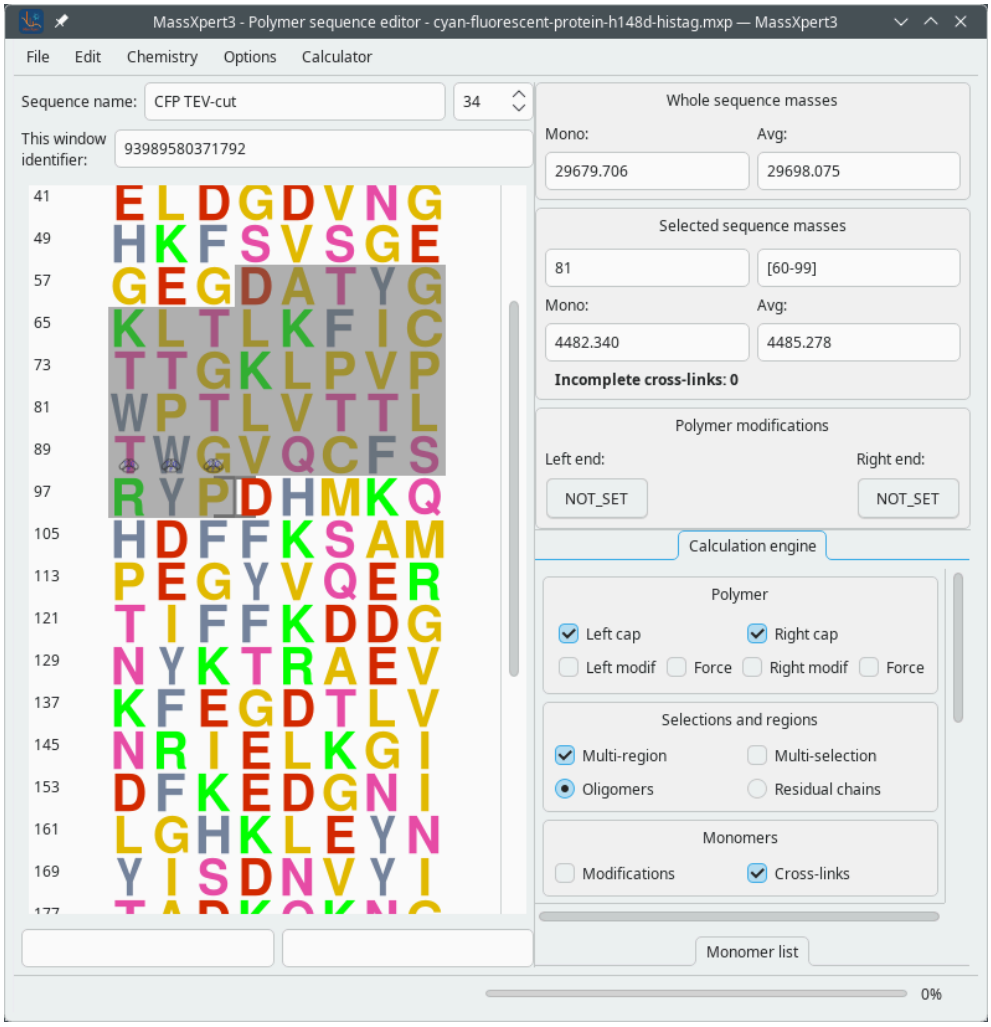


Figure 51: Starting a fragmentation that accounts the cross-links

The polymer sequence in the editor is selected making sure that all the cross-links are included in it (see the calculation engine's Cross-links check box). The calculation engine is configured to account for the cross-links.

If we select the oligomer region [60–99] and that we ask for a fragmentation, the fragmentation results will take into account the cross-link only in the case the generated fragments encompass fully the cross-link.

The following calculation rationale applies:

- Fragments b (left end) from b_1 (D) to b_{29} (D), sequence [^{60}DATY...QCFS^{88}], do not take into account the cross-link as all the monomers involved in it are outside of these be fragments;
- Fragments b_{29} to b_{31} do not take into account the cross-link because not all the monomer involved in it are contained in these fragments;
- Fragments b_{32} (up to C) to b_{40} do take into account the cross-link because all the monomers involved in it are contained in these fragments;
- Likewise, the only y fragments (right end) that do take into account the cross-link are the fragments y_9 and all the remaining, as for these fragments, the cross-link is fully contained since all the monomers involved in it are contained in these fragments.

A more complex cross-linked structure example is shown in Figure Figure 52 – Complicated cross-linking situation, where the 17-mer oligomer has 4 fully-encompassed cross-links and one partial. This oligomer is used as an example of how the fragmentation computation is handled by *MassXpert*.

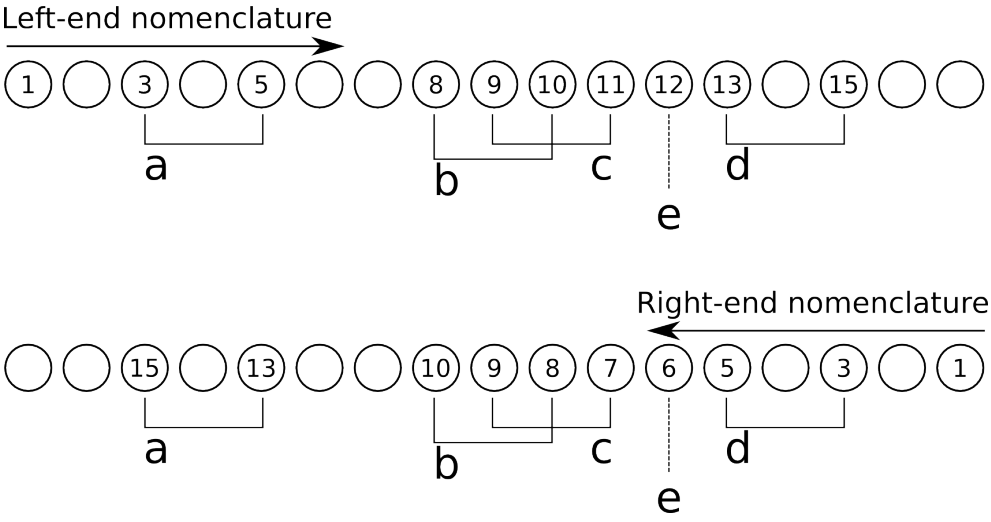


Figure 52: Complicated cross-linking situation

This figure shows a complicated cross-linking situation with an oligomer that has five cross-links, four of which are fully encompassed by the oligomer and one that involves a monomer outside of the oligomer.

The calculation of the fragments for this oligomer involves the following steps:

- Calculate regions of the oligomer that involve cross-links either overlapping or not. The regions are thus the following: [3–5], [8–11] and [13–15]. Note that the cross-link involving monomer 12 is never taken into account as it involves also a monomer outside of the oligomer;

- For fragments that have the left end of the oligomer (⟶ "Left-end nomenclature"), the following rationale is used:
  - Fragments ⟶1 and ⟶2 do not have any cross-link;
  - Fragments ⟶3 to ⟶4 do not account for cross-link *a* because that cross-link is not fully encompassed by the fragments;
  - Fragments ⟶5 to ⟶10 account only for the cross-link *a* as this is the only cross-linked region to be fully encompassed by these fragments;
  - Fragments ⟶11 to ⟶14 account for cross-links *a*, *b* and *c* as they are all fully encompassed in the fragments;
  - Fragments ⟶15 to ⟶16 account for all cross-links, a, b, c, d as they are all fully encompassed in the fragments;
- For fragments that have the right end of the oligomer (⟵ Right-end nomenclature), the following rationale is used:
  - Fragments 1⟵ and 2⟵ do not have any cross-link;
  - Fragments 3⟵ and 4⟵ do not account for cross-link *d* because that cross-link is not fully encompassed by the fragments;
  - Fragments 5⟵ and 6⟵ account for cross-link *d* because it is fully encompassed in these fragments;
  - Fragments 7⟵ to 9⟵ only account for cross-link *d* because cross-links *b* and *c* (which make one cross-linked region) are not fully encompassed by these fragments;
  - Fragments 10⟵ to 14⟵ account for cross-links *d*, *c* and *b*, but not for cross-link a as this last cross-link is not fully encompassed in these fragments;
  - Fragments 15⟵ and 16⟵ account for all the cross-links of the oligomer.

> **Warning**
>
> It is necessary to repeat one more time that cross-links that involve monomer(s) outside of the oligomer are ignored. The user is alerted whenever this situation is encountered.

The various widgets in the Actions groupbox widget are very similar to the ones found in the polymer sequence cleavage window (Figure 49 – Polymer sequence cleavage window).

For oligomer data filtering, please refer to Section 6.13 – Oligomer Data Filtering.

Exactly as described for the cleavage oligomer mass spectrum synthesis, it is possible to synthesize a mass spectrum corresponding to all the fragmentation oligomers currently selected in the table view widget. This process has been described at Mass spectrum synthesis.



Tip

In order to highlight inside of the sequence editor window an oligomer listed in the table view widget, double-click its corresponding row.

6.12 Mass Searching

It may happen that the scientist needs to know if some arbitrary sequence region would have a given mass. *MassXpert* allows for mass searching operations in the polymer sequence. This is done by using the menu **Chemistry⇨Mass Search**. The window in Figure 53 – Searching masses in a polymer sequence shows up and the user enters masses to search for. A number of parameters are to be detailed:

- Targets: should the masses be searched for in the whole sequence or in the currently selected region only?
- Ionization: should different levels of ionization be calculated when calculating masses for the potential oligomers matching the searched mass? For example, one finds in an electrospray ionization experiment mass spectrum a peak at *m/z* 1245. It is not possible to know the ionization level for that ion. On could imagine that this value is for a monoprotonated or for a multiprotonated species. If we wanted to assess this, we might ask that the mass be searched for by computing a range of possible ionization levels between Start level 1 and End level 4 (admitting that for that experiment this is what one would expect).

Once the masses have been searched for, if results are found they are displayed in the same window in the Oligomers table view widgets (the left one for the mono masses and the right one for the avg masses).

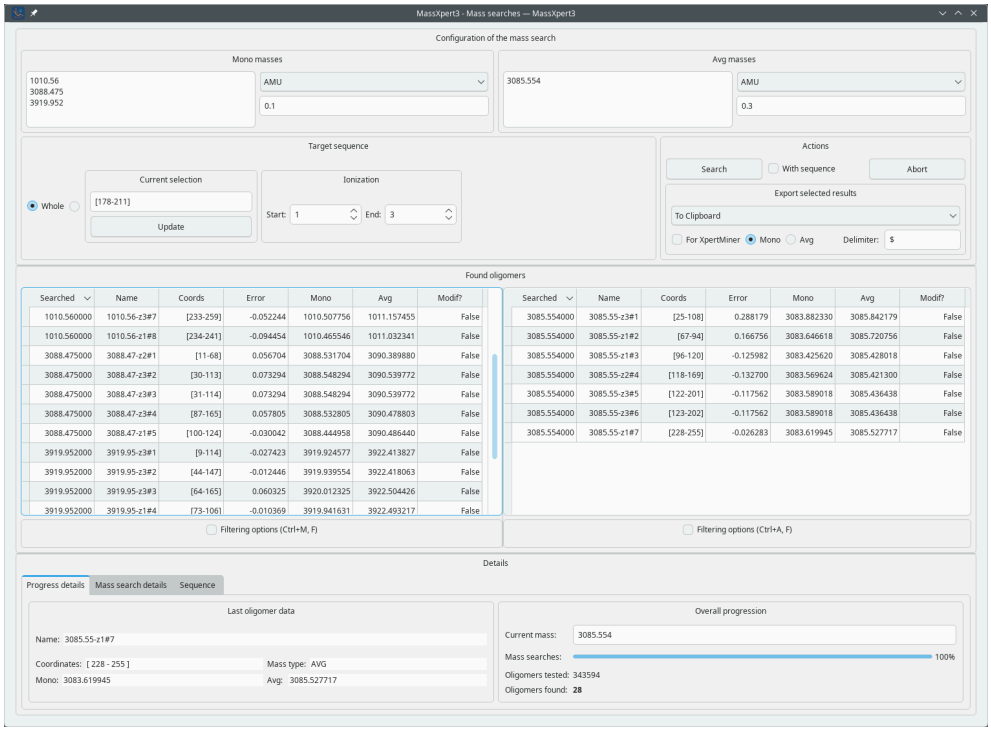


Figure 53: Searching masses in a polymer sequence

This figure shows the window in which to search for masses in a polymer sequence.

The various widgets in the Actions groupbox widget are very similar to the ones found in the polymer sequence cleavage window (Figure 49 – Polymer sequence cleavage window).

For oligomer data filtering, please refer to (Section 6.13 – Oligomer Data Filtering).

> **Tip**
>
> To highlight in the sequence editor window a given oligomer listed in the tableview widget, double-click its item.

6.13 Oligomer Data Filtering

Oligomer-generating simulations, like polymer sequence cleavages or fragmentations or mass searches, produce a very large amount of data. It is often desirable to be able to filter quickly some specific data out of these bunch of data... In all these three simulation contexts, the results that are displayed in the corresponding dialog windows are easily filtered using the mechanism illustrated in Figure 54 – Oligomer data filtering. To enter filtering mode, check Filtering options the line edit widgets will appear for you to start entering filters.

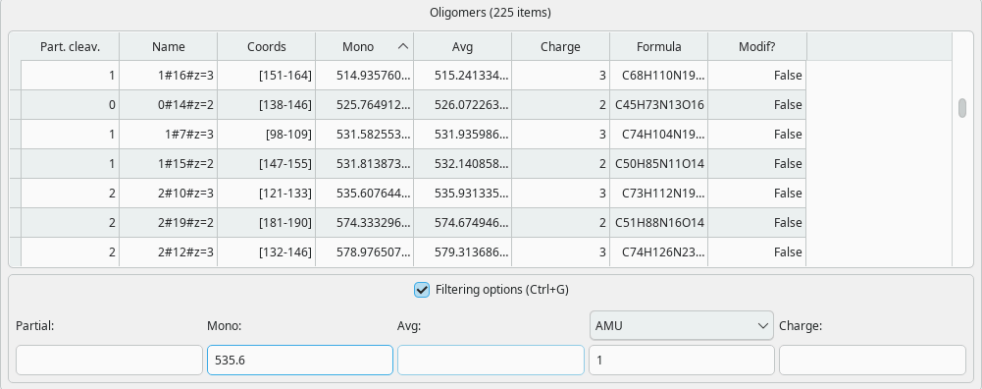


Figure 54: Oligomer data filtering

This figure shows how oligomer data can be filtered. The Filtering options group box contains four line edit widgets where filtering might be triggered: Partial, Mono, Avg, Charge. The filtered data are displayed in the same window (this example is for polymer sequence cleavage simulation data).

Filtering on the data is easily performed by entering the options in the Filtering options group box (Figure 54 – Oligomer data filtering). For any filtering operation, only one criterium can be used, that is, for example, filtering can occur only on the basis of the monoisotopic mass or of the average mass, but not on both masses.

For example, if one wanted to filter a huge set of data against a specific monoisotopic *m/z* of 850 +/− 0.33 u, it would simply be a matter of setting the monoisotopic *m/z* to be *850* with a tolerance of *3 AMU* in the corresponding line edit widgets contained in the Filtering options group box. To perform that filtering action, first set the tolerance value (*0.33*) in its line edit widget below the drop down list that reads AMU and next set the monoisotopic mass value to be *850* in the corresponding Mono line edit widget. While the cursor *is still* in the Mono line edit where *850* was entered, press the keyboard key combination `Ctrl` + `ENTER`.

The filtering will be immediate and the table view will show the data that passed the filter. Note that the combo box widget holding the unit of the tolerance (in the example, that unit is AMU, that is, “atomic mass unit”) and the line edit widget where the tolerance value proper is set (0.33 in the example) do not trigger any filtering by themselves; these widgets are only useful in conjunction with other oligomer data: Mono, Avg, Error line edit widgets (depending on the dialog window where the filtering occurs: cleavage, fragmentation or mass search). In our example, thus, the filtering would be spelled like this:

> — *“Only show the oligomers for which the monoisotopic mass is 850 plus or minus 0.33 atomic mass units”*

.

To exit the data filtering mode, uncheck Filtering options and all the initial data will be displayed, irrespective of any data in the line edit widgets described above.

6.14 Monomeric And Elemental Composition

The **Chemistry⇨Determine Compositions** menu triggers the window shown in Figure 55 – Determination of the compositions. The elemental composition is determined using the calculations engine configuration currently set in the polymer sequence editor window.

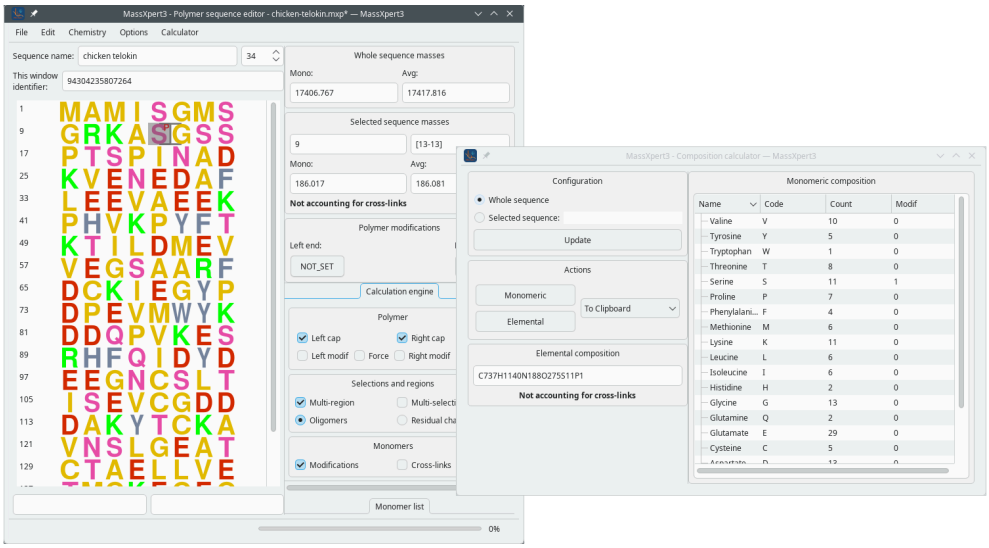


Figure 55: Determination of the compositions

This figure shows how to determine the monomeric and elemental compositions for the whole sequence or the current selection. As visible, the elemental composition formula comprises a phosphorus atom because the Modifications check box in the Monomers group box is checked in the polymer sequence editor window's Calculation engine.

