

Validation of MS/MS identifications and label-free quantification using Proline

Véronique Dupierris, Anne-Marie Hesse, Jean-Philippe Menetrey, David Bouyssié, Thomas Burger, Yohann Couté, Christophe Bruley

► To cite this version:

Véronique Dupierris, Anne-Marie Hesse, Jean-Philippe Menetrey, David Bouyssié, Thomas Burger, et al.. Validation of MS/MS identifications and label-free quantification using Proline. Statistical Analysis of Proteomic Data, 2426, Springer US, pp.67-89, 2023, Methods in Molecular Biology, 10.1007/978-1-0716-1967-4_4. hal-03244240

HAL Id: hal-03244240 https://hal.science/hal-03244240v1

Submitted on 1 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Validation of MS/MS identifications and label-free quantification using Proline

Véronique Dupierris¹, Anne-Marie Hesse¹, Jean-Philippe Menetrey², David Bouyssié³, Thomas Burger², Yohann Couté¹ and Christophe Bruley^{1,†}

¹ Univ. Grenoble Alpes, INSERM, CEA, UMR BioSanté U1292, CNRS FR2048, 38000, Grenoble, France
 ² Univ. Grenoble Alpes, CNRS, INSERM, CEA, FR2048 38000, Grenoble, France
 ³ Institut de Pharmacologie et de Biologie Structurale (IPBS), Université de Toulouse, CNRS, UPS, Toulouse, France
 [†] christophe.bruley@cea.fr

Abstract

In the proteomics field, the production and publication of reliable mass spectrometry (MS)-based label-free quantitative results is a major concern. Due to the intrinsic complexity of bottom-up proteomics experiments (requiring aggregation of data relating to both precursor and fragment peptide ions into protein information, and matching this data across samples), inaccuracies and errors can occur throughout the data-processing pipeline. In a classical label-free quantification workflow, the validation of identification results is critical since errors made at this first stage of the workflow may have an impact on the following steps and therefore on the final result. Although false discovery rate (FDR) of the identification is usually controlled by using the popular target-decoy method, it has been demonstrated that this method can sometimes lead to inaccurate FDR estimates. This protocol shows how Proline can be used to validate identification results by using the method based on the Benjamini-Hochberg procedure and then quantify the identified ions and proteins in a single software environment providing data curation capabilities and computational efficiency.

Key words Mass spectrometry software, Discovery proteomics, Benjamini-Hochberg False Discovery Rate, Statistics, Label-free quantification, Data visualization, Software engineering

1 Introduction

High-throughput mass spectrometry-based proteomics is continuously evolving toward increased complexity through the analysis of larger sample cohorts, with more sophisticated experimental designs, and deeper proteome coverage. This complexity has dramatically increased the data volumes to be processed, so that the processing efficiency has become a major concern for many labs or core facilities. Another effect of this complexity is that the examination of the data and the ability to review and correct the data processing steps requires a solution allowing experimentalists to visualize and navigate a considerable amount of data in a very effective way. To address this, we developed a software suite called Proline [1] which provides an efficient and integrated way to process, visualize and publish proteomic datasets.

The validation process of tandem mass spectrometry (MS/MS) identification in Proline was originally based on predefined filters used to accept or reject a peptide-spectrum match (PSM) and on the widely used target-decoy competition (TDC) method [2] to control the false discovery rate (FDR). We recently proposed an alternative to TDC validation by a totally different method to control the FDR at the PSM, peptide, and protein levels, while benefiting from the theoretical guarantees of the Benjamini-Hochberg (BH) framework [3]. Since version 2.1.2, the BH FDR method is available in the official release cycle of Proline, in addition to the target-decoy competition based method. The BH method provides a simple and pragmatical way to validate identification results without requiring the use of decoy protein databases. Following validation, Proline can then be used to quantify peptides and proteins as shown in this protocol.

2 Material

2.1 Requirements

Proline is a client-server application that can either be installed on a machine running Linux, Windows or MacOS. An all-in-one package called Proline-Zero, including both the client and the server parts as well as the required dependencies is available. These package does not required any complicated installation procedure (*see*

Subheading 2.3). In that particular case, the amount of RAM available is the only requirement: a minimum of 8GB of RAM is recommended for the server especially to be able to run quantification processes; while 1GB of RAM is enough for the client (*see* **Note 1**). This protocol is based on Proline version 2.1.2 installed from the Proline-Zero distribution.

