

I2MASSCHROQ USER MANUAL

FREE AND OPEN SOURCE PROTEIN IDENTIFICATION SOFTWARE

version: 1.2.0

Benoît Valot
Thomas Renne
Michel Zivy

Olivier Langella
Filippo Rusconi

CONTENTS

I	GENERALITIES	7
1.1	History of the project	7
1.2	What does <i>izMassChroQ</i> Stand for?	8
1.3	Transitioning from <i>X!TandemPipeline++</i> to <i>izMassChroQ</i>	8
1.4	General concepts and terminologies	8
1.5	Citing the <i>izMassChroQ</i> software.	10
1.6	Installation of the software	11
2	FUNDAMENTALS IN BOTTOM-UP PROTEOMICS	12
2.1	The Protein Biopolymer: Structure and Chemistry	12
	BIBLIOGRAPHY	15

LIST OF FIGURES

FIGURE 1	PEPTIDIC BOND FORMATION BY CONDENSATION	13
FIGURE 2	END CAPPING CHEMISTRY OF THE PROTEIN POLYMER	13
FIGURE 3	PROTEIN CLEAVAGE BY WATER AND CYANOGEN BROMIDE	14

PREFACE

SOFTWARE FEATURE OFFERINGS AND INTENDED AUDIENCE

This manual is about the *i2MassChroQ* protein identification software project.

i2MassChroQ has the following features:

- Load mass spectrometry data files in the mzXML or mzML format, thanks to the excellent *libpwiz* library of ProteoWizard¹ fame.
- Configure the way the peptide/mass spectrum matches (PSM) are to be performed;
- Configure the database files to be used (target organism databases and contaminant databases);
- Use the MS/MS data in the file to feed the *X!Tandem* program that produces peptide identification results by matching the measured ion masses with peptide fragments calculated *in silico* on the basis of the databases contents;
- Perform the protein inference step that leads to reliable protein identifications on the basis of the peptide identifications performed by *X!Tandem*
- Display the data obtained at any step in powerful ways in a unified graphical user interface to allow the user to inspect the peptide identifications and also control the way these identifications are used to infer the protein identifications.
- Export the data after the results exploration above in a variety of formats.
- Perform quantitative proteomics on the basis of the results obtained at the previous steps.
- Perform bio-statistical analyses on the quantitative proteomics data obtained at the previous step.

FEEDBACK FROM THE USERS

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

The PAPPSO software team might be contacted *via* the following contact page:

http://pappso.inrae.fr/en/travailler_avec_nous/contact/ (search for team members having the “Bioinformatics” specialty mentioned, like Olivier Langella or Filippo Rusconi).

¹<http://proteowizard.sourceforge.net/>

PROGRAM AND DOCUMENTATION AVAILABILITY AND LICENSE

The programs and all the documentation that are shipped along with the *i2MassChroQ* software suite are available at <http://pappso.inrae.fr/en/bioinfo/xtandempipeline/>². Most of the time, a new version is published as source, and as binary install packages for *MS-Windows* (64-bit systems only).

For *GNU/Linux*, binary packages are created locally (see <http://pappso.inrae.fr/en/bioinfo/xtandempipeline/download/>) but are also built in the *Debian*² autobuilders and are uploaded to 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* philosophy, the reader is kindly urged to visit <http://www.gnu.org/philosophy>.

²<http://www.debian.org/>

I GENERALITIES

In this chapter, I wish to introduce some general concepts around the *i2MassChroQ* program, the reference to be used to cite the software in publications, the building and installation procedures.

1.1 HISTORY OF THE PROJECT

i2MassChroQ is the successor of the *X!TandemPipeline-Java* project that has seen the following changes along the years:

- Full rewrite of the *X!TandemPipeline-Java* program from Java to C++17. The Java-based software program had been published in Langella, O., Valot, B., Balliau, T., Blein-Nicolas, M., Bonhomme, L., and Zivy, M. (2017). X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification. *Journal of Proteome Research* 16, 494–503. doi: [10.1021/acs.jproteome.6b00632](https://doi.org/10.1021/acs.jproteome.6b00632)



TIP

Before the integrations described below, the product of the rewrite has been called temporarily *X!TandemPipeline++* (or *xtpcpp*). That name might appear in some places while the code/documentation is being revised to change its name to *i2MassChroQ*.

- Integration into the new software of the *MassChroQ* software project that was developed as a standalone C++ software piece. *MassChroQ* is a software project that was developed to perform quantitative proteomics in a variety of modes (label-free or with labelling).
- Unfinalized integration of the *MCQR* project that was developed as a standalone project. *MCQR* is a GNU R project aimed at performing bio-statistical analyses on the quantification analysis performed by *MassChroQ*.