6.15 pKa, pH, pI and Charges

When preparing biochemical experiments, very often users need to know how many charges a given polymer sequence will bear at any given pH. Equally important is the ability to know at which pH value the polymer sequence will have a net charge near to zero. The pH value for which a given polymer sequence has a net charge near to zero (typically this means that the number of positive charges equals the number of negative charges) is called the isoelectric point—the pI.

Such computations are pretty computer-intensive and require a very precise knowledge of the chemical structure of the different monomers that take part in the definition of the polymer chemistry. A file, called `pka_ph_pi.xml` is located in the polymer chemistry definition directory. This file lists all the chemical groups that are possibly charged; each

monomer of the polymer definition is represented by a "<monomer>" element in which data are defined for any chemical group of that monomer that might bear a charge at any given pH. You can find the listing of the `pka_ph_pi.xml` file in the appendices. We'll discuss any aspect of this file's contents in the next sections with enough detail that the user will be able to write one such file for her specific polymer chemistry.

At the moment, two entities in the polymer chemistry definition might have chemical groups bearing charges: monomers and modifications. We will first review monomers, and modifications next.

6.15.1 Ionized Group(s) In Monomers

Monomers are the building blocks of polymer sequences. These blocks must have at least two reactive groups so that they can be polymerized into a polymer sequence thread. Reactive groups are often chargeable groups; for example, the amino group of amino-acids is such that it gets protonated (positively charged) at a pH inferior to its pKa. Similarly, the carboxylic acid group of amino-acids is deprotonated (negatively charged) at physiological pH.

6.15.1.1 Some Theory First

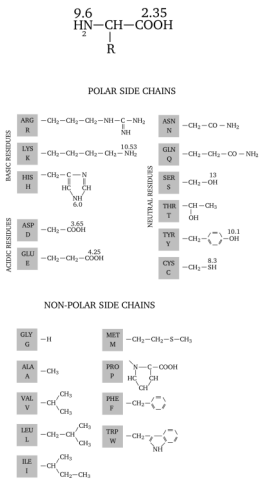


Figure 56: Different pKa values of chemical groups

All of the twenty amino-acids are represented here, which each amino-acid's lateral chain fully represented. Above each chemical group for which the value makes sense from a biological perspective the pKa value is indicated.

For the non-biochemist reader, amino-acids involved in the formation of proteins have always at least two chemical groups that are of inverted electrical charge, at physiological pH values (Figure 56 – Different pKa values of chemical groups).

← *XpertEdit: a Powerful Editor and Simulation Center*

XpertCalc: a Powerful Mass Calculator →

- The amino group (called αNH_2) has a typical pKa value of 9.6. This means that, at physiological pH values (between 6.5 and 7.5), the amino group will find the environment rather acidic, and will thus be protonated, leading to a positively-charged species (αNH_3^+);

- The carboxylic group (called $\alpha COOH$) has a typical pKa value of 2.35. This means that, at physiological pH values, the carboxylic group will find itself in a rather basic environment, and will thus be deprotonated, leading to a negatively-charged species (αCOO^-).

It should be clear that, at physiological pH values the two α-chemical groups have a naught net charge ($z = 0$). But proteins are charged, and this is because some of the twenty common amino-acids have other chemical groups beyond the two others already described. Indeed, some amino-acids have lateral chains that bear groups that might be charged depending on the pH: seryl residues have an alcohol group that has a pKa of 13, for example; that means that it is almost always uncharged (form ROH at physiological pH values). The lateral chain of lysine has a pKa of 10.53, which means that at pH values below this pKa value, the εNH_2 gets protonated, introducing a positive charge in the protein. Similarly, amino-acids glutamate and aspartate do have a lateral chain ending with a $\gamma COOH$ and a $\beta CCOOH$, respectively. Their pKa values are below 4.5, and thus the groups are negatively charged a physiological pH values.

When the net charge of a polymer sequence has to be computed for a given pH condition, the program iterates in the sequence, and for each monomer will check which one of its chemical group(s) is possibly charged. For this to happen, it is required that a number of data be known for each monomer's chemical group that might play a role in the determination of the polymer sequence's electrical charge. Thus, for each chemical group a number of data should be listed in the `pka_ph_pi.xml` file. In the *XML* file format code below, the examples of the protein's glutamate (left column) and lysine (right column) amino-acids are detailed in the next page.

```
<monomer>
  <code>E</code>
  <mnmchemgroup>
    <name>N-term NH2</name>
    <pka>9.6</pka>
    <acidcharged>TRUE</acidcharged>
    <polrule>left_trapped</polrule>
    <chemgrouprule>
      <entity>LE_PLM_MODIF</entity>
      <name>Acetylation</name>
      <outcome>LOST</outcome>
    </chemgrouprule>
  </mnmchemgroup>
  <mnmchemgroup>
    <name>C-term COOH</name>
    <pka>2.36</pka>
    <acidcharged>FALSE</acidcharged>
    <polrule>right_trapped</polrule>
  </mnmchemgroup>
  <mnmchemgroup>
    <name>Lateral COOH</name>
    <pka>4.25</pka>
    <acidcharged>FALSE</acidcharged>
    <polrule>never_trapped</polrule>
    <chemgrouprule>
      <entity>MONOMER_MODIF</entity>
      <name>AmidationGlu</name>
      <outcome>LOST</outcome>
    </chemgrouprule>
  </mnmchemgroup>
</monomer>
```

```
<monomer>
  <code>K</code>
  <mnmchemgroup>
    <name>N-term NH2</name>
    <pka>9.6</pka>
    <acidcharged>TRUE</acidcharged>
    <polrule>left_trapped</polrule>
    <chemgrouprule>
      <entity>LE_PLM_MODIF</entity>
      <name>Acetylation</name>
      <outcome>LOST</outcome>
    </chemgrouprule>
  </mnmchemgroup>
  <mnmchemgroup>
    <name>C-term COOH</name>
    <pka>2.36</pka>
    <acidcharged>FALSE</acidcharged>
    <polrule>right_trapped</polrule>
  </mnmchemgroup>
  <mnmchemgroup>
    <name>Lateral NH2</name>
    <pka>10.53</pka>
    <acidcharged>TRUE</acidcharged>
    <polrule>never_trapped</polrule>
    <chemgrouprule>
      <entity>MONOMER_MODIF</entity>
      <name>Acetylation</name>
      <outcome>LOST</outcome>
    </chemgrouprule>
  </mnmchemgroup>
</monomer>
```

- The chemical group's `<name>` element is required. For example, "αNH₂" or "εNH₂" or "αCOOH"

- The chemical group's `<pka>` element is optional, but is the basis for the charge calculation. For example, `9.6` for the "αNH₂" or `2.35` for "αCOOH"

- The `<acidcharged>` element is required if the `<pka>` element is given. This element is responsible for telling if the chemical group is charged (positively) when the pH is lower than pKa (that is when the medium is acidic with respect to the pKa). For example, an amine is positively charged when it is in its acidic form (protonated); a carboxylic acid is *not* charged when it is in its acidic form;

- There can be none, one, or more `<polrule>` element(s) for each chemgroup. The `<polrule>` element gives informations about the way the chemical group at hand might be "*trapped*" (or not) in the formation of inter-monomer bonds (while the monomer is polymerized into the polymer sequence). The value *left_trapped* means that the chemical group ceases to be involved in charge calculations as soon as it has a monomer at its left end. The value *right_trapped* means the same as above, but when a monomer is polymerized at its right end. For a chemical group that is *left_trapped*, we understand that it is only effectively evaluated if it is at the left end of the polymer sequence, since in this case it does not have a monomer at its left side. Conversely, a chemical group that has a `<polrule>` element with value *right_trapped*, will be evaluated only if the monomer is actually the right end monomer in the polymer sequence. Finally, the typical lateral

chains of amino-acids have a <polrule> element with a value *never_trapped*, as these chemical groups do not take part in the formation of the inter-monomer bond;

- There can be none, one or more <chemgrouprule> element(s) for each chemical group. Such an element should contain the following:
  - There must be an <entity> element that indicates what is the chemical entity being dealt with in the current `guilabel`"" element. Valid values for this element are *LE_PLM_MODIF*, *RE_PLM_MODIF* or *MNM_MODIF*;
  - There must be a <name> element naming the chemical entity properly;
  - There must be an <outcome> element telling what action should be taken when encountering the <entity> on the <chemgroup>. Valid values are either *LOST* or *PRESERVED*.

6.15.2 Ionized Group(s) In Modifications

In the excerpt from the `pka_ph_pi.xml` file below, we see that chemical modifications can also bring charges. The example of the chemical modification *Phosphorylation* shows that when a monomer is phosphorylated, two chemical groups are brought in: the first has a pKa value of 1.2 (that is it will always be deprotonated at physiological pH values), the second has a pKa value of 7 (that is it will be divided by half in a protonated (not charged) form and in an un-protonated (negatively charged) form, leading to a net electrical charge of −0.5.

```
<modif>
  <name>Phosphorylation</name>
  <mdfchemgroup>
    <name>none_set</name>
    <pka>1.2</pka>
    <acidcharged>FALSE</acidcharged>
  </mdfchemgroup>
  <mdfchemgroup>
    <name>none_set</name>
    <pka>6.5</pka>
    <acidcharged>FALSE</acidcharged>
  </mdfchemgroup>
</modif>
```

At this point we should be able to study the way computations are actually performed in the *XpertEdit* module.

6.15.3 pH, pI and Charge Calculations

The user willing to compute charges (positive, negative, net) or the isoelectric point for the current polymer sequence uses the menu **Chemistry⇨pKa pH pI** which triggers the appearance of the window shown in Figure 57 – Acido-basic computations: net charges.

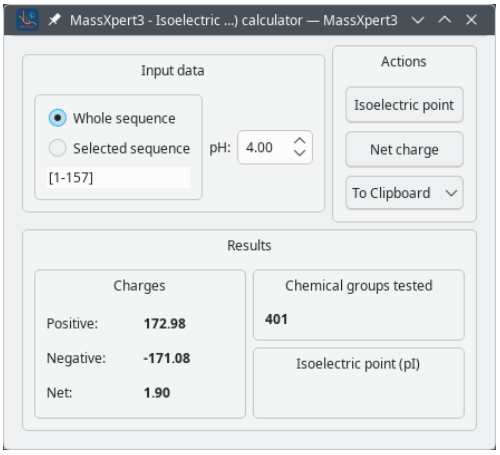


Figure 57: Acido-basic computations: net charges

This figure shows the options that can be set for the calculation of the charges beared by the polymer sequence.

This figure shows that the user can calculate the charges (positive, negative and net) beared by the polymer sequence (either the whole sequence or the current selection) by setting the pH value at which the computation should take place. It is also possible to calculate the isoelectric point by clicking onto the Isoelectric Point button.

Note that the computations might involve the permanent left/right modifications of the polymer sequence, as well as the monomer chemical modifications. To configure the way net charge—or pI—calculations are performed, use the calculation engine configuration of the sequence editor window.

6.16 General Options

One of the options that are valued most by users is to be able to set the number of decimal places used to diplay numbers. The settings should apply in a distinct manner depending on the different entities for which numerical values are to be displayed (see Section 3.1 – Setting the application preferences - general options).

7 *XpertCalc*: a Powerful Mass Calculator

After having completed this chapter you will be able to perform sophisticated polymer chemistry-aware mass calculations.

7.1 *XpertCalc* Invocation

The *XpertCalc* module is easily called by pulling down the ***XpertCalc*⇨Open calculator** menu item from the *MassXpert* program's menu. The user is presented with a window to select the polymer chemistry definition that should be used for the calculations (Figure 58 – Selecting a polymer chemistry definition for use with *XpertCalc*).

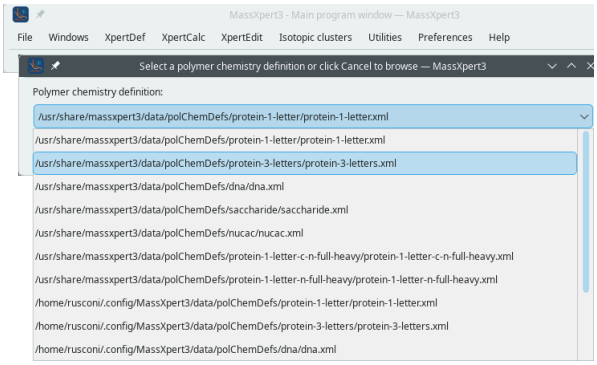


Figure 58: Selecting a polymer chemistry definition for use with *XpertCalc*

> Upon starting a molecular calculator, the user is asked to choose a polymer chemistry definition from a list of available defintitions.

7.2 An Easy Operation

Once the polymer chemistry definition has been correctly selected, it is parsed by the *XpertCalc* module and its entities are automatically made available in the calculator window. The way *XpertCalc* is operated is very easy. This is partly due to the very self-explanatory graphical user interface of the module, which is illustrated in Figure 59 – Interface of the *XpertCalc* module.

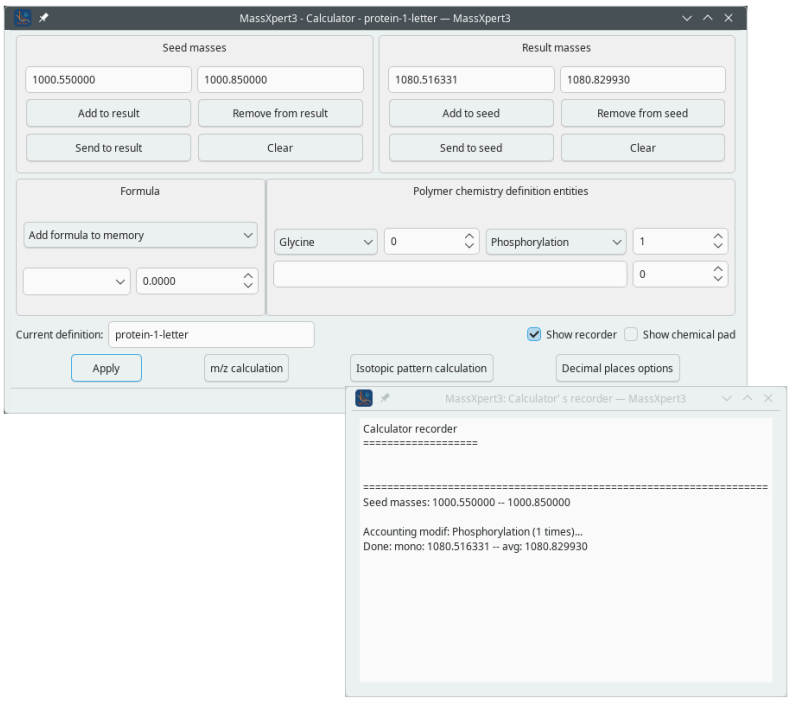


Figure 59: Interface of the *XpertCalc* module

This figure shows that the *XpertCalc* polymer chemistry definition-aware mass calculator can handle atoms, formulæ, monomers, modifications and even polymer sequences for computing masses.

XpertCalc can handle a number of features that are reviewed below:

- The user may (is not obliged to) seed the calculation by setting masses manually in the Seed masses line edit widgets (the left line edit is for mono and the right one for avg);

> **Warning**
>
> Both monoisotopic and average *m/z* values need to be entered.

For example, imagine that a mass spectrum analysis session ends up like this:

> — *"There is a peak at* m/z *= 1000.55,* z *= 1 and another one at roughly 80 Da more. Is it possible that the analyte showing up at* m/z *1000.55 is phopshorylated?"*

As shown in Figure 59 – Interface of the *XpertCalc* module, the mass spectrometry scientist would seed the mass calculator with mass *1000.55* and ask that one *Phosphorylation* modification be added to it by setting *1* in front of the corresponding drop-down widget. Clicking the Apply button triggers the calculation, with the resulting masses being displayed in the Result masses line edit widgets. We can see that the phosphorylation of our analyte shifts its *m/z* value from *1000.55* to *1080.5163*, which allows the

operator to evaluate the phosphorylation hypothesis by comparing the calculated value with the observed one.



Tip

Each time a calculation is triggered by clicking Apply (or the chemical pad's buttons; see below), the values already present in the Result masses line edit widgets are transferred to the Seed masses line edit widgets. This provides a 1-level undo;

- The Formula group box widget contains two widgets: a line edit widget where the formula is typed and a count spin box widget where the user sets the number of times that the formula should be applied. Setting the formula to *H2O* and the count to *2* would hydrate the analyte twice.

- The Polymer Chemistry Definition Entities group box widget contains two drop-down widgets and a line edit widget. The drop-down widget on the left lists all the monomers defined in the protein-1-letter polymer chemistry definition that was selected when starting the calculator; the drop-down widget on the right lists all the modifications defined in the protein-1-letter polymer chemistry definition. Each drop-down widget has its corresponding count spin box widget. In the example, the user asked that one (*1*) *Phosphorylation* modification be applied during the calculation. The line edit widget below the first row of widgets is the polymer sequence widget where the user might enter a sequence of monomers. It is possible to apply many times the sequence by setting the count spin box widget value to something greater than *1* (either positive or negative);

It is possible to perform a set of calculations in one go, that is, the user may ask for a formula, a monomer, a modification, a sequence to be accounted in one single calculation operation. Once all the chemical entities to be taken into account have been set, the user clicks the Apply button: all the entities are parsed in sequence and their mass equivalents are added to the result masses. Other prominent features of *XpertCalc* are described in the following sections.

7.3 The Programmable Calculator

For the scientists who work on molecules that are often modified in the same usual ways, *XpertCalc* features a built-in mechanism which enables them to easily program their calculator. This programming involves the definition of how a *chemical pad* (or *chempad*) may be arranged, exactly the same way as a desktop calculator would arrange its numerical keypad.

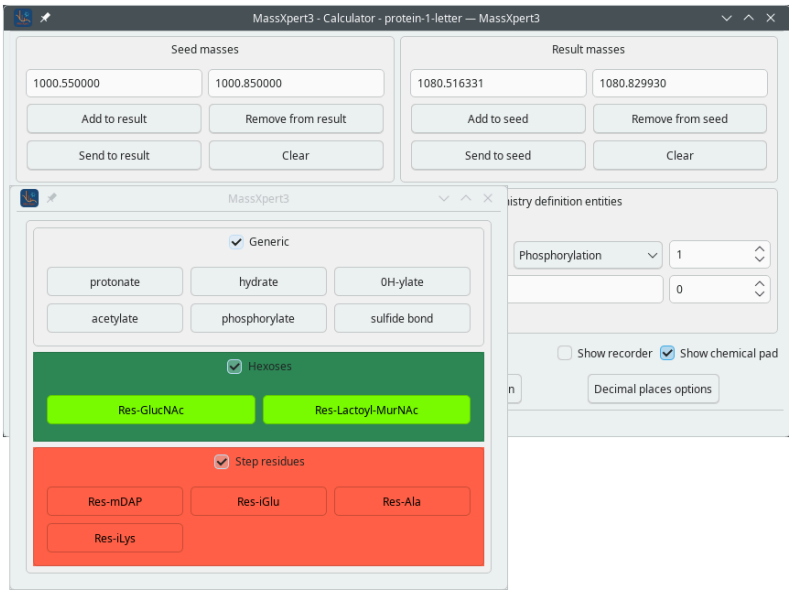


Figure 60: Interface of the chemical pad

This figure shows that the chemical pad is very similar to what a numerical calculator would display. Here, the user has programmed a number of chemical reactions.