For more intensive usage, the server part of Proline can be installed on a centralized server allowing multiple clients to connect to the same Proline server. In this deployment scheme, Java SE 8 Runtime Environment is required for client and server parts and a PostgreSQL database server (versions above 9.4) must be installed and configured on server side.

2.2 Data format

Identification results from different search engines can be imported into Proline in their native data format. This includes .dat files from Mascot, .omx files from OMSSA, .xml files from X!Tandem and folders of .txt files from Maxquant. In addition, Proline supports mzIdentML files, enabling to import results from any search engines that support this standard format.

2.3 Software install

To install Proline from the Proline-Zero distribution, select the correct archive file for your operating system from the Proline website (http://www.profiproteomics.fr/proline/#downloads) and unarchive it. On first launch, the different components are automatically initialized and configured, including the database schema that will be used by Proline.

2.4 Sample data

This protocol focuses on validation of identification results based on the BH method to control the FDR. To do so, input data must be MS/MS identification searches performed by Mascot search engine on databases containing only target proteins: no decoy is needed using this strategy. Identification data from [3] are used throughout this protocol to illustrate the workflow. This dataset (.dat and .raw files) is available on the proteomeXchange repository [4] with the identifier PXD016669. For the sake of simplicity, the subset of data needed for this protocol as well as the Proline-Zero distribution version 2.1.2 are available at ftp://ftp.cea.fr/pub/edyp/Proline/MiMB/. This FTP folder contains:

- 1. Replicates 1 to 4 of the 10 replicates of samples analyzed with a Q-Exactive Plus instrument with HH settings and searched against a database containing only target proteins will be used (files QEx_HH_no_decoy_R1.dat to QEx_HH_no_decoy_R4.dat).
- The Thermo .raw files corresponding to the MS analysis of these four replicates (QEx2_020296.raw, QEx2_020300.raw, QEx2_020302.raw and QEx2_020419.raw) that will be used in the second part of this protocol.
- 3. The protein sequence database (.fasta file) that has been used to perform the MS/MS search.

3 Methods

3.1 Starting and configuration

- 1. To start Proline, double-click Proline-Zero.exe on Windows or type Proline-Zero.sh on Linux-like OS. The different components are started by the launcher. Afterwards, the graphical user interface process (named ProlineStudio) starts.
- 2. At the first start, the default configuration is initialized and a default user as well as a project are created.
- 3. The connection window appears. Fill the requested fields with the default authentication information (host: localhost and user/password: proline/proline).

The initial configuration creates a default user and a first project but it also generates some predefined descriptions such as instrument configurations or software used to generate the peaklist submitted to the search engine. Once connected, multiple projects can be created and some of these definitions can be modified as described below:

- 4. To create a new project, click on the \ddagger icon in the left panel. In the creation dialog, indicate the name and the description of the project. Click OK.
- 5. Click on the File Admin menu. A dialog appears showing four tabs allowing to: create new users; create a new peaklist software definition; view all existing projects with their associated data; and add new fragmentation rule sets (*see* Note 2).

By default all data used and generated by Proline-Zero are located in a sub-folder of the installation folder named \bigcirc data. For security reasons, the server side is only allowed to browse the content of a restricted list of folders that can however be manually modified by editing the server configuration. By default the folders added to this list are sub-folders of \bigcirc <proline-install>/data. Files must be manually copied into the following folders to be accessible to the server:

- 6. Copy the Mascot .dat files into the <proline-install>/data/mascot sub-folder.
- 7. Copy the fasta file into <proline-install>/data/fasta sub-folder.
- 8. Copy the Thermo .raw files into 🕾 <proline-install>/data/mzdb sub-folder.

3.2 Import search results

Importing a search engine results consists in parsing the search results file to extract information and metainformation and store them into the Proline databases. Neither filtering nor thresholding is applied at this stage: all submitted spectra, peptide-spectrum matches and protein hits are stored in the database, enabling the subsequent validation of putative identifications. As Mascot is a prerequisite to validate identification results using the BH method, this protocol focuses on Mascot identifications results. However, Proline also supports other search engines (*see* **Note 3**) that can be validated through the classical target-decoy approach.