The *i2MassChroQ* project encompasses three main quantitative proteomics fields of endeavour:

- Database search, peptide identification and protein inference. The database search is actually performed by *X!Tandem* and is started seamlessly by *i2MassChroQ*. Protein grouping is performed by original code in *i2MassChroQ*.
- Quantitative proteomics, mainly based on area-under-the-curve processes (requires the full mass data set to extract ion current chromatograms, XIC). This part was historically performed by the *MassChroQ* software program.
- Bio-statistical analysis of the quantification data. This part was historically performed by the *MCQR* GNU R-based package (unpublished software as of yet).

1.2 WHAT DOES *i2MASSCHROQ* STAND FOR?

The *i2MassChroQ* software project aims at providing users with an integrated software solution for quantitative proteomics. As described in detail in another chapter of this book, quantitative proteomics involve a number of steps that can be enumerated in sequence below:

- Search databases to connect MS/MS spectra to peptide sequences. This step is called *identification*;
- Apply logic to reliably identify proteins based on the peptides identified at the previous step. This step is called *inference*;
- Optionally perform quantification of the *identified* peptides and *inferred* proteins. *i2MassChroQ* has area-under-the-curve quantitative proteomics capabilities that are based on precursor peptide ion current extraction from the mass spectrometric data. The extracted ion currents are then plot like chromatograms: intensity as a function of retention time. This analytical process thus somehow involves “*Mass Chromatograms*” for the *Quantification*.

From the sequence above, the `&i2mcq;` name becomes self-explanatory!



TIP

It is however possible (and encouraged) to mentally read *i2MassChroQ* as “*I too MassChroQ !*”

1.3 TRANSITIONING FROM *X!TANDEMPIPELINE++* TO *i2MASSCHROQ*

The previous *X!TandemPipeline++* version of this software did store configuration data in the local configuration directory and in the `PAPPS0/xtpcpp.conf` file. In order to preserve these configuration data after having transitioned from *X!TandemPipeline++* to *i2MassChroQ*, please, rename that configuration file to `PAPPS0/i2masschroq.conf`.

1.4 GENERAL CONCEPTS AND TERMINOLOGIES

This section describes the general concepts at the basis of the analysis of proteomics data that one needs to grok in order to properly assimilate the workings of the *i2MassChroQ* software.

1.4.1 BOTTOM-UP PROTEOMICS OR TOP-DOWN PROTEOMICS?

Proteomics is a mass spectrometry-based field of endeavour that is aimed at characterizing the “protein complement” of a given genome. The protein complement of a genome is the set of proteins that are expressed at a given instant in the life of a cell, a tissue or an organ, for example. Characterizing that protein complement actually means identifying the proteins expressed by a given living cell or tissue or organ. Optionally, if feasible, the characterization of post-translational modifications might be desirable.

There are two main variants of proteomics: “bottom-up” proteomics and “top-down” proteomics:

- The first variant—bottom-up proteomics—identifies proteins on the basis of the identification of all the peptides obtained by first digesting all the proteins of the sample using an enzyme of known specificity. In this variant, the sample that is injected in the mass spectrometer is the resulting peptide mixture (first resolved by high performance liquid chromatography). The identification of the proteins contained in the initial sample is performed in a number of steps that are actually the focus of *i2MassChroQ*. Indeed the *i2MassChroQ* software is a bottom-up-oriented software program.
- The second variant—top-down proteomics—identifies proteins on the basis of intact proteins directly injected in the mass spectrometer. Of course, it might be necessary to fragment the proteins in the mass spectrometer and to use the fragments to actually identify the protein. However, the fact that the protein is first detected and analyzed as one entity (and not as set of peptides), allows for some very useful discoveries, like the identity and number of post-translational modifications, for example.



NOTE

At the moment, *i2MassChroQ* does not handle top-down proteomics data: it is a bottom-up proteomics software project.

1.4.2 TYPICAL CYCLE OF A MASS SPECTROMETER DATA ACQUISITION

Once the initial sample, containing all the proteins to identify, has been digested using a protease of known cleavage specificity (trypsin, typically), the peptidic mixture (that might be highly complex) needs to be resolved as much as possible using chromatography. In the vast majority of the proteomics experimental settings, the chromatography setup is connected to the mass spectrometer so that when the gradient is developed, all the peptides are immediately injected “on line” to the mass spectrum ion source.