The chemical pad can be shown/hidden by using the `Show Chemical Pad` check box. An example of such a chemical pad is shown in Figure 60 – Interface of the chemical pad, where a "protein-1-letter" polymer chemistry definition-associated chempad is featured. As shown, the user has programmed a number of chemical reactions that may be applied to the masses in the *XpertCalc* calculator window by simply clicking on their respective button. The configuration of the chemical pad is very easy.



Tip

It is recommended to copy one of the `chemPad.conf` configuration files in any of the polymer chemistry definitions distributed within *MassXpert* and to modify it according to the instructions at the top of the file.

One example of such a configuration file is shown below along with explanations:

```
color%aliceblue%240,248,255
color%antiquewhite%250,235,215
color%aqua%0,255,255
chempad_columns%3
chempadgroup%Generic
chempadkey=protonate%+H1%adds a proton
chempadkey=hydrate%+H2O1%adds a water molecule
chempadkey=OH-ylate%+O1H1%adds an hydroxyl group
chempadkey=acetylate%-H1+C2H3O1%adds an acetyl group
chempadkey=phosphorylate%-H+H2PO3%add a phosphate
group
chempadkey=sulfide bond%-H2%oxydizes with loss of hy-
drogen
chempadgroup%Hexoses%[seagreen]
chempadkey%Res-1bRE-Hexose%C6H11O6%residue Hexose
(1bRE)%[lawngreen,black]
chempadkey%Res-1bRE-Hexalditol%C6H12O6%residue Hexa-
lditol (1bRE-ol)%[lawngreen,black]
```

The first three lines of the code above are color definitions. One line specifies the arrangement of the keys on the chemical pad. The remaining lines are definitions for groups of keys and the keys themselves. Below are the feature descriptions:

- It is possible to define as many colors as necessary according to the standard red, blue, green format, on a scale of 0–255).
- The calculator chemical pad shall have its buttons organized in three columns.
- This separator will create a group box widget labelled "Generic" that will be populated with all the items found below (until another separator is encountered).
- A button definition is introduced by the *chempadkey=* string. Separated by % characters, follow the name of the chemical reaction that will label the button (*protonate*), the chemical formula of the reaction (*+H2O1*) and finally the tooltip text that displays when the cursor stays on the button.
- *chempagroup* is a separator that starts a new button group box, labelled *Hexoses* in the example. This syntax allows for the coloring of the group box widget (*seagreen* in the example).
- The last two lines describe buttons by also specifying their color. *lawngreen* is the background color and *black* is the color of the text.

These buttons might be used in two distinct ways:

- Upon clicking the button, its formula is evaluated and the corresponding masses are added to (or subtracted from) the Result masses
- If the Ctrl key is maintained pressed upon clicking on a chemical pad button, its formula is inserted into the Formula line edit widget of the calculator, so that the formula is not directly evaluated and the Result masses are not modified.

> **Tip**
>
> The Formula line edit widget has a drop down menu above it that enables the user to store the formula in memory, to recall it or to erase it.

7.4 The Log Book Recorder

Each time an action that is chemically relevant—from a molecular mass point of view—is performed, the program dumps the calculations to the *XpertCalc* recorder window (Figure 61 – The *XpertCalc* recorder window). The recorder can be shown/hidden by using the Show Recorder check box widget. The text in the recorder window is editable for the user to edit the *XpertCalc* output, and selectable also, so that pasting to text editors or word processors is easy *via* the clipboard.

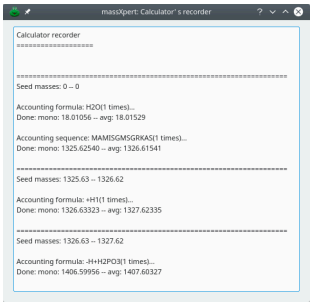


Figure 61: The *XpertCalc* recorder window

This figure shows that the recorder window is a simple text edit widget that records all the mass-significant operations in the *XpertCalc* calculator. The text in the recorder may be selected and later used in an electronic logbook or printed.

7.5 The *m/z* Ratio Calculator

It very often happens that the mass spectrometrist doing electrospray analyzes is faced with a challenging task: to compute by mind all the *m/z* ratios for a given family of charge peaks. To ease that daunting task, *XpertCalc* contains a *m/z* ratio calculator that is called

by clicking onto the m/z calculation button.

7.6 *m/z* Ratio Calculations

In electrospray ionization, a given polymer sequence might be charged a large number of times. The tool shown in Figure 62 – Calculation of ranges of m/z ratios details how to compute a range of *m/z* ratios starting from one *m/z* value for a given charge and a given ionization agent. It is also possible to switch ionization agent on-the-fly.

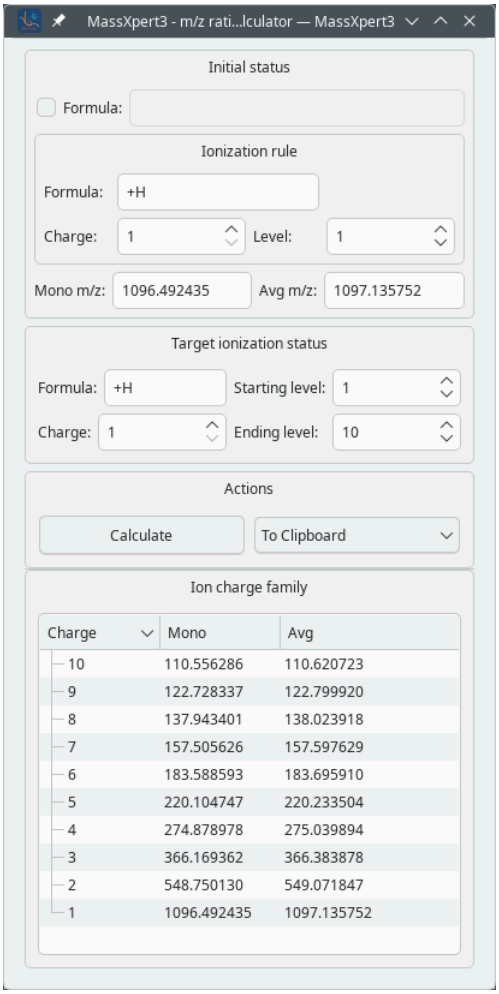


Figure 62: Calculation of ranges of m/z ratios

This figure shows the window in which to perform the calculation of different *m/z* ratios starting from one *m/z* value with a given ionization rule.

The sequence with which data should be entered in the upper part of the window is the following:

- Ensure that the Ionization rule is correct;

- Enter the Mono *m/z* and or the Avg *m/z* values
- Set the Target ionization status
- Click the Calculate button.

8 Isotopic cluster calculations

When the resolving power of the mass spectrometer is good, zooming-in on a mass peak may reveal that a given ion has given rise not to one peak but to a set of peaks that together form an "*isotopic cluster*". The ions responsible for the different cluster peaks are called "*isotopologues*" because they are homologues of the analyte that have different numbers (and therefore positions) of the different isotopes.

While the richness of mass spectra acquired with high resolving power machines has been neglected for many years, it is now well established that using the resolved isotopic cluster peaks for the detailed analysis of mass spectra is of great use.

When performing mass spectrometric data analyses, it is very often necessary to calculate the isotopic cluster of a given analyte. Indeed, the isotopic cluster can be a criterion helping to establish if a given detected signal (the acquired isotopic cluster) does indeed correspond to a given chemical entity, and matching the experimental isotopic cluster with one that was simulated for the chemical entity under scrutiny might prove useful. This way of ascertaining the assignment of mass peaks to specific analytes is becoming more and more desirable as mass spectrometers have ever-greater resolving power capabilities, thus providing the user with neatly resolved isotopic clusters even for oligomers of intermediate masses. In this section, the process of calculating the isotopic cluster of a given analyte defined by its chemical composition (also called elemental composition or formula) and optionally its charge is described in detail.

It is possible to predict how a given ion (of known elemental composition *and* charge) is supposed to be revealed as an isotopic cluster in a high resolution mass spectrum. One such cluster is shown in Figure 63 – Simulation of the isotopic cluster for a protein, for a protein in its deca-protonated form $[M+10H^+]^{10+}$, of ionic elemental composition $C_{769}H_{1222}N_{210}O_{218}S_2$.



Figure 63: Simulation of the isotopic cluster for a protein

Calculated isotopic cluster for a deca-protonated protein.

MassXpert provides a sophisticated interface to the `libIsoSpec++` library.

> IsoSpec: Hyperfast Fine Structure Calculator
>
> Mateusz K. Łącki, Michał Startek, Dirk Valkenborg, and Anna Gambin
>
> *Analytical Chemistry*, 2017, 89, 3272–3277
>
> DOI: https://doi.org/10.1021/acs.analchem.6b01459

This library performs high-resolution isotopic cluster calculations. In order to run the calculations, it is necessary to have the following items ready:

- The elemental composition of the analyte (for example, C769H1222N210O218S2). This formula needs to account for the ionization agent that is involved in the ionization of the analyte prior to its detection in the mass spectrometer. If an analyte is deca-protonated, 10 protons need to be added to its neutral-state formula.

- A detailed isotopic configuration of all the chemical elements that are found in the elemental composition. *MassXpert* provides three interfaces to define the isotopic characteristics of the chemical elements. These will be described in the following sections.

8.1 Generating an isotopic cluster simulation in *MassXpert*

To perform an isotopic cluster simulation in *MassXpert*, select the **Isotopic clusters⇨ Isotopic cluster generator** menu (Figure 64 – Isotopic clusters main menu).

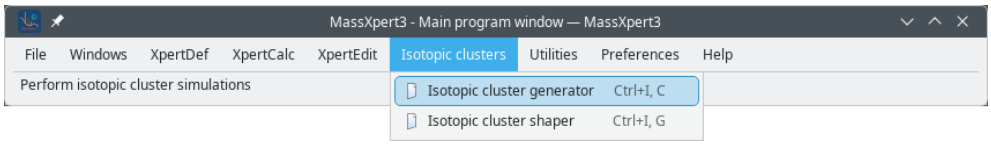


Figure 64: Isotopic clusters main menu

Two menu items are available, one to compute the isotopic cluster centroids starting from a formula and a charge (if any) and one to shape the mass peaks around the centroids, so that their profile matches the expected mass spectrum acquired on an instrument.

The window shown in Figure 65 – Isotopic cluster generator window contains all the interface elements required to configure the simulation starting from a formula.

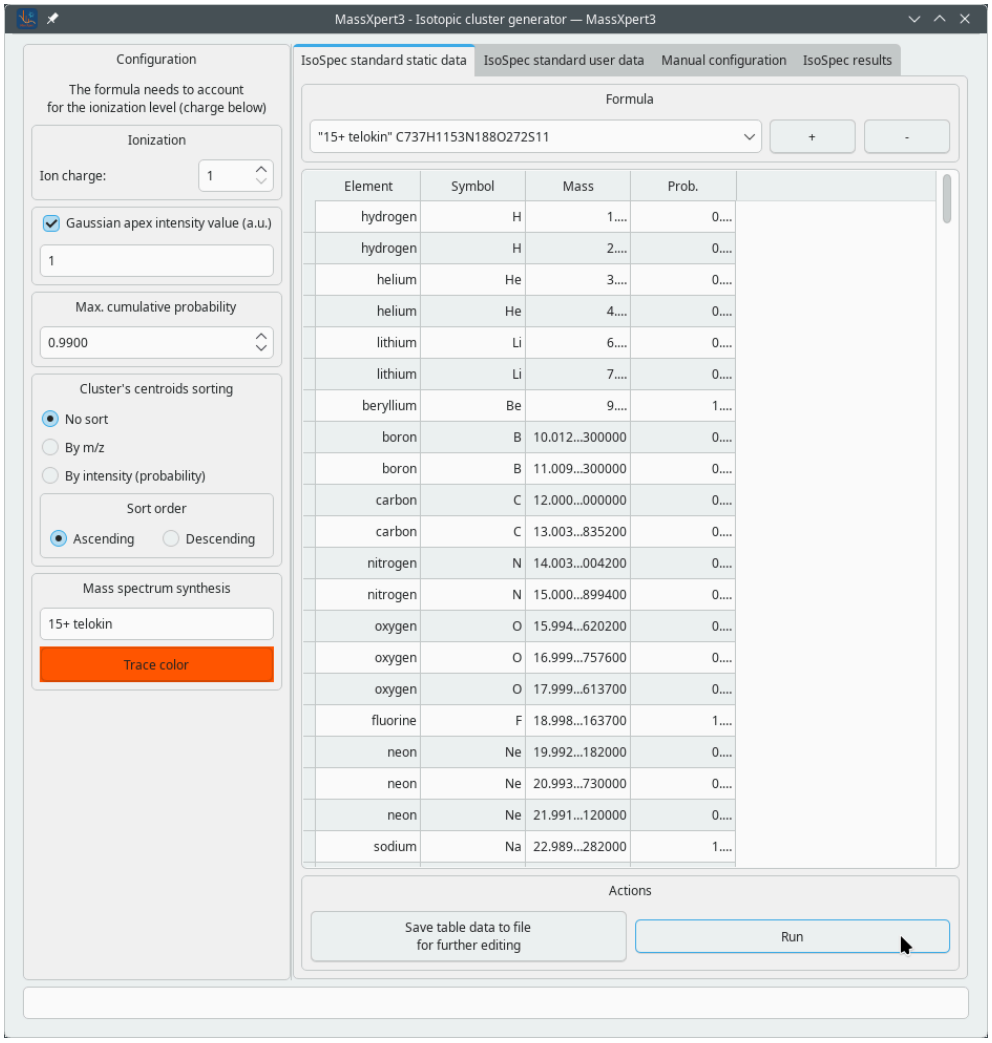


Figure 65: Isotopic cluster generator window

The window contains two panels. The left hand side panel configures the charge for which the calculation is to be carried over, the intensity of the highest peak in the isotopic cluster, and the maximum cumulative isotopic presence probability that *IsoSpec* must reach during the calculation. The right hand side panel contains a tab widget to hold the configuration tabs and the results tab.

An isotopic cluster calculation is most probably performed with the aim of simulating an expected isotopic cluster for an analyte that is being analyzed by mass spectrometry. It is thus logical that the analyte be in an ionized form. The way that the analyte has been ionized needs to be taken into account in the chemical formula that describes the ion for which the isotopic cluster is being calculated. For example, when determining the elemental composition of a protein in the positive ion mode, the number of protons used to ionize the protein needs to be taken into account in the formula.

> **Warning**
>
> The calculation is "charging agent-agnostic" in the sense that it does not know what chemical group in the chemical formula is re-

sponsible for the ionization of the analyte (if any). The ionization *level* of the analyte can be handled by *MassXpert* if that information is set to the Ion charge spin box widget. By default, the charge state of the analyte is *1*.

The Gaussian apex intensity value (a.u.) check box widget allows the user to set the intensity (y axis) that the isotopic cluster will have for its highest peak. This is useful when the isotopic cluster is to be simulated for direct comparison with an isotopic cluster in an experimental mass spectrum.

The Max. cumulative probability spin box widget serves to configure the extent to which *IsoSpec* simulates the theoretically expected isotopic cluster. A value of *0.99* tells the software to simulate enough combinations of the analyte isotopes to represent 99 % of the theoretically expected combinations.



Tip

For large biopolymers, it might be wise to start with a relatively low value for Max. cumulative probability, because setting this value too high, that is, near 1, would notably increase the calculation duration.

The Cluster's centroid sorting group box widget contains a few settings to configure the way the generated cluster's (centroid, intensity) pairs are sorted in the output.

The Mass spectrum synthesis group box widget contains a few settings to configure the title and the color of the generated spectrum (the actual generation of the spectrum on the basis of the *IsoSpec* output is described in detail below). The title and color of the mass spectrum are used when the spectrum is displayed in the companion *MineXpert* software program. The process by which the mass spectrum is transferred from *MassXpert* to *MineXpert* is described at Section 9.1 – Configuration of the *MassXpert* and *MineXpert* intercommunications.

To perform isotopic cluster calculations, the simulation software needs to be aware of the whole set of isotopes of all of the chemical elements that enter in the composition of the ionized analyte. An isotope is defined by its mass and by the probability that it is found in nature. Carbon has two major isotopes that can be found in nature: the ^{12}C most abundant isotope and the ^{13}C least abundant isotope (the ^{14}C isotope is irrelevant for conventional mass spectrometry because its abundance in nature is way too low).

There are three ways that the user might define the isotopic data to be used for the simulations. Two of them involve a configuration preparation on the part of the user. The

third one, that we'll describe first, does not involve any chemical element configuration. The other ways are reviewed in the following sections.

8.1.1 Static Standard Element Tables Shipped within the *IsoSpec* Library

In order to document all the chemical elements' isotopes' characteristics, the *IsoSpec* library has, in its own software code headers, a number of arrays that *MassXpert* automatically loads up when opening the Figure 65 – Isotopic cluster generator window dialog window.[9] These data are displayed in the IsoSpec standard static data tab. The table view widget data are loaded from source code each time the window is opened, hence the *static* qualifier in the tab name, meaning that the user won't be able to change these data and keep the changes between sessions.

In this tab, apart the configuration on the left hand side pane of the dialog window, all that is left to do is enter in the Formula combo box widget (which looks like a line edit widget with an arrow at its right side) the elemental composition for which the isotopic cluster calculation is to be performed.



Tip

If a formula is used often, the user might want to store it with a label that explains what the elemental composition describes. To add a label to the formula, enter the label between quotes before the formula itself. To store the "titled" formula so that it will be available across sessions, click the + push button. To remove a stored formula, select it in the combo box widget and click the – push button.

Once the formula has been set to its combo box widget, the charge of the ion needs to be set (Ion charge spin box widget) along with the Max. cumulative probability. To run the calculation, click Run.