- 1. [Right-click] on the **P** Identifications node in the upper left panel and select [Import Search Result].
- 2. The import dialog appears (see Figure 1).
- 3. Select the files to import by clicking on the \square file button.
- 4. Set the different options:
 - (a) Software engine Mascot: the search engine used to generate the file to import.
 - (b) Instrument Q EXACTIVE (A1=FTMS F=HCD A2=FTMS): the mass spectrometer used to perform the MS/MS analysis.
 - (c) Fragmentation Rule Set ESI-TRAP: The fragmentation rules specified in the search engine. The button on the right can be used to visualize all rules of a specific rule set and new rules can be created through the admin dialog (see Subheading 3.1, Step 5). This information is required to generate theoretical ion fragments and then annotated MS/MS spectra.
 - (d) Peaklist software Mascot Distiller: the software used to generate the peaklist submitted to the search engine. This information is used during the optional quantification step to extract scan number or retention time from the MS/MS spectrum titles. Additional software can be configured in the admin dialog (see Subheading 3.1, Step 5).
 - (e) Decoy No Decoy: the target-decoy strategy used during the search: No Decoy if the search was performed using a database containing only target protein sequences, Concatenated Decoy if the search was performed against a database containing both target and decoy proteins or Software Engine Decoy if the decoy search is performed on the fly by the search engine.

If a wrong fragmentation rule set or a wrong peaklist software is selected at this step, they can be modified afterward.

- 5. Click OK.
- 6. Processes are all executed on the server side of Proline. The graphical user interface communicates with the server by submitting the tasks to be performed and modifies the user interface status while waiting for the tasks completion. The submitted and running tasks can be displayed in the logs tab: submitted tasks are represented by an hourglass icon (^{III}), running tasks by a blue arrow (^{III}), successfully completed tasks by a green tick (^{III}) or a red cross (^{III}) when a task failed.
- 7. Wait for the task completion.

Once imported, the search result appears in the upper left panel under the Identification node. User's actions on these results are reachable by right-clicking on the icon \bigcirc representing them in the left panel. By default a search result is designated by the name of the imported file but can be renamed.

3.3 Combine identification results

Identification results of different MS/MS spectrum searches, for example from replicates or sample fractionation, can be combined or merged to generate a non-redundant list of identified peptides and proteins. This merge can be performed before (on search results) or after validation (on identification summaries) and can be applied recursively, leading in a hierarchical organization of the datasets.

If applied before validation, all PSMs identified by the search engine are taken into account and are combined to create new PSMs into the aggregated dataset. This is useful to combine analytical replicates or fractions since peptides belonging to a single protein could be spread across different fractions. Then the created dataset can be validated as if it were a single search engine identification result.

🗫 Import Search Results	N	×
Files Selection	6	
mascot_data_all\FDR\QEx_ mascot_data_all\FDR\QEx_ mascot_data_all\FDR\QEx_ mascot_data_all\FDR\QEx_ mascot_data_all\FDR\QEx_	HH_no_decoy_R1.dat HH_no_decoy_R2.dat HH_no_decoy_R3.dat HH_no_decoy_R4.dat	
Parameters		
Software Engine :	Mascot	\sim
Instrument :	Q EXACTIVE (A1=FTMS F=HCD A2=FTMS)	\sim
Fragmentation Rule Set :	ESI-TRAP	~
Peaklist Software :	Mascot Distiller	\sim
Decoy Parameters		
Decoy :	No Decoy	\sim
Decoy Accession Regex :		
Parser Parameters		
Subset Threshold : 1.0		
ave Load	🖌 OK 🔀 Car	ncel

Figure 1: Import dialog.

- 1. To create a new dataset, Right-click on the Identification node and select Add Dataset.
- 2. Name the new dataset.
- 3. Each imported file can be reused with no need to import it again. The All Imported node allow to retrieve the already imported files. Right-click on All Imported and then Display List (or double click on All Imported). Select the search results to be merged, drag and drop them in the newly created dataset. The search results are now hierarchically organized: the newly created dataset is the parent of the hierarchy and the dragged and dropped search results are the children.
- 4. To merge the search results, right-click on the parent dataset and select Merge Datasets Union and wait for the completion of the merge task (see Note 4).
- 5. Once the merge is completed, the icon representing the dataset changes to give a feedback of the type of combination the dataset is made of. While \bigcirc represents an empty dataset, \bigcirc is used for an imported result set or a dataset just merged from non validated result sets. The right part of the icon indicates properties before validation whereas the left part indicates properties after validation. A letter is added in the blue half circle to indicate the type of combination that has been made: U for Union and A for Aggregate (see Note 4.). In our case, since the merge is done by union and before validation, the icon changes to \bigcirc .