The mass spectrometer runs an analysis cycle that can be summarized like the following:

- Acquire a full scan mass spectrum of the whole set of ions at a given chromatography retention time. This kind of mass spectrum is called a MS spectrum;
- Enter a loop during which ions having the most intense signal are subjected in turn to collision-induced dissociation (CID), that is, are fragmented by accelerating them against gas molecules in a fragmentation cell. The mass spectra that are collected at each one of these fragmentation acquisitions are called MS/MS spectra because they are obtained after two mass analysis events: the first event is the measurement of the intact peptide ion’s m/z value (full scan mass spectrum) and the second event is the measurement of all the obtained fragments’ m/z values (MS/MS scan).

Each instrument records all the MS and MS/MS spectra in a raw data format file that is specific of the vendor. Free Software developers cannot know the internal structure of the files. To use the mass spectrometric data, they need to rely on a specific software that performs the conversion from the raw data format to an open data format (mzML). That program is called *msconvert*, from the *ProteoWizard* project.



NOTE

Mass spectrometrists used to call ions that were analyzed in full scan mass spectra “parent ions”. They also used to call fragment ions arising upon fragmentation of a parent ion “daughter ions”. This terminology has been deprecated and has been replaced with “precursor ion” and “product ion”, respectively. In our document, we thus use the new terminology.

1.4.3 OUTLINE OF AN *i2MASSCHROQ* WORKING SESSION

i2MassChroQ loads mzXML- and mzML-formatted files and needs for its operations to have access to all the MS and MS/MS spectra. Once data files have been loaded, *i2MassChroQ* allows the user to perform the following tasks, that will be detailed in later chapters:

- Configure the *X!Tandem* database searching software (that is, the software, external to *i2MassChroQ* that actually performs the peptide-mass spectrum matches);
- Run the *X!Tandem* software and load its results;
- Display the results to the user in a way that they can be scrutinized and checked. The peptide identification results serve as the basis for another processing step that is integrally performed by *i2MassChroQ*: the “protein inference”. That step aims at using the peptide identifications to actually craft a list of proteins identities. The user is provided with various means to control that step in various ways.
- Optionally start the *MassChroQ* module to perform the quantitative proteomics on the identification data checked at the previous step.
- Optionally start the *MassChroQ* module to perform the bio-statistical analysis of the quantitative proteomics data obtained at the previous step.

1.5 CITING THE *i2MASSCHROQ* SOFTWARE.

Please cite the latest article :

Langella, O., Renne, T., Balliau, T., Davanture, M., Brehmer, S., Zivy, M., et al. (2024). Full Native timsTOF PASEF-Enabled Quantitative Proteomics with the *i2MassChroQ* Software Package. *Journal of Proteome Research* 23, 3353–3366. doi: [10.1021/acs.jproteome.3c00732](https://doi.org/10.1021/acs.jproteome.3c00732)

Former citation was :

Langella, O., Valot, B., Balliau, T., Blein-Nicolas, M., Bonhomme, L., and Zivy, M. (2017). X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification. *Journal of Proteome Research* 16, 494–503. doi: [10.1021/acs.jproteome.6b00632](https://doi.org/10.1021/acs.jproteome.6b00632)

1.6 INSTALLATION OF THE SOFTWARE


The installation material is available at <http://pappso.inrae.fr/en/bioinfo/xtandempipeline/download/>.

1.6.1 INSTALLATION ON MS WINDOWS AND MacOS SYSTEMS

The installation of the software is extremely easy on the MS-Windows and macOS platforms. In both cases, the installation programs are standard and require no explanation.

1.6.2 INSTALLATION ON DEBIAN- AND UBUNTU-BASED SYSTEMS

The installation on Debian- and Ubuntu-based GNU/Linux platforms is also extremely easy (even more than in the above situations). ; is indeed packaged and released in the official distribution repositories of these distributions and the only command to run to install it is:

```
$ sudo apt install <package_name> 
```

In the command above, the typical *package_name* is in the form `i2masschroq` for the program package and `i2masschroq-doc` for the user manual package.

Once the package has been installed the program shows up in the *Science* menu. It can also be launched from the shell using the following command:

```
$ i2masschroq 
```

2 FUNDAMENTALS IN BOTTOM-UP PROTEOMICS

This chapter is an optional chapter which the reader might be referred to upon reading other part of this manual.