8.1.2 User-Modified Standard Element Table Shipped within the *IsoSpec* Library

While the previous section showed how to use the static element table from the *IsoSpec* library, this section shows a way to customize these elemental data for a specific experiment.

[9] That process is performed at software build time and is thus not configurable at run time, which is why I refer to this dataset as *static*.

The user-modified version of the standard static isotopic data is available in the IsoSpec standard user data tab of the dialog window, which, upon opening has an empty table view widget, as shown in Figure 66 – User-modified standard element table.

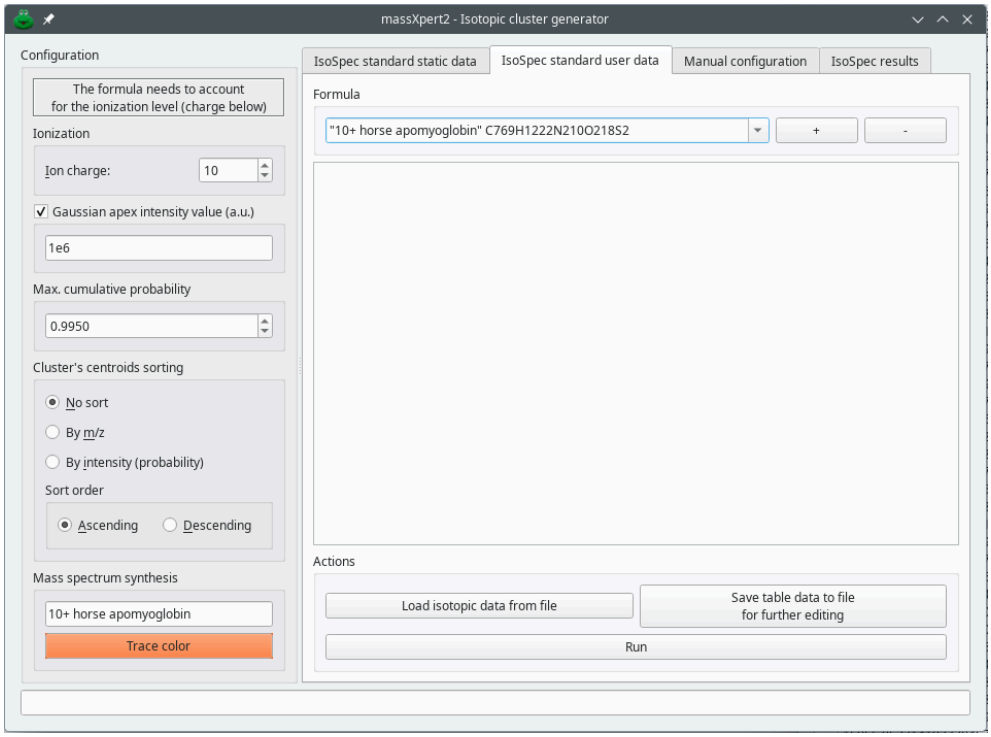


Figure 66: User-modified standard element table

When started, the IsoSpec standard user data tab of the dialog window shows an empty table view widget. The way to populate the table view is described in the text.

The simplest way to fill in the user-modified standard isotopic table is to export the isotopic data from the IsoSpec standard static data tab to a text file (Save table data to file for further editing button). The data are output to the file in the *CSV* (comma-separated-values) format, in a layout that closely mimicks the data visible in the table view. An example of the exported text is shown below after modifications applied to the carbon and nitrogen isotopes for heavy labeling. One would typically modify the isotope definition file with a simple text editor.

Isotope definition for full labelling with heavy C and N isotopes

```
carbon,C,12.000000,0.007000000
carbon,C,13.0033548,0.992999999
nitrogen,N,14.003074004200,0.00700000
nitrogen,N,15.0001088994,0.992999999
```

Now that the modifications have been performed, the file needs to be saved under the same format. It is then available for loading in the IsoSpec standard user data tab of the

dialog window by clicking the Load isotopic data from file push button. The table view now shows the new isotope mass-abundance pairs. From now on, this tab may be used exactly like described above for the standard isospec *static* data-based work.

> **Tip**
>
> The cells of the table view widget are editable. The cells' contents may be edited by double-clicking the cell and entering the new value. The table view widget contents might then be exported to a file exactly as described above.

Once the formula has been set to its combo box widget, the charge of the ion needs to be set (Ion charge spin box widget) along with the Max. cumulative probability. To run the calculation, click the Run button.

8.1.3 User hand-made isotopic configuration of an elemental composition

The last way the user might configure the chemical elements to be used for an isotopic cluster calculation is based on the fully manual description of the isotopes present in the elemental composition of interest. That configuration is performed in the Manual configuration tab of the dialog window. This method is slightly more involved than the previous one but provides also for a much greater flexibility: it allows one to create "new chemical elements" that might be required in specific labelling experiments. The manual configuration is carried over in the Manual configuration tab of the dialog. Upon starting, this tab is empty, as shown in Figure 67 – Hand-made user isotopic configuration of an elemental composition (the working sheet is blank).

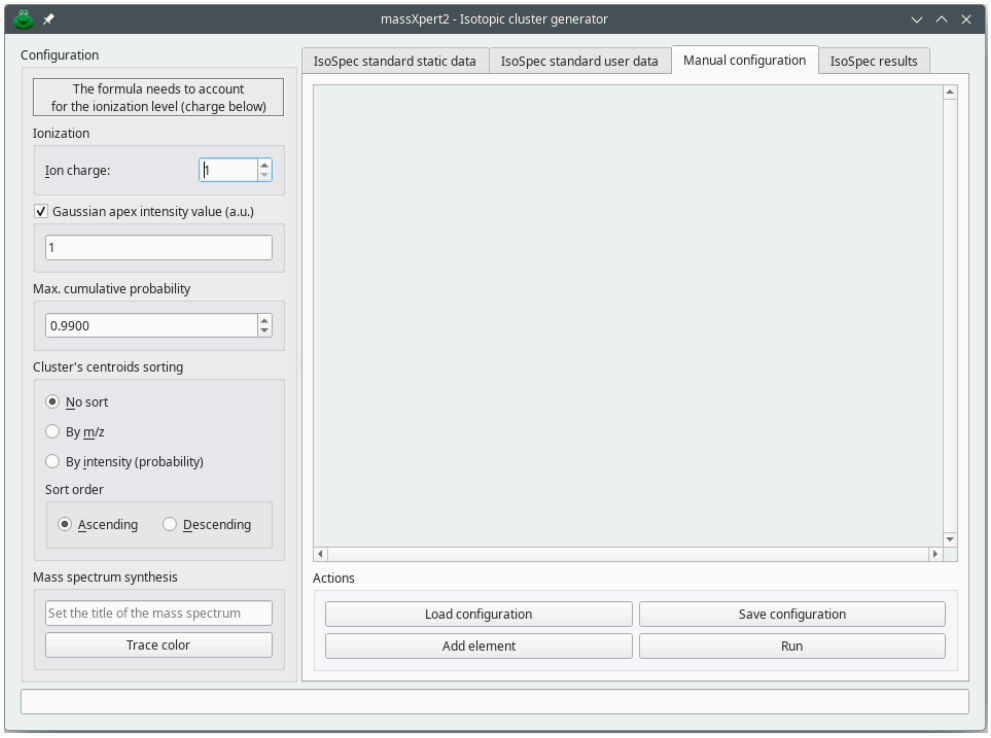


Figure 67: Hand-made user isotopic configuration of an elemental composition

When the window is created, the tab is empty. To start creating element definitions, click Add element.

Upon creation of the window, the Manual configuration tab is empty, with only two rows of buttons at the bottom of the tab. To start configuring chemical elements, click Add element to create an "element group box" that will contains a number of widgets organized in two rows, as described in Figure 68 – Hand-made user isotopic configuration of an elemental composition.

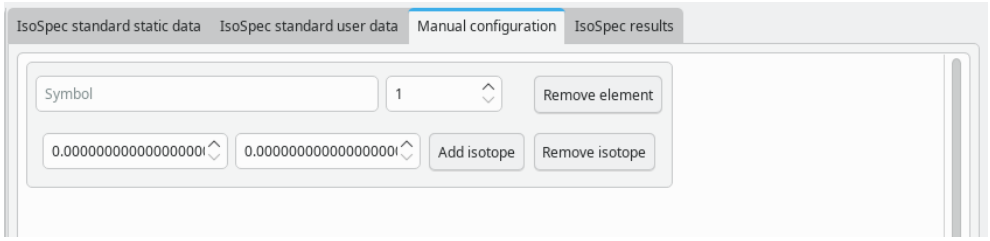


Figure 68: Hand-made user isotopic configuration of an elemental composition

Upon clicking the Add element button a new pristine element is created.

- Top row, a line edit widget to receive the chemical element symbol, like C, for example;
- On the same row, a spin box widget in which to set the number of such atoms in the formula for which the isotopic cluster is being calculated;
- The Remove element removes all the "element group box" in one go;

- The bottom row contains a set of widgets to configure the first isotope of the current element. The left spin box widget is for the mass and the right one is for the abundance;
- To add another isotope for the current chemical element, click the Add isotope button. To remove that isotope row, click the Remove isotope button. There is no limitation on the number of isotopes that may be created to fully qualify a chemical element.

> **Note**
>
> It is not possible to remove all the isotope frames from an element group box, otherwise that element group box would become useless.

Once all the isotopes for a given chemical element have been defined, a new element might be needed. For this, click Add element and start the configuration of the new element as described above.

The manual isotopic configuration of the chemical elements required to perform an isotopic cluster calculation for a given formula is tedious. The user may want to save a given configuration to a file (click the Save configuration button) so that it is easier to recreate automatically all the widgets upon loading of that saved configuration (click Load configuration).

The final configuration is shown in Figure 69 – Typical manual configuration of the isotopic composition of the chemical elements. The experiment that was configured above is the labelling of a glucose molecule with Cz, an imaginary chemical element that is like carbon but that has a ^{14}C. The glucose molecule (normal formula: C6H12O6) is labelled on one single carbon atom with an efficieny of 95 %. This means that, when the labelling fails (in 5 % of the cases) the carbon atom has its isotopes with usual probabilities (compounded by the fact that the unlabeled atom is found at that position only in 5 % of the cases). The isotopic abundances for the Cz element are thus:

- For ^{12}C: 0.05 * natural ^{12}C abundance;
- For ^{13}C: 0.05 * natural ^{13}C abundance;
- For ^{14}C: 0.95;

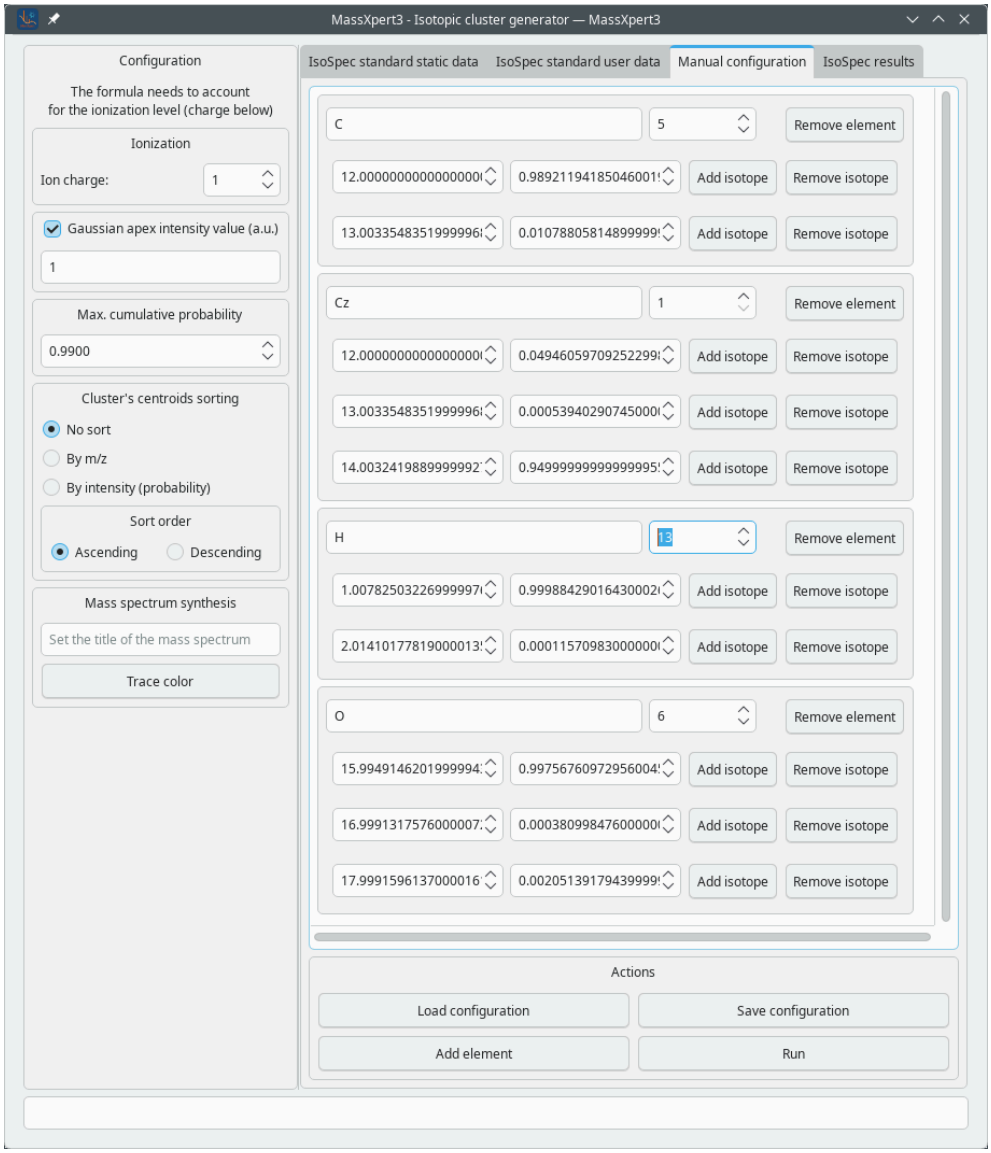


Figure 69: Typical manual configuration of the isotopic composition of the chemical elements

The user has configured a labelling experiment where a protonated glucose molecule (13 H atoms) is labelled at a single carbon position with a ^{14}C atom (the efficiency of the labeling is 95 %).

> **Note**
>
> Note that the count of unlabeled carbon atoms is 5 (and not 6), that the hydrogen count is 13 (and not 12, because the glucose is protonated), and that the ^{14}C-labelled carbon atom is present only once.

The example of has been exported and is available as a text file depicted in Figure 70 – The user manual configuration might be exported to a text file and loaded back.

```
# This file contains isotopic data in a format that can accomodate
# comments in the form of lines beginning with the '#' character.
[Element]
        symbol C count 5
        [Isotopes] 2
                mass 12.0 prob 0.98921194185046
                mass 13.0033548352 prob 0.010788058149
[Element]
        symbol Cz count 1
        [Isotopes] 3
                # This position is 5% unoccupied by radioactive C14, thus, its
                # position has natural abundances for C12 and C13 compounded by 0.05.
                # C12: prob = 0.98921194185046 * 0.05
                mass 12.0 prob 0.049460597092523
                # C13: prob = 0.010788058149 * 0.05
                mass 13.0033548352 prob 0.000539402907450
                # This one is fixed at 0.95 because the vendor says incorporation of
                # C14 is 95%.
                mass 14.003241989 prob 0.95
[Element]
        symbol H count 13
        [Isotopes] 2
                mass 1.00782503227 prob 0.9998842901643
                mass 2.01410177819 prob 0.00011570983
[Element]
        symbol O count 6
        [Isotopes] 3
                mass 15.9949146202 prob 0.99756760972956
                mass 16.9991317576 prob 0.00038099847600
                mass 17.9991596137 prob 0.0020513917944
```

Figure 70: The user manual configuration might be exported to a text file and loaded back

Use the Save configuration push button to export the configuration to a text file and the Load configuration push button to load a configuration from file.

As described above, once the analyte's elemental composition has been defined, as just described, click Run.

8.1.4 The results of the computation are centroids

Once the configurations have been terminated, the isotopic cluster calculations can finally be performed. In the *manual configuration* setting, the formula is automatically handled, since each chemical element that is defined goes along with the count of the corresponding atoms. In the case of the standard *IsoSpec* configuration (either modified or not), the user has to enter the chemical formula of the analyte in the Formula line edit widget.

Click the Run button. If the configuration was correct and the calculation could run properly, then the window switches to the IsoSpec results tab (Figure 71 – Results from the isotopic cluster calculation). That tab contains a text edit widget in which the results are displayed.



Tip

IsoSpec does not know anything about ion charge. The results are thus "corrected" by *MassXpert* for the charge level that was specified in the left panel of the dialog window prior to their display in the results tab (Figure 71 – Results from the isotopic cluster calculation).