3.4 Validate identification results

Validation of a search result can be performed at PSM, peptide and protein levels independently. At each level, a set of predefined filters can be combined and applied to accept or reject PSMs or proteins based on a user specified threshold value applied to various properties such as rank, peptide length, minimum score for PSM and minimum score or number of peptides for protein sets. In addition to these filters, a procedure can be applied to limit

Table 1: PSM filters.

Filter	Description
Minimum peptide length	PSMs corresponding to peptide sequences shorter than the cut-off stipulated will be discarded when this parameter is applied.
Pretty rank	This filter is applied after having temporarily joined target and decoy PSMs corresponding to the same query. For each query, target/decoy PSMs are then sorted by score. As in Mascot, a pretty rank is computed for each PSM depending on their ranking: PSM with almost equal score (difference < 0.1) are assigned the same rank. All PSMs with a pretty rank greater than the cut-off specified are discarded.
Single PSM per rank	This filter selects only one PSM per pretty rank, which is already the case when a given pretty rank is associated with a single PSM. When multiple PSMs have the same pretty rank, Proline retains the peptide associated with the protein that has the highest number of MS/MS events. Thus, if this filter is combined with the "Pretty rank" filter, the result obtained should be identical to the result of the "Single PSM per MS query" filter.

the false discovery proportion. This validation step can be done by using the popular target-decoy competition method [2] (*see* **Note 5**) or by using a method based on the Benjamini-Hochberg procedure, as proposed in [3].

- 1. Right-click on the node representing the merged search results and select Validate.
- 2. In the validation dialog (*see* Figure 2), add filters to accept only PSM of rank 1 and length >= 7. Ensure that only one PSM is accepted per MS/MS spectrum (also named a query) by adding a Single PSM per Query filter (details of these filters are described in Table 1).
- 3. Select the BH procedure to control the PSM FDR and set the expected FDR value to 1%.
- 4. Select the |BH| procedure to control the Peptide FDR and set the expected FDR value to 1%.
- 5. Add a filter to invalidate protein set identified without at least one specific peptide (Specific peptides correspond to peptides identifying a unique protein set at the dataset level).
- 6. Select the BH procedure to control the Protein FDR and set the expected FDR value to 1%.
- 7. Choose Fisher as Scoring Type that relies on Fisher's test to calculate protein scores and p-values that will be used by the BH procedure (*see* Note 6).
- 8. Check the Propagate PSM filtering option. When validating a merged dataset, the applied filters as well as the threshold determined by the FDR control method can be propagated to the child datasets of the hierarchy that have been merged. Doing this, the global FDR at the top level of the hierarchy is controlled, but the user can still explore the individual validated search results, relying on the same validation criteria.
- 9. Move on to the Typical Protein Parameters tab. In the Rule Opanel, fill the Typical protein match text field with sp* on Protein Description. A set of identified peptides can match to multiple proteins that cannot be distinguished [5]. This depends on the sample but also on the redundancy of the protein database used for the identification. The typical protein is the protein selected to be the representative of the identified group. The selection can be customized by specifying selection rules. In that particular example, the selection tries to select proteins from the SwissProt database instead of Trembl one if any (see Note 7).
- 10. Click OK and wait for the completion of the validation task. Once validated, the left half-circle is highlighted in orange ((1)) (see Note 8).
- 11. Protein sequences are usually absent from the search engine identification result files. However, Proline can retrieve amino acid sequences from the fasta file (*see* Note 9). Once validation is completed Right-click on the parent dataset and select Retrieve Protein Sequences (*see* Note 10).

3.5 Navigate through identification summaries

MS/MS identification results before and after validation can be selected to display their content. Before validation, the navigation path through the data can start from submitted MS/MS spectrum, PSMs or proteins. After validation, the validated content can alternatively be displayed and the navigation then starts from PSMs, peptides or protein sets. In any case, the selected data are displayed in views composed of one or more tables. The

dentification Validation X	Identification Validation X
Validation Parameters Typical Protein Parameters	Validation Parameters Typical Protein Parameters
PSM	Set Typical Protein Match
Propagate PSM filtering to child Search Results	Using rules (in priority order):
Prefilter(s)	Rule 0
Pretty Rank <= 1	Typical Protein Match : sp* on Protein Description v
AND Length >= 7	advanced RegEx
< Select > v	Rule 1
EDR PSM Eilter	Typical Protein Match : * on Protein Accession v
✓ ensure BH	advanced RegEx
Optimisation based on Score V Basic V	Rule 2
	Typical Protein Match : * on Protein Accession v
Peptide	advanced RegEx
Filter(s)	(* = any string : ? = any character)
< Select > v	
FDR Peptide Filter	
☑ BH ✓ Peptide FDR <=	
Protein Set	
Propagate ProteinSets filtering to child Search Results (Warning FDR Validation will not be propagated !	
Filter(s)	
Specific Peptides >= 1	
< Select > v	
FDR Protein Filter	
BH V Protein FDR <= 1.0 %	
Scoring Type: Fisher	
asve 🚺 Load 🗸 CK 🗶 Cancel 😧	Save Dad V OK X Cancel

(a) Validation parameters tab.