2.1 THE PROTEIN BIOPOLYMER: STRUCTURE AND CHEMISTRY

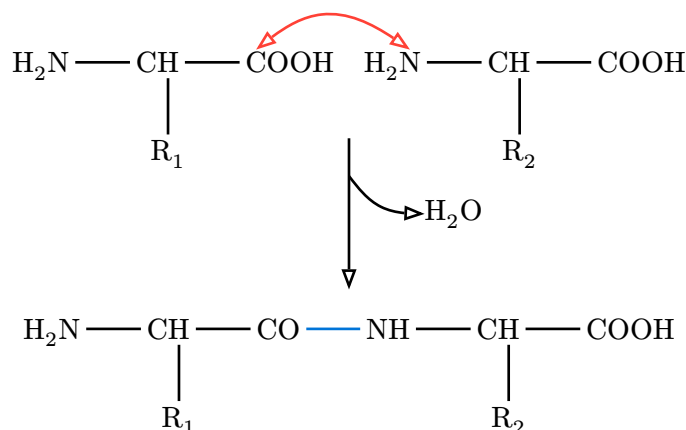
This section introduces the basics in protein polymer chemistry. The way this topic is going to be covered is admittedly biased towards mass spectrometry and proteins. Moreover, the aim of this chapter is to provide the reader with the specialized words that will later be used to describe and explain the (inner) workings of the `&i2mcq;` program. This manual is not a “crash course” in biochemistry.

2.1.1 PROTEIN BIOSYNTHESIS

Proteins 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 to the protein realm.

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 1 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; amino-terminal end* and a right end, called *C-terminus; carboxy-terminal end*. The left end is an amino group ($\text{H}_2\text{N}-$) corresponding to the non-reacted α -amino group of the very first amino acid of the protein sequence. Upon condensation of a new entering amino acid onto the first N-terminal one, the amino group of the entering amino acid reacts (nucleophilic attack) with the α -carboxyl group of the N-terminal 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 un-reacted α -carboxyl group of the last amino acid to have been “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 1, 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).



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.

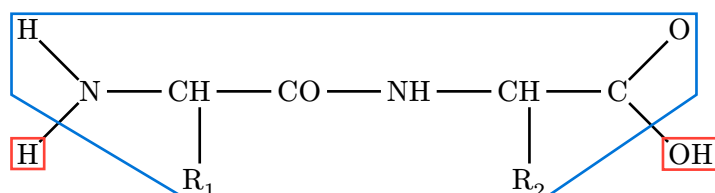
Figure 1: Peptidic Bond Formation by Condensation



NOTE

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 an amino acid has been polymerized into a protein, it is no more called an amino acid, but is called a *residue* instead. 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 2, 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 protein polymer molecule in its “finished state”: by *capping* the left end with a proton *cap* (so as to complete the amino group) and the right end with a hydroxyl cap (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.



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.

Figure 2: End Capping Chemistry of the Protein Polymer

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, its sequence). Of course, proteins have three-dimensional structures, but this is of no interest to a program like *massXpert* (Rusconi, 2009), 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 ($n = 7$)

2.1.2 PROTEIN DISRUPTING CHEMISTRIES

The “polymer chain disrupting chemistry” was mentioned earlier as a complex subject that was of *enormous* importance to the mass spectrometrists. 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; *cleavages* and *fragmentations*:

Cleavages These are chemical processes by which a cleaving agent will act directly on the protein residual chain making it fall into at least two separated pieces (the peptides).



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

2.1.2.1 PROTEIN CLEAVAGE

Upon cleavage of a protein, the cleaving molecule reacts with it, and by doing so directly or indirectly “*dissolves*” an inter-residue bond. A protein cleavage always occurs in such a way as to generate a set of *true* finished polymerization state “proteins” (smaller in size than the parent polymer, evidently, which is why they are called *oligopeptides*, or *peptides*). Indeed, let us take the example shown in Figure 3, 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:

Figure 3: Protein Cleavage by Water and Cyanogen Bromide

BIBLIOGRAPHY

- Langella, O., Renne, T., Balliau, T., Davanture, M., Brehmer, S., Zivy, M., et al. (2024). Full Native timsTOF PASEF-Enabled Quantitative Proteomics with the i2MassChroQ Software Package. *Journal of Proteome Research* 23, 3353–3366. doi: [10.1021/acs.jproteome.3c00732](https://doi.org/10.1021/acs.jproteome.3c00732) 
- Langella, O., Valot, B., Balliau, T., Blein-Nicolas, M., Bonhomme, L., and Zivy, M. (2017). X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification. *Journal of Proteome Research* 16, 494–503. doi: [10.1021/acs.jproteome.6b00632](https://doi.org/10.1021/acs.jproteome.6b00632) 
- Rusconi, F. (2009). massXpert 2 : a cross-platform software environment for polymer chemistry modelling and simulation/analysis of mass spectrometric data. *Bioinformatics* 25, 2741–2742. doi: [10.1093/bioinformatics/btp504](https://doi.org/10.1093/bioinformatics/btp504) 