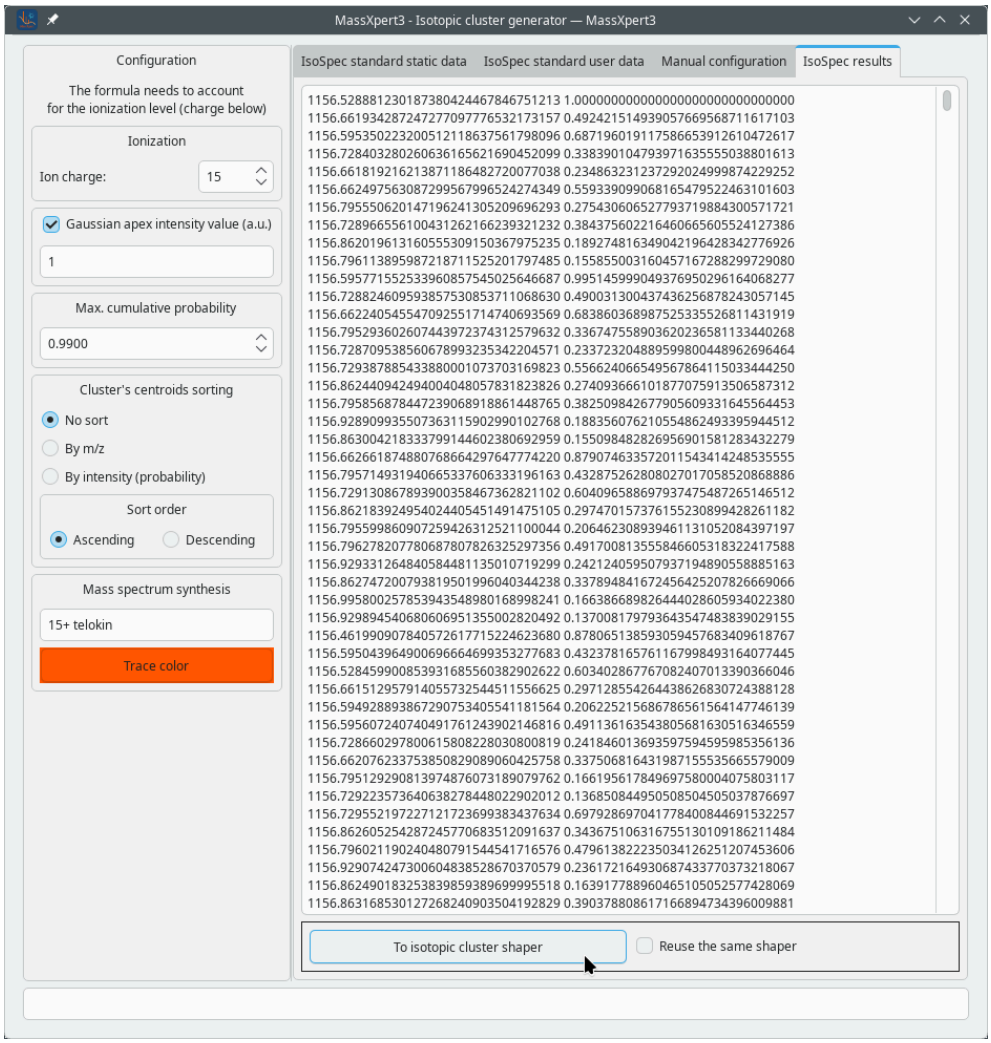


Figure 71: Results from the isotopic cluster calculation

The computation generates the probability of the various combinations of all the isotopes that make the elemental composition submitted to it. The results are in the form of peak centroid values along with corresponding probabilities. The sum of the probabilities corresponds to the Max. cumulative probability value that was set by the user.

The results output by *IsoSpec* represent the centroids of the peaks that make the isotopic cluster. The results are thus a set of (m/z, i) pairs that have not the characteristic shape (the profile) that is found in mass spectra. *MassXpert* features the ability to give a shape to the centroid peaks. For that, click the To isotopic cluster shaper button to open the Isotopic cluster shaper window preloaded with the list of peak centroids.

← *Isotopic cluster calculations*

Utilities →

> **Note**
>
> The Isotopic cluster shaper window's features might be used by themselves, not necessarily as the last step of an isotopic cluster simulation. To shape any peak starting from a centroid value and a corresponding intensity value, select the **Utilities⇨Isotopic cluster shaper** menu item. The left hand side Data centroid points (format: mzi) text edit widget appears empty. Paste the centroid data into it, one centroid per line, with the *m/z* value separated from the intensity by a space. Make sure to register the new centroid list by clicking the Import from text push button.

The workings of this peak shaping feature is described in Section 8.2 – Shaping the centroids of the isotopic cluster peaks into well-profiled peaks.

8.2 Shaping the centroids of the isotopic cluster peaks into well-profiled peaks

As mentioned earlier, the isotopic cluster calculation produces the isotopic cluster in the form of a set of (*m/z*,i) pairs that represent the centroids of the isotopic cluster's peaks and their relative intensity (probability). There are other data simulations or analysis processes that lead to having mass peaks defined by a single centroid *m/z* value and a corresponding intensity. Plotted to a graph, a centroid mass peak yields a bar, which is inadequate if the user is willing to simulate isotopic clusters to later compare these with experimental profiled mass peaks. The features in this dialog window aim at shaping each mass peak by creating a shape around the centroid (in specialized terms, this is the profile of the peak).

The shape of mass peaks is typically Gaussian or Lorentzian (or a mix thereof). In order to convert mass peak centroids into something that resembles a real "profile" mass peak, a mathematical formula can be applied, with some parameters to configure the shapes generated. *MassXpert* makes the shaping of mass peaks accessible *via* the **Isotopic clusters⇨Isotopic cluster shaper** menu item (Figure 64 – Isotopic clusters main menu). The window that opens up is shown in Figure 72 – Setting-up of the centroid mass peak shaping process.

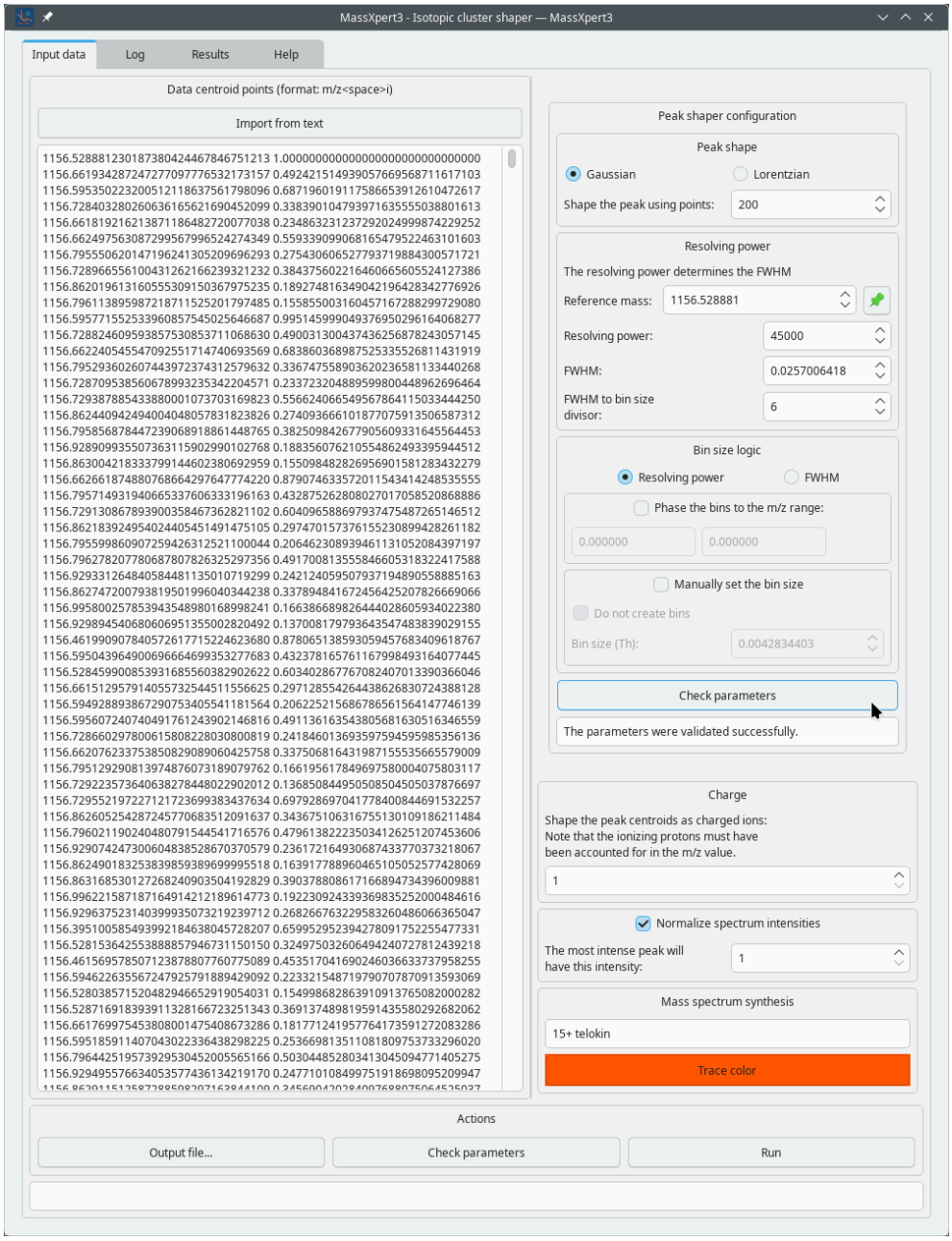


Figure 72: Setting-up of the centroid mass peak shaping process

This dialog window allows one to configure the shaping of mass centroid peaks. Setting a spectrum name in the Mass spectrum name line edit widget and choosing a plotting color will help recognize the result mass spectrum once displayed in the Mass spectra window in the *MineXpert* companion software (see below).

The mass centroid peaks are listed in the Data centroid points (m/z,i) text edit widget. These values may come from different origins. They can be set automatically from the isotopic cluster calculation dialog window (see Section 8.1.4 – The results of the computation are centroids); they can come from the user editing them in place; finally they can come from the user pasting text from the clipboard.

The width of the "profile" mass peak is determined either by setting the resolving power of the instrument or by setting the width of the peak at half maximum of its height (*FWHM*).

The profile mass peaks generated can be either of a Gaussian or of a Lorentzian shape. That parameter is configured by selecting the corresponding radio button widget. The number of points used to actually craft the shape of the peak is configurable in the Shape the peak using points spin box widget.

There are two ways to configure the width of the Gaussian/Lorentzian shape around the centroid peak. Either one knows the resolving power of the instrument, or one can measure the full width at half maximum (*FWHM*) of the experimental peak. The Resolving power group box widget contains all the widgets required to define that *FWHM*. Because the *FWHM* actually depends on the resolving power and the resolving power is computed using the *m/z* value, there needs to be a Reference mass provided for the calculation. If the data in the text edit widget on the left hand side pane of the window is filled with data from an isotopic cluster calculation, the reference mass is the *m/z* value of the most intense peak. If the data come from the user either pasting or editing text in the text edit widget, then it is their responsibility to fill in the reference mass.

To define the *FWHM* value, one can just enter it in the corresponding spin box widget. When the editing is finished and the focus leaves the spin box widget, the Bin size logic group box widget will have the FWHM radio button widget selected. The Resolving power spin box widget now contains the proper resolving power value, as computed using the *FWHM* value set to the Reference mass spin box widget.

To define the resolving power value, one can just enter it in the Resolving power spin box widget. When the editing is finished and the focus leaves the spin box widget, the Bin size logic group box widget will have the Resolving power radio button widget selected. The FWHM spin box widget now contains the proper *FWHM* value, as computed using the resolving power value and the value set to the Reference mass spin box widget.

The end point of the configuration procedures described above is to have a proper *m/z* bin size for the mass spectrum that is generated by the combination of the various peak shapes. The greater the resolving power of the instrument, the smaller the *FWHM* value and thus the smaller the bin size. The bin size that is computed is displayed in the Bin size (Th) spin box widget inside of the Manually set the bin size group box widget. If checked, this group box widget becomes active to let the user set the bin size manually. Its computation does not occur anymore upon modification of the values of *FWHM* or of the resolving power or of the reference mass.

The successful configuration of the bin size, *via* the definition of the *FWHM* or of the resolving power takes a bit of patience. The Check parameters push button triggers a

verification of the configuration and provides feedback about it in the line edit widget below the push button.

The Charge spin box widget allows setting the charge of the ions of which the centroid (*m/z*, i) pairs are listed in the text edit widget on the left hand side of the dialog window.

> **Use caution with the value entered in the Charge spin box widget**
>
> When the peak centroid data have been filled in by the isotopic cluster calculation dialog window, do not change this value (let it be 1, as the default). This is because the isotopic cluster calculation has already taken into account the charge that was set at that step.
>
> If the peak centroid data have been filled by editing in place or by pasting text, with *m/z* values not accounting for any charge state, then the charge value needs to be set according to the actual charge of the ions for which the centroid peaks are being shaped.

Once the peak shaping parameters have been checked, click the Run button. The dialog window shifts tab to Results, as shown in Figure 73 – Shaped peaks mass spectral data.

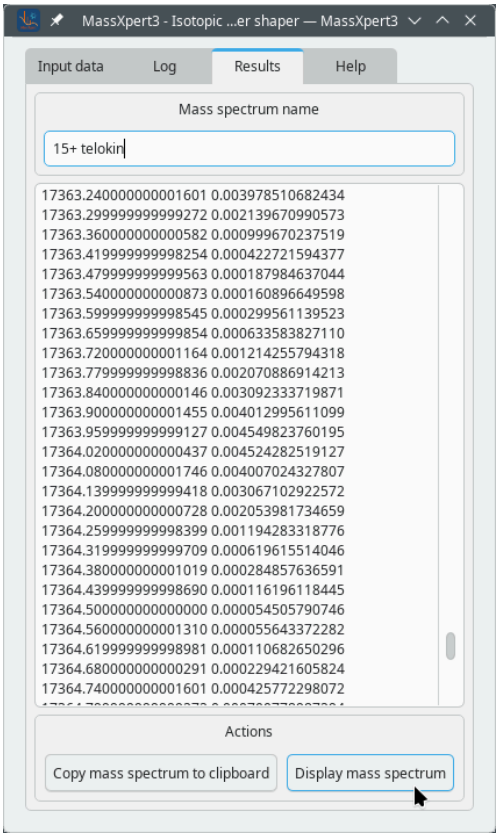


Figure 73: Shaped peaks mass spectral data

The mass spectral data corresponding to the combination of each individual "shaped" centroid peak are displayed in this tab. Plotting these data will produce nicely shaped profile peaks matching what would be seen by recording a mass spectrum on an instrument.



Tip

To ensure that various isotopic cluster calculation-based mass spectra can later be recognized, the user is advised, for the different simulations, to insert distinct names in the Mass spectrum name line edit widget. This name will be used later when creating the mass spectrum if the user asks that the mass spectrum be displayed by clicking Display mass spectrum. In the eventuality that the mass spectrum name is not changed from a simulation to the other, a safeguard process ensures that names are absolutely unambiguous by appending to the mass spectrum name the time at which the mass spectrum is displayed.

Once the mass spectrum has been generated by the peak shaper, it might be copied to the clipboard, for visualization in the companion *MineXpert* software or any other software capable of reading space-separated-values files. It is also possible to actually trigger the

automatic display of the spectrum in *MineXpert* by clicking Display mass spectrum. In this case, *MassXpert* will check if the *MineXpert* program is available on the system, and if so, starts it automatically and transmits *via* the network the mass spectral data to be plotted.

The mass spectrum corresponding to the combination of all the shaped centroid peaks is shown in Figure 74 – Isotopic cluster automatically displayed in *MineXpert*.

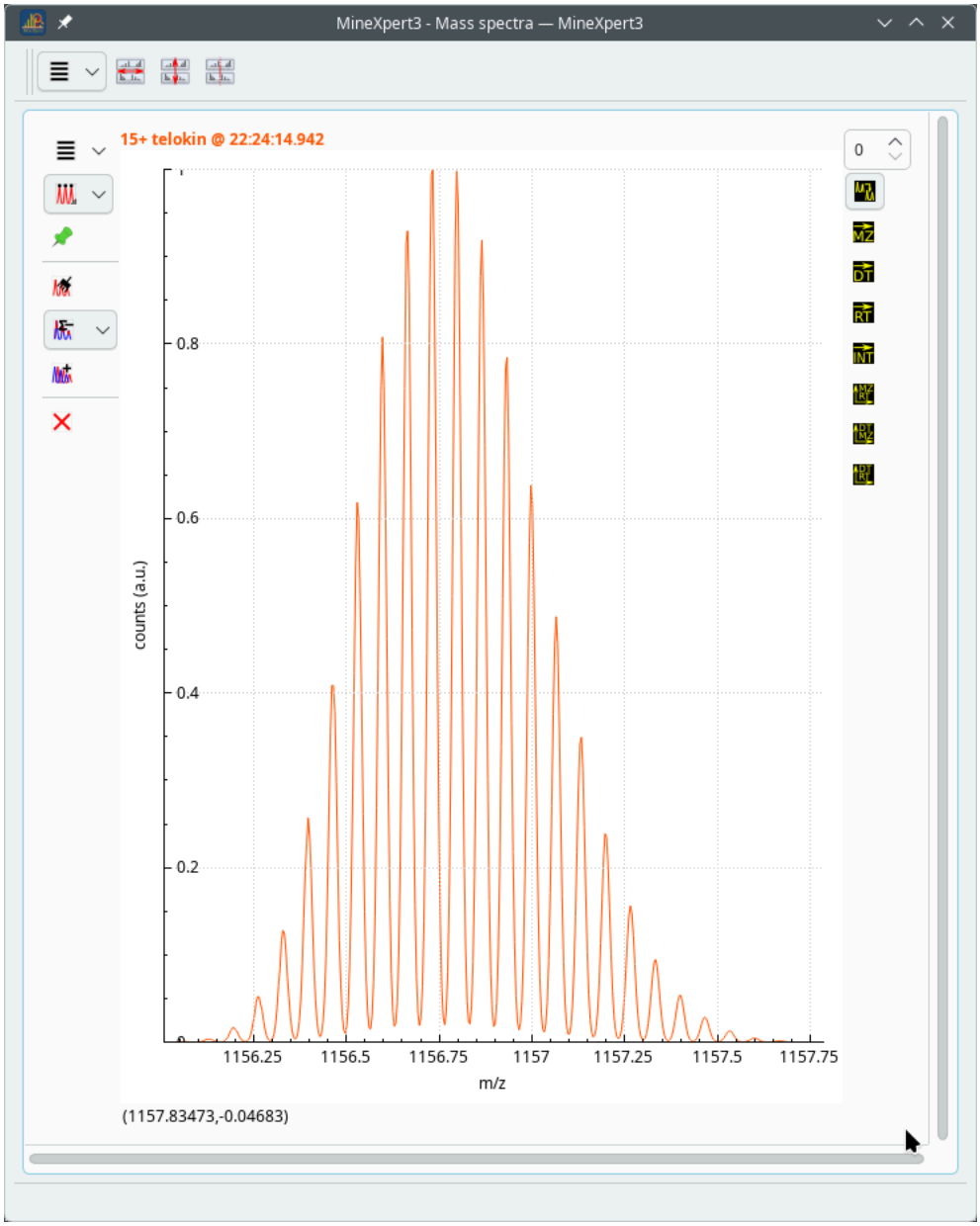


Figure 74: Isotopic cluster automatically displayed in *MineXpert*

If the *MineXpert* program is installed on the system, *MassXpert* starts it and displays the isotopic cluster mass spectrum.

8.3 Combining multiple shaped isotopic clusters' peaks into a single mass spectrum

It often occurs that one wants to run multiples isotopic cluster calculations in order to combine the multiple obtained shaped mass spectra into a single one. This requirement has some tricky implications. Because the formula for which the various isotopic clusters are computed is going to be different each time, the very first *m/z* value of the resulting mass spectrum is going to be different each time. This means that when the *m/z* bins are calculated, the very first bin will be rooted at a different *m/z* value for each one of the mass spectra obtained for the different isotopic cluster calculations. Upon combination of the various mass spectra, even if the resolving power was the same, there will be artifacts visible in Figure 75 – Combination of mass spectra obtained for different isotopic cluster calculations. This is called the "unphased bins" artifact.