(b) Change typical protein tab.

Figure 2: Validation dialog.

navigation principle is consistent across the graphical interface: the first table contains the data to start from, while the content of the following tables of the view is induced by the selected row in the preceding table. Each table can be searched, filtered and sorted and the column visibility can be modified to focus on the information of interest.

- 1. Right-click on the parent dataset node, select Display Identification Summary Protein sets. The same view as represented in Figure 3 appears in the right panel of the main window. The first table shows the identified protein sets of the identification summary. Each row is a protein set represented by one of its protein. The menu bar on the left allows to search for a specific value in the table (m), to filter rows (T), to modify the display (column visibility, order and width) (\gtrsim) or to export the content of the table to a CSV or XLS file (\square) (see Note 11).
- 2. Select or search for the FDOG_ECOLI protein set in the list of identified protein sets in the first table: Click on the search button, indicate FDOG* in the search popup panel.
- 3. The second and third tables of the view show the content of the selected protein set: the different proteins identified by the same set of peptides (or by a subset of them) is represented in the second table, while the set of peptides identifying the proteins is displayed in the third table (*see* Figure 3). If a protein is selected in the second table, for example a protein set marked as a subset by the icon (2), the set of peptide displayed in the third table is restricted to peptides identifying the selected protein.
- 4. The third table shows all peptides identifying the selected protein. In addition to the sequence and the post-translational modifications associated to each peptide, information related to the highest scoring PSM identifying the peptide are shown. Click on a peptide in the third table to change the selection.
- 5. The protein sequence retrieved by Proline from the fasta file (*see* Subheading 3.4, **Step 11**) is displayed in the last panel at the bottom of the view. Amino acids matched by an identified peptide are represented on a gray background and the selected peptide is highlighted in blue. Underneath the amino-acid sequence, a graphical display of the sequence layed out on a single line shows the position of identified peptides on the sequence.
- 6. Move on to the Spectrum tab representing the MS/MS spectrum submitted to the search engine. Click on