Figure 75: Combination of mass spectra obtained for different isotopic cluster calculations

The top two spectra correspond to two isotopic cluster calculations for mono-charged analytes differing in mass by rougly 1 Da. The bottom spectrum correspond to the combination of the blue trace into a copy of the red trace. The "unphased bins" artifact is clearly visible, with what appears as spikes combined on top of the initial peak shape or at its left.

One way to avoid the "unphased bins" artifact is to manage the bin calculation prior to starting the isotopic cluster centroids shaping process. This is shown in Figure 76 – Reusing the isotopic cluster shaper dialog window.

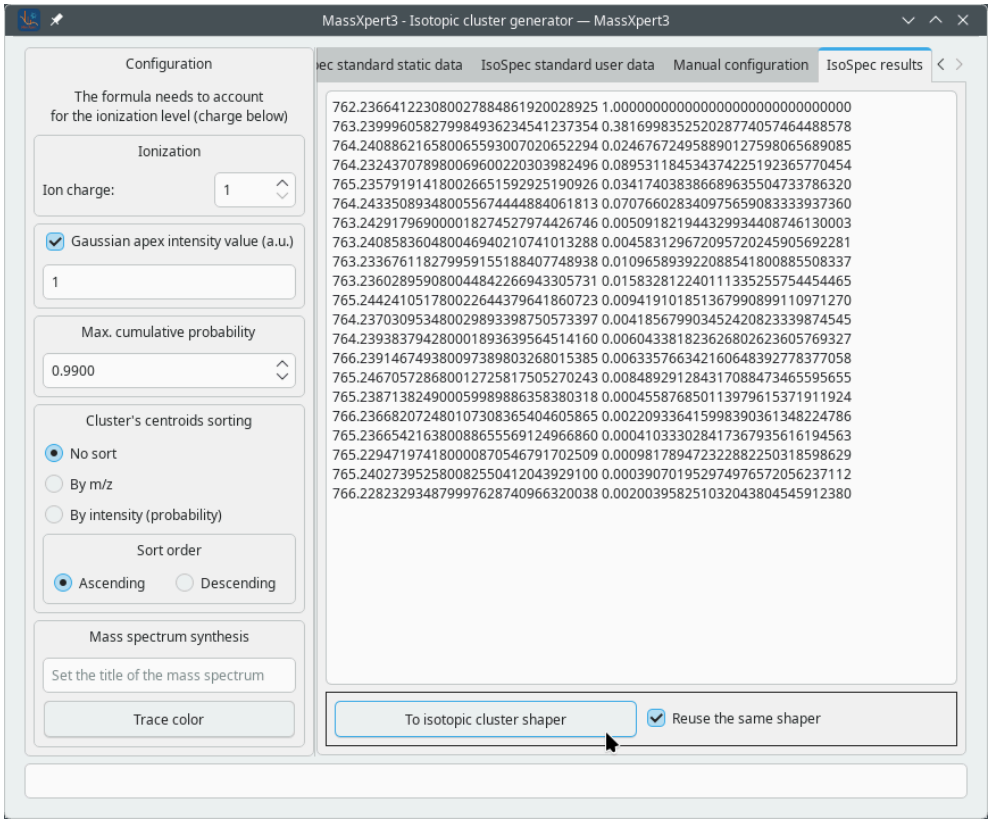


Figure 76: Reusing the isotopic cluster shaper dialog window

When the isotopic cluster centroid peaks have been computed, right before sending them to the isotopic cluster shaper dialog window, check the Reuse the same shaper check box widget.

This has the effect of replacing the centroid peaks data to the same isotopic cluster shaper dialog window, so that other configuration data in it are not modified.

In the isotopic cluster shaper dialog window, there are some configuration steps that will ensure that all the isotopic cluster mass spectra share the same *m/z* bin width and the same set of bins (they are thus "phased"). The configuration steps are illustrated in Figure 77 – Ensuring that the bins are "phased" in the various isotopic cluster mass spectra.

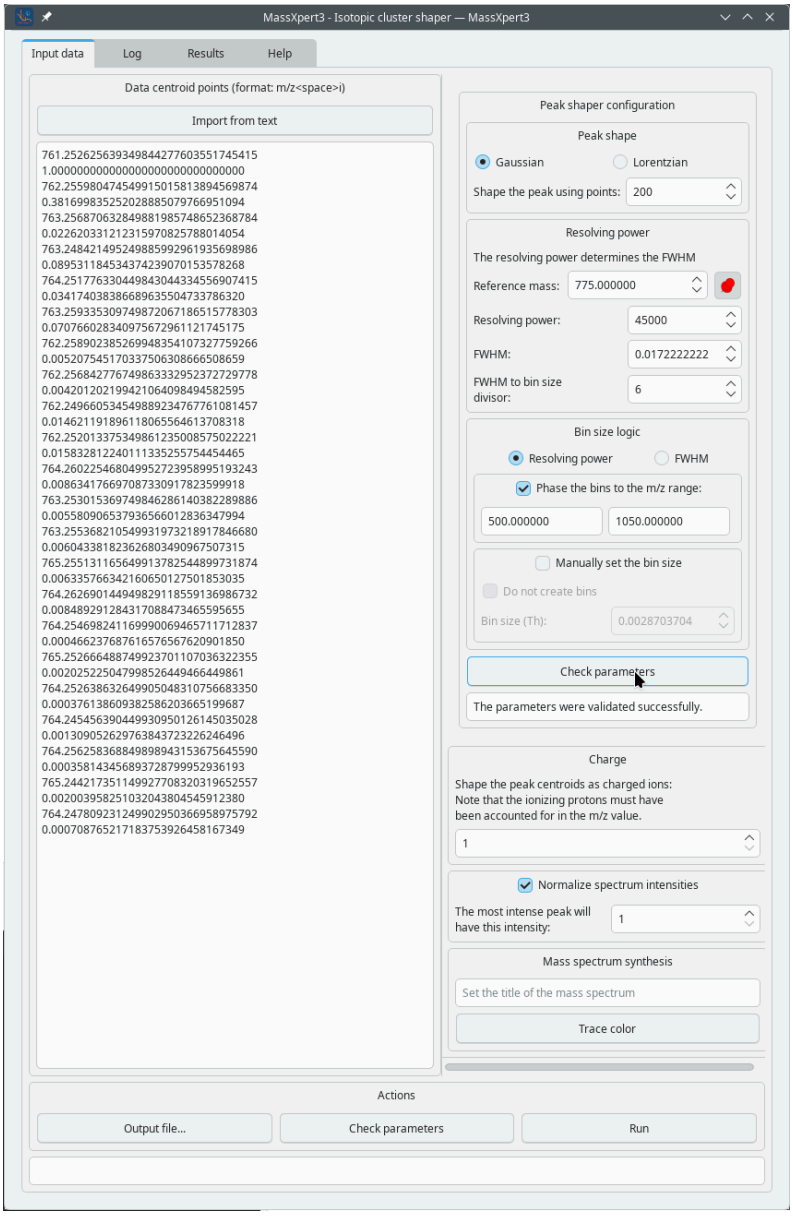


Figure 77: Ensuring that the bins are "phased" in the various isotopic cluster mass spectra

By setting the *m/z* range for which the bins have to be computed according to the *FWHM* value, the user ensures that multiple isotopic cluster mass spectra will combine without the "unphased bins" artifact.

In the figure, the user has pushed the pin on the right hand size of the Reference mass value, so that upon loading a new set of isotopic cluster centroids from the isotopic cluster generator window that value is not modified.

Another setting that is crucial is the *m/z* range set in the Phase the bins to the m/z range group box widget. The start and stop values descring the *m/z* range need to encompass the whole set of *m/z* values in all the mass spectra generated during the isotopic cluster generation session.

Upon editing of the second *m/z* range value (the range end value), the Reference mass *m/z* value is updated to correspond to the middle value of that range. If that value is not correct, the user may modify it right in the double spin box and that value is not going to change.

By using the settings described above, one can ensure that the combination of all the isotopic cluster mass spectra works without the artifact, as shown in Figure 78 – Avoiding the combination artifact due to having "unphased bins".

> **Note**
>
> In order for the combination to work perfectly, it is important that the *m/z* integration parameters in the mass spectrum composite widgets in *MineXpert* be set to No binning. This is because the bins are already set and phased correctly.

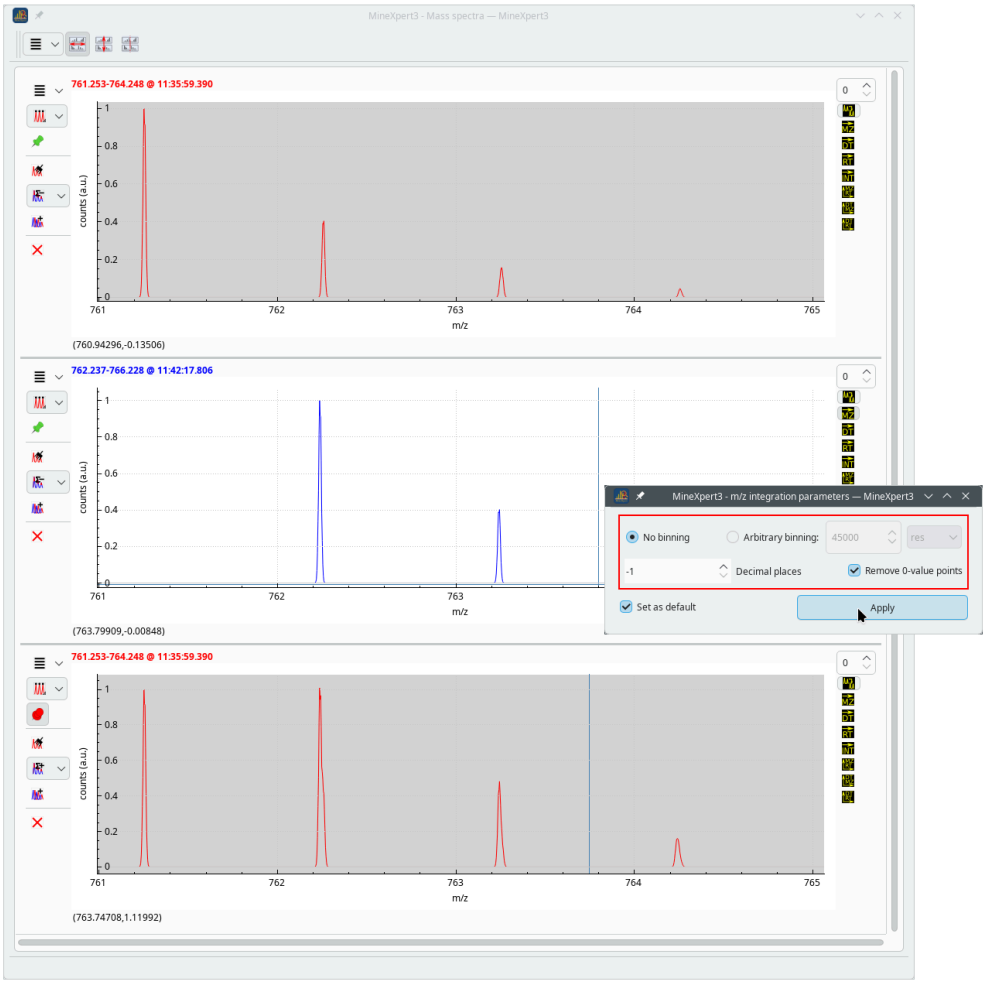


Figure 78: Avoiding the combination artifact due to having "unphased bins"

As visible, the combination of the top two mass spectra does not show the artifact.

9 Utilities

The utilities made available in the **Utilities** menu of the main program window are shown in Figure 79 – Utilities. These are described in detail in the following sections.

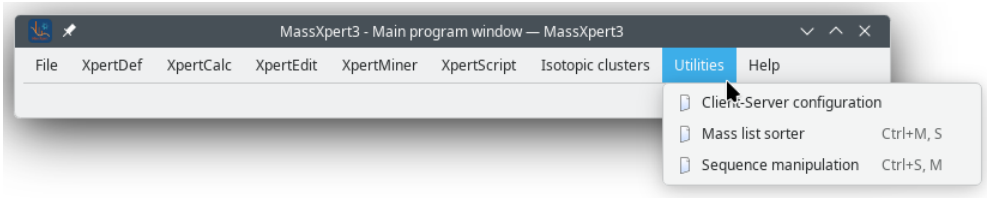


Figure 79: Utilities

The **Utilities** menu provides useful features, in particular the network-based communications with the companion software *MineXpert*.

9.1 Configuration of the *MassXpert* and *MineXpert* intercommunications

The two programs of the *MsXpertSuite* software suite, *MassXpert* and *MineXpert* can exchange data via the network, provided that both are set up for *TCP/IP*-based communications. In *MassXpert*, the configuration of these communications is performed by using the **Utilities⇨Client-Server configuration** menu item. The configuration is performed in the dialog window shown in Figure 80 – Setting-up of the centroid mass peak shaping process.

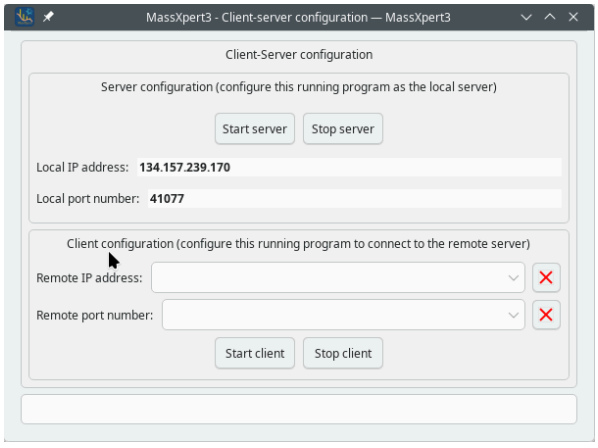


Figure 80: Setting-up of the centroid mass peak shaping process

This dialog window allows one to configure the *TCP/IP*-based communications between the two software pieces of the *MsXpertSuite* software suite: *MassXpert* and *MineXpert*.

To start the server, click the Start server button and the IP address and port number will be displayed. To start the client, insert the IP address and port number in the Client configuration group box and click Start client.

9.1.1 Dynamic configuration of the *MassXpert* and *MineXpert* intercommunications

Each time a task occurring in *MassXpert* requires to display data in *MineXpert* the former program checks if the latter program is already running and if an active connection between the two is up.

If an active connection between *MassXpert* and *MineXpert* is alive, the data are immediately dislayed. Instead, if the *MineXpert* program is effectively running, but there is no active connection between *MassXpert* and *MineXpert*, the user is prompted with a dialog box to suggest configuring a connection (see Figure 81 – Configuration of the dynamic communication between *MassXpert* and *MineXpert*).

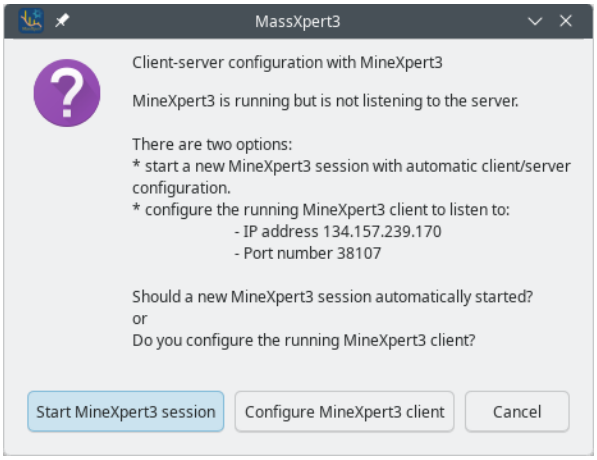


Figure 81: Configuration of the dynamic communication between *MassXpert* and *MineXpert*

> If *MassXpert* finds that *MineXpert* is running but that there is no active connection between the two program, this dialog box allows the user to configure *MineXpert* so that a connection made. If this is done, *MassXpert* will send the data to *MineXpert* for it to display them.

Whenever *MassXpert* cannot find a *MineXpert* process, it will automatically try to start if it finds the *MineXpert* executable installed on the computer. In that case, if the *MineXpert* program can be started, the data are transferred as soon as the process and a communications connection can be set up.

10 Data Customizations

In this chapter, the user will be walked trough an example of how new polymer chemistry definition data can be generated and included in the automatic "data detection system" of *MassXpert* (that is how new polymer chemistry definitions should be registered with the system).

Customization is typically performed by the normal user (not the Administrator nor the Root of the machine) and as such new data are typically stored in the user's "home" directory. On *UNIX* machines, the "home" directory is usually the `/home/username` directory, where username is the login username. On *MS-Windows*, that directory is typically the `C:/Windows/users/username`, once again with username being the login username.



Tip

Although *MS-Windows* pathnames use a back slash ("\"), in this book these are composed using forward slashes for a number of valid reasons. The reader only needs to replace back lashes with the forward variant.

In the next sections we will refer to that "home directory" (be it on *UNIX* or *MS-Windows* machines) as the `$HOME` directory, as this is the standard environment variable describing that directory in *GNU/Linux*.

When *MassXpert* is executed, it automatically tries to read data configuration files. First, the user's *MassXpert* configuration settings are located in the `.config/MassXpert3` directory, in the `configSettings.ini` file. The excerpt below, for user *Rusconi* shows two important stanzas:

```
[SystemFileSystemHierarchyPreferences]
dataDir=/usr/share/massxpert3/data
docDir=/usr/share/doc/massxpert3
polChemDefsDir=/usr/share/massxpert3/data/polChemDefs
polSeqsDir=/usr/share/massxpert3/data/polSeqs

[UserFileSystemHierarchyPreferences]
dataDir=/home/rusconi/.config/MassXpert3/data
polChemDefsDir=/home/rusconi/.config/MassXpert3/data/polChemDefs
polSeqsDir=/home/rusconi/.config/MassXpert3/data/polSeqs
```

This excerpt shows that the system polymer chemistry definitions shipped with the software during install are located at `/usr/share/massxpert3/data`. Inside that directory two subdirectories contain the polymer chemistry definitions installed alongside the program and

the polymer sequence examples, `/usr/share/massxpert3/data/polChemDefs` and `/usr/share/massxpert3/data/polSeqs`, respectively.

MassXpert loads all the available data from these two directories to make them available to the user upon terminating the program initialization.