🏷 Pi Eile V	oline (indo	eStudio 2.2 Mile ow <u>H</u> elp	estone (2021-03-09T09:06:53Z_C	3205360)						G						- 0	×
	ogs	× () HH_dat	aset Protein Set 🗙													•	
		73 Id	Protein Set 5 282 SYP_ECOLI	Description sp P16659 SYP_ECO	Score 1480.2	Proteins 27 1	(1 🗖 , 0 🚄)	Peptides	Observable 28	Peptides 33	Spectral Coun	t Spec	ific Spectral C 102	Sequence Count	Mass 26	63 693	
	7	74 75 78	5 261 ALF_ECOLI 5 292 SYN_ECOLI 5 273 SYE ECOLI	sp P0AB71 ALF_EC sp P0A8M0 SYN_EC sp P04805 SYE_ECO	1456.7 1445.8 1432.7	75 1 89 1 73 1	(1 , 0) (1 , 0) (1 , 0)		24 27 26	18 28 27		134 95 99	134 95 99		22 26 24	39 147 52 570 53 816	
		77 78	5 286 NUSA_ECOLI 5 271 FDOG_ECOLI	sp P0AFF6 NUSA_E sp P32176 FDOG_E	1430. 1429.9	14 1 38 2	(1 , 0) (1 , 1 ≤)		29 31	30 56		88 79	88 79		28 31	54 871 112 549	
•		79 80 81	5 269 SYT_ECOLI 5 269 SYT_ECOLI 5 299 PLITA_ECOLI	sp P06720 AGAL_EC sp P0A8M3 SYT_EC solP09546 PLITA_EC	1425. 1425. 1424	15 1 17 1	(1 , 0) (1 , 0)		25 30 30	46		94 105 93	94 105 93		24 27 30	50 657 74 014 143 815	~
2	Ту	ypical Protein:	sp P32176 FDOG_ECOLI Formate	lehydrogenase-O major	subunit OS=Escherich	ia coli (strain	K12) OX=83	333 GN=fdoG PE=1	SV=5								
	1	Protein	Description DLI sp P32176 FDOG_	Sameset /	Subset	Score	1	Peptides		Observa	able Peptides	56	Sequence Count	Ma 31	155	112549.00	₽ ^
	2	FDNG_ECC	sp P24183 FDNG_	ECOLI Fo				54.04		1		66		1		112963.00	Ļ
3	2	Prev. AA	Peptide	Next AA	Length	PTMs		PSM Score	Start		Stop		Calc. Mass	Exp. MoZ		Ppm	R.
	S 1. 1	1 R 2 K	LPSTCFAEENGSIVNSGR GADAPGIALTDGEILSG.	W L		18 Carbam 21	idomethyl (Cl	5)	138.99 113.77		614 639	631 659	1	936.89 085.11	969.45 696.04		^ !
2		3 K 4 K	HALLNAILQPEQFVEIG TASFLYALGWTQHSVG	.т		23 21			106.70 102.49		908 403	930	2	533.35 319.18	845.46 774.07		
		5 K	FLVTIDPLNTETSNFW	I		29			97.75		578	606	3	333.58	1112.20		
		7 R	GHSNIQGLTDLGLLSQ	Q		28			92.47		447	474	2	955.50	986.17		
		8 R	EDYGFEDGLFTGYDAEK	R		17			87.30		314	330	1	954.81	978.42		
	1	9 K	APGSDKWQQISWEEA	W		18			79.91		517	132	2	520.76	811 39		
L	1	I R	SGTDIAFLSGVLLYLLN	F		20			77.81		277	296	2	166.16	723.06		
L	1	12 K	GAGLVDFIHSESR	L		13			73.37		95	107	1	386.69	463.24		
I.	1	13 R	IFKDDAEALGK	A		11			68.36		875	885	1	205.63	402.88		~
H																	
	Prote	MQVSRRQFF RDANYIAQN IAFLSGVLL NIRTMAMIC	pectrum Spectrum Error Fragmi K ICAGGMAGTT AAALGFAE IA EGVTVNRWLS TGMLCASA Y LLINBERPIRE YTEAYTMA IL LLGNMGMAGG GVNALRGE	NTATION TABLE SPECTRU SV ALAETRQYKL I SS NETGYLTQKF S SL IVREDYGFED (SN IQGLTDLGLL S	n Values LRTRETRNTC TYC SRALGMLAVD NQA GLFTGYDAEK RKY SQSLPGYMTL PSE	SVGCGLL RVUHGPT DKSSWTY	MYSLGDG VASLAPT ELDENGF YLTANTP	AKN AKASIFHI FGR GAMTNHWV AKR DTTLQHPR KPL LEGQVNYW	EG DPDHPVNI DI KNANLVV CV WNLLKQHV GN YPKFFVSM	RGA LCPH /MG GNA/ /SR YTPI /MK AFF(RGAGLVD F AEAHPVG FI DVVENI <mark>C</mark> G' GDRATAE N	IHSESRLKF RWAMEAKIH FPKDAFLKV SWGFDWLPK	PEYRAPGSDK NGAKLIVIDP CEYIAETSAH WDKGYDVLQY	WQQISWEEAF RFTRTAAVAD DKTASFLYAL FEMMKEGKVN	DRIAKLMK YYAPIRSG GWTQHSVG GYICQGFN	SD FD AQ	^
		ASFPNKNKV ALADITDPA KMAEGPFPE GVAKKGFIA	I GCLSKLKFLV TIDPLNTE T GAVIVKKGQQ LSSFAQLE H YEPFETPLGT NPLHPNVI N TLTPFVGDAN TQTPEFKS	TS NFWQNHGELN I DD GTTSCGCWIF J SN PAARIFKDDA I FL VNVEKV	EVDSSKIQTE VFF AGSWTPEGNQ MAF EALGKADKFP YVC	LPST <mark>CFA</mark> RDNADPS TTYRLTE	EENGSIV GLGNTLG HFHYWTK	NSG RWLQWHWR WAW AWPLNRRI HAL LNAILQPE	GA DAPGIALT LY NRASADPO QF VEIGESL	TDG EILS QGN PWDI ANK LGIA	SGIFLRL RI PKRQLLK WI AQGDTVK V	KMYAEQGGA OGTKWTGWD SSNRGYIKA	NPDQVLNMTW IPDYSAAPPG KAVVTKRIRT	NYAIPHEPSS SGVGPFIMQQ LKANGKDIDT	EEVAMESN EGMGRLFA IGIPIHWG	GR LD (E	
		N/C-te	r PTM 🗧 AA PTM	N/C-ter and	d AA PTM												
		Coverage FDOG ECOLIF	: 48.82% Protein Seguence Coverage 1016 a	nino acid													~
						-				Ð						_	-
Ľ																	ŀ