Once this is done, *MassXpert* reads all the data configuration files in the installation directory (typically, on *GNU/Linux* that would be the configuration data in the `/usr/share/MassXpert/data` directory or, on *MS-Windows*, the `c:/Program Files/MassXpert/data` directory).

We said above that *MassXpert* tries to read the data configuration files from the home directory. But upon its very first execution, right after installation, that directory does not exist, and in fact *MassXpert* creates that directory for us to populate it some day with interesting new data.

The `$HOME/.MassXpert` directory should have a structure mimicking the one that was created upon installation of the software, that is, it should contain the following two directories:

- `polChemDefs`
- `polSeqs`

Those are the directories where the user is invited to store their personal data. In order to start a new definition, one might simply copy in the `polChemDefs` one of the polymer chemistry definitions that are shipped with *MassXpert*. What should be copied? An entire polymer chemistry definition directory, like for example the following:

`/usr/local/share/MassXpert/polChemDefs/protein-1-letter`

or

`C:/Program Files/MassXpert3/polChemDefs/protein-1-letter`

Once that polymer chemistry definition is copied, one may start studying how it actually works. This directory contains the following kinds of files:

- `protein-1-letter.xml`: the polymer chemistry definition file. This is the file that is read upon selection of the corresponding polymer chemistry definition name in *XpertDef*. If the polymer chemistry definition is not yet registered with the system (described later), then open that file by browsing to it by clicking Cancel (see Section 5);
- *SVG* files: *scalar vector graphics* files used to render graphically the sequence in the sequence editor. For example, `arginine.svg` contains the graphical representation of the arginine monomer. There are such graphics files also for the modifications (like, for example, the `sulphation.svg` contains the graphical representation of the sulphation modification.

Figure 82: The polymer chemistry definition directory

shows two examples of *SVG* files belonging to two distinct polymer chemistry definitions;

- `chemPad.conf`: configuration file for the chemical pad in the *XpertCalc* module;
- `monomer_dictionary`: file establishing the relationship between any monomer code of the polymer chemistry definition and the graphical *SVG* file to be used to render graphically that monomer in the sequence editor;
- `modification_dictionary`: file establishing the relationship between any monomer modification (see Section 6.8.1 – Selected Monomer(s) Modification) and the graphical *SVG* file to be used to render graphically that modification onto the modified monomer in the sequence editor;
- `cross_linker_dictionary`: file establishing the relationship between any cross-link (see Section 6.9 – Monomer Cross-linking) and the graphical *SVG* file to be used to render graphically that cross-link onto the cross-linked monomers in the sequence editor;
- `pka_ph_pi.xml`: file describing the acido-basic data (see Section 6.15 – pKa, pH, pI and Charges) pertaining to ionizable chemical groups in the different entities of the polymer chemistry definition;

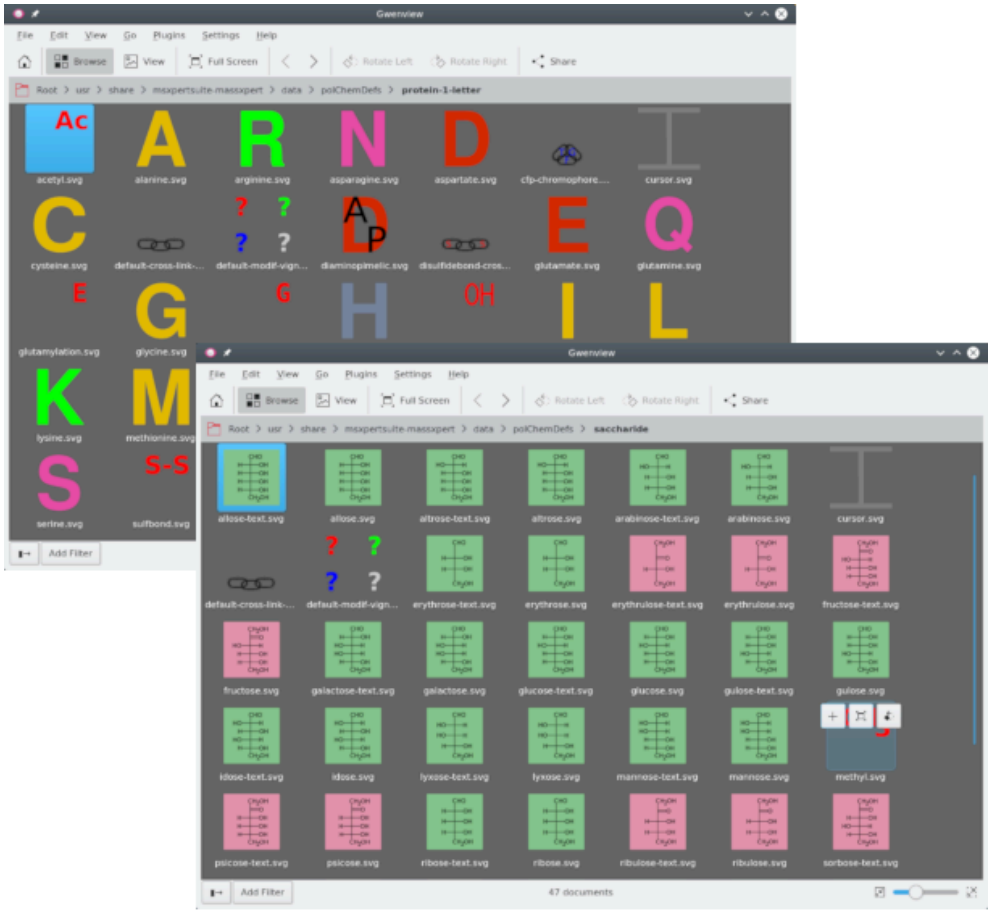


Figure 82: The polymer chemistry definition directory

Each monomer of the polymer chemistry definition ought to have a corresponding *SVG* file with which it has to be rendered graphically should that monomer be inserted in the polymer sequence. This example shows two *SVG* files corresponding to two monomers each belonging to a different polymer chemistry definition.

The polymer sequence editor is not a classical editor. There is no font in this editor: when the user starts keying-in a polymer sequence in the editor, the small *SVG* graphics files are rendered into raster *vignettes* at both the proper resolution and screen size and displayed in the sequence editor. The user is totally in charge of designing the *SVG* graphics files for each of the monomers defined in the polymer sequence editor. Of course, reusing material is perfectly possible.

> **Tip**
>
> One powerful software to edit *SVG* files is Inkscape, on any platform.

There is one constraint: that the `monomer_dictionary` file lists with precision "what monomer code goes with what *SVG* graphics file". That file has the following contents, for example, for the "protein-1-letter" polymer chemistry definition, as shipped in the *MassXpert* package:

```
# This file is part of the MassXpert3 project.
# The "MassXpert3" project is released —in its entirety— under the
# GNU General Public License and was started (in the form of the
# GNU polyxmass project) at the Centre National de la Recherche
# Scientifique (FRANCE), that granted me the formal authorization to
# publish it under this Free Software License.
# Copyright (C) 2006,2023 Filippo Rusconi
# This is the monomer_dictionary file where the correspondences
# between the codes of each monomer and their graphic file (pixmap
# file called "image") used to graphically render them in the
# sequence editor are made.
# The format of the file is like this :
#
#————————————-
# A%alanine.svg
# where A is the monomer code and alanine.svg is a
# resolution-independent svg file.
# Each line starting with a '#' character is a comment and is ignored
# during parsing of this file.
# This file is case-sensitive.
A%alanine.svg
C%cysteine.svg
D%aspartate.svg
[...]
T%threonine.svg
V%valine.svg
W%tryptophan.svg
Y%tyrosine.svg
```

What one sees from the contents of the file is that each monomer code has an associated *SVG* file. For example, when the user has to key-in a valine monomer, they key-in the code V and *XpertEdit* knows that the monomer vignette to show has to be rendered using the `valine.svg` file.

For the monomer modification graphical rendering, the situation is somewhat different, as seen in the `modification_dictionary` file:

```
" # This file is part of the MassXpert project.
# The "MassXpert3" project is released —in its entirety— under the
# GNU General Public License and was started (in the form of the GNU
# polyxmass project) at the Centre National de la Recherche
# Scientifique (FRANCE), that granted me the formal authorization to
# publish it under this Free Software License.
# Copyright(C) 2006,2023 Filippo Rusconi
# This is the modification_dictionary file where the correspondences
# between the name of each modification and their graphic file (pixmap
# file called "image") used to graphicallly render them in the
# sequence editor are made. Also, the graphical operation that is to
# be performed upon chemical modification of a monomer is listed ('T'
# for transparent and 'O' for opaque). See the manual for details.
# The format of the file is like this :
# ————————————-
Phosphorylation%T%phospho.svg
# where Phosphorylation is the name of the modification. T indicates
# that the visual rendering of the modification is a transparent
# process (O indicates that the visual rendering of the modification
# is a full image replacement 'O' like opaque).
# phospho.svg is a resolution-independent svg file.
# Each line starting with a '#' character is a comment and is ignored
# during parsing of this file.
# This file is case-sensitive.
Phosphorylation%T%phospho.svg
Sulphation%T%sulpho.svg
AmidationAsp%O%asparagine.svg
Acetylation%T%acetyl.svg
AmidationGlu%O%glutamine.svg
Oxidation%T%oxidation.svg"
```

There are two ways to render a chemical modification of a monomer:

- *Opaque* rendering: the initial monomer vignette is replaced using the one listed in the file for the modification. One example is the "AmidationGlu" modification:

  AmidationGlu%O%glutamine.svg

  Upon amidation of a glutamyl residue ("AmidationGlu" is the name of a modification in the current polymer chemistry definition), the graphical representation of the modification involves the *replacement* of the glutamyl residue vignette in the sequence editor with the new one, that happens to be in the `glutamine.svg` file. In other words, the process involves an "Opaque" overlay of the vignette for unmodified Glu with a vignette rendered by using the `glutamine.svg` file.
- *Transparent* rendering: the initial monomer vignette is overlaid with a new vignette that is read from an *SVG* file that has a transparent background. One example is the "Phosphorylation" modification:

  Phosphorylation%T%phospho.svg

The monomer undergoing a phosphorylation has its vignette overlaid with a "Transparent" one showing a small red "P" that is read from the `phospho.svg` file.

When designing vignettes, the best thing to do is to convert the text to path, so that the rendering is absolutely perfect.

> **Warning**
>
> It is absolutely essential, for the proper working of the sequence editor, that the *SVG* files be square (that is, width = height).

Once the new polymer chemistry has been correctly defined, it is time to register that new definition to the system. To recap: all the files for that definition should reside in a same directory, exactly the same way as the files pertaining to a given polymer chemistry definition are shipped in *MassXpert* altogether in one directory. The name of the new polymer chemistry definition should be unambiguous, with respect to other registered polymer chemistry definitions.

The way personal polymer chemistry definitions are registered is by creating a personal polymer chemistry definition catalogue file, which must comply with both following requirements:

- Be named `xxxxx-polChemDefsCat`, with "xxxxx" being a discretionary string (this might well be your login username). The requirement is that `-polChemDefsCat` be the last part of the filename.

  > **Tip**
  >
  > Please *DO NOT USE* spaces or diacritical signs in your filenames. *RESTRICT* yourself to ASCII characters between [a-z], [0-9], "_" and "-".
  >
  > This is actually something very general as a recommendation in order to not suffer from severe headaches when you least expect them...

- Be located in the `$HOME/.MassXpert/data/polChemDefs` directory and have the following format:

  dna=/path/to/definition/directory/dna/dna.xml

  In this example, the "dna" polymer chemistry definition is being registered as a file `dna.xml` located in the `dna` directory, itself located in the `/path/to/definition/directory` directory;

Note that if a new polymer chemistry definition should be made available system-wide, then it is logical that its directory be placed along the ones shipped with *MassXpert* and a new local catalogue file might be created to register the new polymer chemistry definition.

At this point the new polymer chemistry definition might be tested. Typically, that involves restarting the *MassXpert* program and creating a brand new polymer sequence of the new definition type. The first step is to check if the new definition is successfully registered with the system, that is, it should show up a an available definition upon creation of the new polymer sequence. If not, then that means that the catalogue file could not be found or parsed correctly.

When problems like this one occurs, the first thing to do is to ensure that the console window (on *MS-Windows* it is systematically started along with the program; on *GNU/Linux* the way to have it is to start the program from the shell) so as to look with attention at the different messages that might help understanding what is failing.

Please, do not hesitate to submit bug reports (See Feedback to the developers).

Appendix

GNU General Public License version 3

Version 3, 29 June 2007

Copyright © 2007 Free Software Foundation, Inc. https://fsf.org/

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

Preamble

The *GNU* General Public License is a free, copyleft license for software and other kinds of works.

The licenses for most software and other practical works are designed to take away your freedom to share and change the works. By contrast, the *GNU* General Public License is intended to guarantee your freedom to share and change all versions of a program—to make sure it remains free software for all its users. We, the Free Software Foundation, use the *GNU* General Public License for most of our software; it applies also to any other work released this way by its authors. You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for them if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs, and that you know you can do these things.

To protect your rights, we need to prevent others from denying you these rights or asking you to surrender the rights. Therefore, you have certain responsibilities if you distribute copies of the software, or if you modify it: responsibilities to respect the freedom of others.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must pass on to the recipients the same freedoms that you received. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

Developers that use the *GNU GPL* protect your rights with two steps: (1) assert copyright on the software, and (2) offer you this License giving you legal permission to copy, distribute and/or modify it.

For the developers' and authors' protection, the *GPL* clearly explains that there is no warranty for this free software. For both users' and authors' sake, the *GPL* requires that modified versions be marked as changed, so that their problems will not be attributed erroneously to authors of previous versions.

Some devices are designed to deny users access to install or run modified versions of the software inside them, although the manufacturer can do so. This is fundamentally incompatible with the aim of protecting users' freedom to change the software. The systematic pattern of such abuse occurs in the area of products for individuals to use, which is precisely where it is most unacceptable. Therefore, we have designed this version of the *GPL* to prohibit the practice for those products. If such problems arise substantially in other domains, we stand ready to extend this provision to those domains in future versions of the *GPL*, as needed to protect the freedom of users.

Finally, every program is threatened constantly by software patents. States should not allow patents to restrict development and use of software on general-purpose computers, but in those that do, we wish to avoid the special danger that patents applied to a free program could make it effectively proprietary. To prevent this, the *GPL* assures that patents cannot be used to render the program non-free.

The precise terms and conditions for copying, distribution and modification follow.

TERMS AND CONDITIONS

0. Definitions.

"This License" refers to version 3 of the *GNU* General Public License.

"Copyright" also means copyright-like laws that apply to other kinds of works, such as semiconductor masks.

"The Program" refers to any copyrightable work licensed under this License. Each licensee is addressed as "you". "Licensees" and "recipients" may be individuals or organizations.

To "modify" a work means to copy from or adapt all or part of the work in a fashion requiring copyright permission, other than the making of an exact copy. The resulting work is called a "modified
version" of the earlier work or a work "based on" the earlier work.

A "covered work" means either the unmodified Program or a work based on the Program.

To "propagate" a work means to do anything with it that, without permission, would make you directly or secondarily liable for infringement under applicable copyright law, except executing it on a computer or modifying a private copy. Propagation includes

copying, distribution (with or without modification), making available to the public, and in some countries other activities as well.

To "convey" a work means any kind of propagation that enables other parties to make or receive copies. Mere interaction with a user through a computer network, with no transfer of a copy, is not conveying.

An interactive user interface displays "Appropriate Legal
Notices" to the extent that it includes a convenient and prominently visible feature that (1) displays an appropriate copyright notice, and (2) tells the user that there is no warranty for the work (except to the extent that warranties are provided), that licensees may convey the work under this License, and how to view a copy of this License. If the interface presents a list of user commands or options, such as a menu, a prominent item in the list meets this criterion.

1. Source Code.

The "source code" for a work means the preferred form of the work for making modifications to it. "Object code" means any non-source form of a work.

A "Standard Interface" means an interface that either is an official standard defined by a recognized standards body, or, in the case of interfaces specified for a particular programming language, one that is widely used among developers working in that language.

The "System Libraries" of an executable work include anything, other than the work as a whole, that (a) is included in the normal form of packaging a Major Component, but which is not part of that Major Component, and (b) serves only to enable use of the work with that Major Component, or to implement a Standard Interface for which an implementation is available to the public in source code form. A "Major Component", in this context, means a major essential component (kernel, window system, and so on) of the specific operating system (if any) on which the executable work runs, or a compiler used to produce the work, or an object code interpreter used to run it.

The "Corresponding Source" for a work in object code form means all the source code needed to generate, install, and (for an executable work) run the object code and to modify the work, including scripts to control those activities. However, it does not include the work's System Libraries, or general-purpose tools or generally available free programs which are used unmodified in performing those activities but which are not part of the work. For example, Corresponding Source includes interface definition files associated with source files for the work, and the source code for shared libraries and dynamically linked subprograms that the work is specifically designed to require, such as by intimate data communication or control flow between those subprograms and other parts of the work.

The Corresponding Source need not include anything that users can regenerate automatically from other parts of the Corresponding Source.

The Corresponding Source for a work in source code form is that same work.

2. Basic Permissions.

All rights granted under this License are granted for the term of copyright on the Program, and are irrevocable provided the stated conditions are met. This License explicitly affirms your unlimited permission to run the unmodified Program. The output from running a covered work is covered by this License only if the output, given its content, constitutes a covered work. This License acknowledges your rights of fair use or other equivalent, as provided by copyright law.

You may make, run and propagate covered works that you do not convey, without conditions so long as your license otherwise remains in force. You may convey covered works to others for the sole purpose of having them make modifications exclusively for you, or provide you with facilities for running those works, provided that you comply with the terms of this License in conveying all material for which you do not control copyright. Those thus making or running the covered works for you must do so exclusively on your behalf, under your direction and control, on terms that prohibit them from making any copies of your copyrighted material outside their relationship with you.

Conveying under any other circumstances is permitted solely under the conditions stated below. Sublicensing is not allowed; section 10 makes it unnecessary.

3. Protecting User's Legal Rights From Anti-Circumvention Law.

No covered work shall be deemed part of an effective technological measure under any applicable law fulfilling obligations under article 11 of the WIPO copyright treaty adopted on 20 December 1996, or similar laws prohibiting or restricting circumvention of such measures.

When you convey a covered work, you waive any legal power to forbid circumvention of technological measures to the extent such circumvention is effected by exercising rights under this License with respect to the covered work, and you disclaim any intention to limit operation or modification of the work as a means of enforcing, against the work's users, your or third parties' legal rights to forbid circumvention of technological measures.

4. Conveying Verbatim Copies.

You may convey verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice; keep intact all notices stating that this License and any non-permissive terms added in accord with section 7 apply to the code; keep intact all notices of the absence of any warranty; and give all recipients a copy of this License along with the Program.

You may charge any price or no price for each copy that you convey, and you may offer support or warranty protection for a fee.

5. Conveying Modified Source Versions.

You may convey a work based on the Program, or the modifications to produce it from the Program, in the form of source code under the terms of section 4, provided that you also meet all of these conditions:

1. The work must carry prominent notices stating that you modified it, and giving a relevant date.
2. The work must carry prominent notices stating that it is released under this License and any conditions added under section 7. This requirement modifies the requirement in section 4 to "keep intact all notices".
3. You must license the entire work, as a whole, under this License to anyone who comes into possession of a copy. This License will therefore apply, along with any applicable section 7 additional terms, to the whole of the work, and all its parts, regardless of how they are packaged. This License gives no permission to license the work in any other way, but it does not invalidate such permission if you have separately received it.
4. If the work has interactive user interfaces, each must display Appropriate Legal Notices; however, if the Program has interactive interfaces that do not display Appropriate Legal Notices, your work need not make them do so.

A compilation of a covered work with other separate and independent works, which are not by their nature extensions of the covered work, and which are not combined with it such as to form a larger program, in or on a volume of a storage or distribution medium, is called an "aggregate" if the compilation and its resulting copyright are not used to limit the access or legal rights of the compilation's users beyond what the individual works permit. Inclusion of a covered work in an aggregate does not cause this License to apply to the other parts of the aggregate.

6. Conveying Non-Source Forms.

You may convey a covered work in object code form under the terms of sections 4 and 5, provided that you also convey the machine-readable Corresponding Source under the terms of this License, in one of these ways:

1. Convey the object code in, or embodied in, a physical product (including a physical distribution medium), accompanied by the Corresponding Source fixed on a durable physical medium customarily used for software interchange.
2. Convey the object code in, or embodied in, a physical product (including a physical distribution medium), accompanied by a written offer, valid for at least three years and valid for as long as you offer spare parts or customer support for that product model, to give anyone who possesses the object code either (1) a copy of the Corresponding Source for all the software in the product that is covered by this License, on a durable physical medium customarily used for software interchange, for a price no more than your reasonable cost of physically performing this conveying of source, or (2) access to copy the Corresponding Source from a network server at no charge.
3. Convey individual copies of the object code with a copy of the written offer to provide the Corresponding Source. This alternative is allowed only occasionally and noncommercially, and only if you received the object code with such an offer, in accord with subsection 6b.
4. Convey the object code by offering access from a designated place (gratis or for a charge), and offer equivalent access to the Corresponding Source in the same way through the same place at no further charge. You need not require recipients to copy the Corresponding Source along with the object code. If the place to copy the object code is a network server, the Corresponding Source may be on a different server (operated by you or a third party) that supports equivalent copying facilities, provided you maintain clear directions next to the object code saying where to find the Corresponding Source. Regardless of what server hosts the Corresponding Source, you remain obligated to ensure that it is available for as long as needed to satisfy these requirements.
5. Convey the object code using peer-to-peer transmission, provided you inform other peers where the object code and Corresponding Source of the work are being offered to the general public at no charge under subsection 6d.

A separable portion of the object code, whose source code is excluded from the Corresponding Source as a System Library, need not be included in conveying the object code work.

A “User Product” is either (1) a “consumer product”, which means any tangible personal property which is normally used for personal, family, or household purposes, or (2) anything designed or sold for incorporation into a dwelling. In determining whether a

product is a consumer product, doubtful cases shall be resolved in favor of coverage. For a particular product received by a particular user, "normally used" refers to a typical or common use of that class of product, regardless of the status of the particular user or of the way in which the particular user actually uses, or expects or is expected to use, the product. A product is a consumer product regardless of whether the product has substantial commercial, industrial or non-consumer uses, unless such uses represent the only significant mode of use of the product.

"Installation Information" for a User Product means any methods, procedures, authorization keys, or other information required to install and execute modified versions of a covered work in that User Product from a modified version of its Corresponding Source. The information must suffice to ensure that the continued functioning of the modified object code is in no case prevented or interfered with solely because modification has been made.

If you convey an object code work under this section in, or with, or specifically for use in, a User Product, and the conveying occurs as part of a transaction in which the right of possession and use of the User Product is transferred to the recipient in perpetuity or for a fixed term (regardless of how the transaction is characterized), the Corresponding Source conveyed under this section must be accompanied by the Installation Information. But this requirement does not apply if neither you nor any third party retains the ability to install modified object code on the User Product (for example, the work has been installed in *ROM*).

The requirement to provide Installation Information does not include a requirement to continue to provide support service, warranty, or updates for a work that has been modified or installed by the recipient, or for the User Product in which it has been modified or installed. Access to a network may be denied when the modification itself materially and adversely affects the operation of the network or violates the rules and protocols for communication across the network.

Corresponding Source conveyed, and Installation Information provided, in accord with this section must be in a format that is publicly documented (and with an implementation available to the public in source code form), and must require no special password or key for unpacking, reading or copying.

7. Additional Terms.

"Additional permissions" are terms that supplement the terms of this License by making exceptions from one or more of its conditions. Additional permissions that are applicable to the entire Program shall be treated as though they were included in this License, to the extent that they are valid under applicable law. If additional permissions apply only to part

of the Program, that part may be used separately under those permissions, but the entire Program remains governed by this License without regard to the additional permissions.

When you convey a copy of a covered work, you may at your option remove any additional permissions from that copy, or from any part of it. (Additional permissions may be written to require their own removal in certain cases when you modify the work.) You may place additional permissions on material, added by you to a covered work, for which you have or can give appropriate copyright permission.

Notwithstanding any other provision of this License, for material you add to a covered work, you may (if authorized by the copyright holders of that material) supplement the terms of this License with terms:

1. Disclaiming warranty or limiting liability differently from the terms of sections 15 and 16 of this License; or
2. Requiring preservation of specified reasonable legal notices or author attributions in that material or in the Appropriate Legal Notices displayed by works containing it; or
3. Prohibiting misrepresentation of the origin of that material, or requiring that modified versions of such material be marked in reasonable ways as different from the original version; or
4. Limiting the use for publicity purposes of names of licensors or authors of the material; or
5. Declining to grant rights under trademark law for use of some trade names, trademarks, or service marks; or
6. Requiring indemnification of licensors and authors of that material by anyone who conveys the material (or modified versions of it) with contractual assumptions of liability to the recipient, for any liability that these contractual assumptions directly impose on those licensors and authors.

All other non-permissive additional terms are considered "further restrictions" within the meaning of section 10. If the Program as you received it, or any part of it, contains a notice stating that it is governed by this License along with a term that is a further restriction, you may remove that term. If a license document contains a further restriction but permits relicensing or conveying under this License, you may add to a covered work material governed by the terms of that license document, provided that the further restriction does not survive such relicensing or conveying.

If you add terms to a covered work in accord with this section, you must place, in the relevant source files, a statement of the additional terms that apply to those files, or a notice indicating where to find the applicable terms.

Additional terms, permissive or non-permissive, may be stated in the form of a separately written license, or stated as exceptions; the above requirements apply either way.

8. Termination.

You may not propagate or modify a covered work except as expressly provided under this License. Any attempt otherwise to propagate or modify it is void, and will automatically terminate your rights under this License (including any patent licenses granted under the third paragraph of section 11).

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, you do not qualify to receive new licenses for the same material under section 10.

9. Acceptance Not Required for Having Copies.

You are not required to accept this License in order to receive or run a copy of the Program. Ancillary propagation of a covered work occurring solely as a consequence of using peer-to-peer transmission to receive a copy likewise does not require acceptance. However, nothing other than this License grants you permission to propagate or modify any covered work. These actions infringe copyright if you do not accept this License. Therefore, by modifying or propagating a covered work, you indicate your acceptance of this License to do so.

10. Automatic Licensing of Downstream Recipients.

Each time you convey a covered work, the recipient automatically receives a license from the original licensors, to run, modify and propagate that work, subject to this License. You are not responsible for enforcing compliance by third parties with this License.

An "entity transaction" is a transaction transferring control of an organization, or substantially all assets of one, or subdividing an organization, or merging organizations. If propagation of a covered work results from an entity transaction, each party to that transaction who receives a copy of the work also receives whatever licenses to the work the party's predecessor in interest had or could give under the previous paragraph, plus a right to possession of the Corresponding Source of the work from the predecessor in interest, if the predecessor has it or can get it with reasonable efforts.

You may not impose any further restrictions on the exercise of the rights granted or affirmed under this License. For example, you may not impose a license fee, royalty, or other charge for exercise of rights granted under this License, and you may not initiate litigation (including a cross-claim or counterclaim in a lawsuit) alleging that any patent claim is infringed by making, using, selling, offering for sale, or importing the Program or any portion of it.

11. Patents.

A "contributor" is a copyright holder who authorizes use under this License of the Program or a work on which the Program is based. The work thus licensed is called the contributor's "contributor version".

A contributor's "essential patent claims" are all patent claims owned or controlled by the contributor, whether already acquired or hereafter acquired, that would be infringed by some manner, permitted by this License, of making, using, or selling its contributor version, but do not include claims that would be infringed only as a consequence of further modification of the contributor version. For purposes of this definition, "control" includes the right to grant patent sublicenses in a manner consistent with the requirements of this License.

Each contributor grants you a non-exclusive, worldwide, royalty-free patent license under the contributor's essential patent claims, to make, use, sell, offer for sale, import and otherwise run, modify and propagate the contents of its contributor version.

In the following three paragraphs, a "patent license" is any express agreement or commitment, however denominated, not to enforce a patent (such as an express permission to practice a patent or covenant not to sue for patent infringement). To "grant" such a patent license to a party means to make such an agreement or commitment not to enforce a patent against the party.

If you convey a covered work, knowingly relying on a patent license, and the Corresponding Source of the work is not available for anyone to copy, free of charge and under the terms of this License, through a publicly available network server or other readily accessible means, then you must either (1) cause the Corresponding Source to be so available, or (2)

arrange to deprive yourself of the benefit of the patent license for this particular work, or (3) arrange, in a manner consistent with the requirements of this License, to extend the patent license to downstream recipients. "Knowingly relying" means you have actual knowledge that, but for the patent license, your conveying the covered work in a country, or your recipient's use of the covered work in a country, would infringe one or more identifiable patents in that country that you have reason to believe are valid.

If, pursuant to or in connection with a single transaction or arrangement, you convey, or propagate by procuring conveyance of, a covered work, and grant a patent license to some of the parties receiving the covered work authorizing them to use, propagate, modify or convey a specific copy of the covered work, then the patent license you grant is automatically extended to all recipients of the covered work and works based on it.

A patent license is "discriminatory" if it does not include within the scope of its coverage, prohibits the exercise of, or is conditioned on the non-exercise of one or more of the rights that are specifically granted under this License. You may not convey a covered work if you are a party to an arrangement with a third party that is in the business of distributing software, under which you make payment to the third party based on the extent of your activity of conveying the work, and under which the third party grants, to any of the parties who would receive the covered work from you, a discriminatory patent license (a) in connection with copies of the covered work conveyed by you (or copies made from those copies), or (b) primarily for and in connection with specific products or compilations that contain the covered work, unless you entered into that arrangement, or that patent license was granted, prior to 28 March 2007.

Nothing in this License shall be construed as excluding or limiting any implied license or other defenses to infringement that may otherwise be available to you under applicable patent law.

12. No Surrender of Other's Freedom.

If conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot convey a covered work so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not convey it at all. For example, if you agree to terms that obligate you to collect a royalty for further conveying from those to whom you convey the Program, the only way you could satisfy both those terms and this License would be to refrain entirely from conveying the Program.

13. Use with the #acronym(GNU) Affero General Public License.

Notwithstanding any other provision of this License, you have permission to link or combine any covered work with a work licensed under version 3 of the *GNU* Affero General Public License into a single combined work, and to convey the resulting work. The terms of this License will continue to apply to the part which is the covered work, but the special requirements of the *GNU* Affero General Public License, section 13, concerning interaction through a network will apply to the combination as such.

14. Revised Versions of this License.

The Free Software Foundation may publish revised and/or new versions of the *GNU* General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies that a certain numbered version of the *GNU* General Public License "or any later version" applies to it, you have the option of following the terms and conditions either of that numbered version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of the *GNU* General Public License, you may choose any version ever published by the Free Software Foundation.

If the Program specifies that a proxy can decide which future versions of the *GNU* General Public License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Program.

Later license versions may give you additional or different permissions. However, no additional obligations are imposed on any author or copyright holder as a result of your choosing to follow a later version.

15. Disclaimer of Warranty.

THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFEC-

TIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

16. Limitation of Liability.

IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MODIFIES AND/OR CONVEYS THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

17. Interpretation of Sections 15 and 16.

If the disclaimer of warranty and limitation of liability provided above cannot be given local legal effect according to their terms, reviewing courts shall apply local law that most closely approximates an absolute waiver of all civil liability in connection with the Program, unless a warranty or assumption of liability accompanies a copy of the Program in return for a fee.

END OF TERMS AND CONDITIONS

How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively state the exclusion of warranty; and each file should have at least the “copyright” line and a pointer to where the full notice is found.

> *one line to give the program's name and a brief idea of what it does*
>
> Copyright (C) *year — name of author*

> This program is free software: you can redistribute it and/or modify it under the terms of the *GNU* General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.
>
> This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the *GNU* General Public License for more details.
>
> You should have received a copy of the *GNU* General Public License along with this program. If not, see https://www.gnu.org/licenses/.

Also add information on how to contact you by electronic and paper mail.

If the program does terminal interaction, make it output a short notice like this when it starts in an interactive mode:

> *program* Copyright (C) *year* — *name of author*
>
> This program comes with ABSOLUTELY NO WARRANTY; for details type "`show w`". This is free software, and you are welcome to redistribute it under certain conditions; type "`show c`" for details.

The hypothetical commands "`show w`" and "`show c`" should show the appropriate parts of the General Public License. Of course, your program's commands might be different; for a GUI interface, you would use an "about box".

You should also get your employer (if you work as a programmer) or school, if any, to sign a "copyright disclaimer" for the program, if necessary. For more information on this, and how to apply and follow the *GNU GPL*, see https://www.gnu.org/licenses/.

The *GNU* General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the *GNU* Lesser General Public License instead of this License. But first, please read https://www.gnu.org/licenses/why-not-lgpl.html.

Colophon

About the authors

Filippo Rusconi is a senior research scientist at the French national research council (*Centre national de la Recherche scientifique*, *CNRS*). Filippo has a background in biochemistry and organic chemistry and was trained during his Ph.D. as a bioanalytical chemist. He has extensive knowledge of analytical techniques involved in the study of biopolymers.

Filippo Rusconi is the author of a handbook about mass spectrometry for biochemists (French). The book was published by the French sci/tech publisher Lavoisier Tech&Doc.

Book preparation

The look of this book (*PDF* file) is the result of me having read many books from the O'Reilly publisher.

The typesetting of the book has been done on a Debian GNU/Linux computer using only Free Software. The *Typst* application was used for rendering the document code.

The main font used is EBGaramond and the symbol/mathematical font was from the STIX project (font: STIX2Math).

The screen shots were taken with Spectacle, the screen capture program shipped along with my KDE desktop environment and resampled either with *Gimp* (the GNU image manipulation program) or *Krita*. Illustrations were done in Inkscape, a vectorial drawing software.

Index

A

Action formula 26
Application
  Preferences
    Configuration 30
    Filesystem hierarchy 32
    General options 30
      Decimals 30
      Threads 30
    Shortcuts 31

C

Calculator 116
Centroids 134
  Shaping into profiled peaks 136
Cleavage agent
  Definition 61
Cross-linker
  Definition 60
Customization 149
  Data files 150
    Modification dictionary 154
    Monomer dictionary 153

D

Data
  Customization 149
    Data files 150
      Modification dictionary 154
      Monomer dictionary 153
Decimal places 30

F

File format
  XML 27
Filesystem hierarchy 32
Formula
  Action formula 26

Fragmentation
Pathway
Definition 65
Complex patterns 68
Even more complex patterns 70
Simple patterns 67
Process 102
Accounting cross-links 103
Mass calculation 104

I

IsoSpec library
Isotopic data 124
Isotopes 55
Isotopic cluster
Calculations 123
Centroids 134
Shaping into profile peaks
Mass spectrum 140
Shaping into profiled peaks 136
Generating a cluster 125
Static IsoSpec library data 128
User manual configuration 131
User-modified IsoSpec library data 129
IsoSpec library 124
Isotopic data
IsoSpec library 124

L

Load polymer sequence 74
Load sample polymer sequence 73

M

Modification
Definition 59
Molecular calculator 116
Features 116
Calculator of m/z Ratios 121
Programmable mode 118
Recorder 120

N

Nucleic acids 37

O

Oligomer
  Data
    Filtering 108

P

Polymer chemistry 33
  Cleavage 41, 42
  Definition
    Saving 71
  Definitions 26, 53
    Entities 56
      Cleavage agents 61
      Cross-linkers 60
      Fragmentation pathways 65
      Modifications 59
      Monomers 56
    Isotopes 55
  Fragmentation 42, 45
  Nucleic acids
    Fragmentation 49
  Nucleic acis 37
  Protein
    Fragmentation 46
  Proteins 34
  Saccharides 39
Polymer sequence
  Charge, pKa, pI 109
    Calculation 114
    Ionized groups in monomers 113
    Theory 110
  Composition
    Elemental 109
    Monomeric 109
  Load 74
  Load sample 73
Polymer sequencen
  Search masses 106
Programmable calculator 118
  Recorder 120
Proteins 34

S

Saccharides 39
Sequence editor
  Accounting ionization 77
  Accounting monomers 76
    Cross-links 76
    Modifications 76
  Accounting polymer
    Caps 75
    Modifications 75
  Calculation engine 74
  Editing
    Erroneous monomer codes 81
    Monomer codes 79
  Find motif 83
  Import sequences 83
    Fix failures 84
  Main menu 77
  Oligomer sequence
    Fragmentation 102
  Polymer sequence
    Cross-linking monomers 97
    Monomer modification 88
    Whole polymer modification 96
  Regions 76
    Multi-selection 87
  Selections 76
    Multi-region 85
  Simlified editing 82
Shortcuts 31

T

Threads 30

U

Utilities 147
  Client-server communications
    MineXpert 147

X

XML 27
XpertCalc 115

XpertDef 53
XpertEdit 72