Figure 3: Navigation through identification summary data.

the Generate and Store Spectrum Match icon (\square) on the left. In the dialog, select ESI-TRAP as the fragmentation rule set and click Ok. The fragments matching the theoretical spectrum generated using the specified fragmentation rules are displayed over the MS/MS spectrum (*see* Figure 4). Move on to the Fragmentation table tab to display the fragment matches in a tabular view (*see* Note 12).

7. Proline systematically stores and keeps track of metadata from processing steps, used parameters and generated data. These metadata are available in different views and export outputs. Metadata from multiple dataset can be easily displayed, allowing to compare them (*see* Figure 5). Select the four replicates in the left panel (left-click on the first dataset, hold the Shiff key and then left-click on the last replicate) then right-click and select Properties. In the Properties view, each row is a property or a metadata (grouped by theme) and each column represents a dataset. In the second column (reffered to as Type) the property name is displayed on with an orange-light color if the values in the row are different. As an example, in the Search properties group, the Result File Name is different in each dataset but the other search parameters are all identical.

3.6 MS1 label-free quantification

Once validated, identification summaries can be used for label-free quantification, either using spectral counting, or after detection of chromatographic peaks from MS1 signals. The different steps of the quantification process, the algorithms implemented in Proline and their parameters have been extensively described in [1]. In this protocol, the focus is made on how to start a MS1 quantification from an identification summary.

- The signal extraction is based on the mzDB format [6]. Raw files must be first converted by using the conversion tool which is embedded in Proline-Zero (*see* Note 13). In the left panel move on the MS Files tab. Browse the local folders to the Proline installation path, then into the coversion/coversi



Figure 4: Annotated MS/MS spectrum and fragmentation table.

- 3. Click OK and wait for the end of the conversion task (see Subheading 3.2, Step 6).
- 4. Move back to the Projects tab and select the parent dataset. [Right-click] on the dataset and select Quantify Label-Free (see Note 14).
- 5. In the dialog (see Figure 6a), the first step is the experimental design definition. Groups and samples corresponding to conditions to compare and to biological samples can be created manually by right-clicking on the Quant node. However, a more automated solution is available: drag and drop the parent dataset node from the right panel into the left panel, on the Quant node. An experimental design is automatically created and appears under the sample node (in this case, it contains a single group, a single sample and the four replicates). Right-Click on the Quant node then select Rename. Modify the default dataset name and click OK. Click Next.
- 6. In step 2 (Associate MS files to sample analyses), mzdb files must be associated to their respective datasets (*see* Figure 6b). In the right panel, browse the mzdb_files folder to display the four mzdb files. Select these files (Shift+left-click) and drag and drop them into the panel named Drop Zone. Proline takes advantage of the available metadata, especially the name of the peakelist associated to each dataset to match the peaklist name and the name of the mzdb files dropped in the Drop Zone. In this case, the names perfectly match so that there is nothing more to do. Click Next].
- 7. The last step consists in setting the parameters used during the quantification process (*see* Figure 6c). By default the dialog shows a simplified set of parameters with default values suitable for high resolution mass spectrometry analysis: the mass over charge (moz) tolerance is set to 5.0 ppm, the alignment process is performed by using a maximum shift of retention time of 600.0s and the cross assignment is enabled between all runs with 60.0s tolerance to match predicted retention time (*see* Note 15).
- 8. Click OK to start the quantification and wait for the completion of the quantification task.
- 9. Once completed, the quantification dataset node icon changes to 🐱, the dataset is ready for browsing and navigation by right-clicking on this icon.
- 10. By default, peptide ion measurements are summarized as protein abundances using a simple sum operation. However, additional operations such as excluding peptides or ions based on their characteristics (missed cleavages, variable modifications, sequence specificity, etc.) or normalizing peptide and protein abundances between runs can be performed. These post-processing steps can be executed on-demand using different parameters or methods; there is no need to repeat the whole quantification process when changes are made. Right-click on the quantification dataset node then select Compute Post Processing on Abundances. Choose how ions and peptides abundances are combined into peptide and protein abundances in the parameter dialog and click OK. Wait for the task completion.

3.7 Navigate through quantification datasets

1. Select the created quantification dataset and right-click then select Display Abundances Protein Sets. The predefined view opens (*see* Figure 7). This view starts from the quantified list of protein sets shown in the first table. Below this table, the quantified peptides of the selected protein set are represented in a tabular display on the left hand side; and in a graphical view on the right one. The graphical view represents the

dov	v <u>H</u> elp				4	
	× 🕕 HH_dataset Protein Set 🛛 🗴	Properties x				٩
	Group	Туре	QEx_HH_no_decoy_R1	QEx_HH_no_decoy_R2	QEx_HH_no_decoy_R3	QEx_HH_no_decoy_R4
1	General Information	Raw File Name	QEx2_020296.mgf	QEx2_020300.mgf	QEx2_020322.mgf	QEx2_020419.m
3		Fasta Files	/local/mascot/mascot_2.6//sequence/UP_K1	/local/mascot/mascot_2.6//sequence/UP_K1	/local/mascot/mascot_2.6//sequence/UP_K1	/local/mascot/mascot_2.6//sequence/UP_K1
4		Search Result Name	YOC_SP-Tr_Cerevisiae	YOC_SP-Tr_Cerevisiae	YOC_SP-Tr_Cerevisiae	YOC_SP-Tr_Cerevisi
6		Instrument Name	Q EXACTIVE (A1=FTMS F=HCD A2=FTMS)	Q EXACTIVE (A1=FTMS F=HCD A2=FTMS)	Q EXACTIVE (A1=FTMS F=HCD A2=FTMS)	Q EXACTIVE (A1=FTMS F=HCD A2=FTM
6		Fragmentation Rule Set	ESI-TRAP	ESI-TRAP	ESI-TRAP	ESI-TR.
7		Target Decoy Mode				
8	Search Properties	Peaklist Software	Mascot Distiller	Mascot Distiller	Mascot Distiller	Mascot Disti
10	Search Properties	Result File Name	QEx_HH_no_decoy_R1.dat	QEx_HH_no_decoy_R2.dat	QEx_HH_no_decoy_R3.dat	QEx_HH_no_decoy_R4.c
11		Search Date	8 février 2019	8 février 2019	8 février 2019	8 février 20
12		Software Name	Mascot	Mascot	Mascot	Maso
13		Software Version	2.6.0	2.6.0	2.6.0	2.6
14		Taxonomy	All entries	All entries	All entries	All entri
15		Enzyme	Trypsin/P	Trypsin/P	Trypsin/P	Trypsin
16		Max Missed Clivage	2	2	2	
17		Fixed Modifications	Carbamidomethyl(C)	Carbamidomethyl(C)	Carbamidomethyl(C)	Carbamidomethyl
18		Variable Modifications	Acetyl(Protein N-term), Oxidation(M)	Acetyl(Protein N-term), Oxidation(M)	Acetyl(Protein N-term), Oxidation(M)	Acetyl(Protein N-term), Oxidation(
19		Fragment Mass Tolerance	25.0 mmu	25.0 mmu	25.0 mmu	25.0 m
20		Peptide Charge States	2+ and 3+	2+ and 3+	2+ and 3+	2+ and 3
21		Peptide Mass Error Tolerance	10.0 ppm	10.0 ppm	10.0 ppm	10.0 p
22		Fragment Charge States	2+ and 3+	2+ and 3+	2+ and 3+	2+ and 3
23		Fragment Mass Error Tolerance	25.0 mmu	25.0 mmu	25.0 mmu	25.0 m
24	Search Result Information					
25		Queries Number	26604	26520	26103	245
26		PSM Number	13469	13499	13312	146
27		Protein Number	1887	1864	1857	20
28		PSM Decoy Number				
29		Protein Decoy Number				
30	Identification Summary Information	Destain Cata Number	1007	1205	1201	10
51		Protein Sets Number	1307	1295	1291	12
		Pontide Number	10345	10443	10451	109
		Protoio Soto Docou Number	0104	0123	0104	00
24		Protein Sets Decoy Number				
26		Pentide Decov Number				
7	Validation Parameters	replace beeby Hamber				
38	Tailed dort 1 ai aine tera	protein filters / description	protein set filter on specific peptide (protein	protein set filter on specific peotide (protein	protein set filter on specific peotide (protein	protein set filter on specific peptide (protein
39		protein filters / parameter	SPECIFIC PEP	SPECIFIC PEP	SPECIFIC PEP	SPECIFIC P
0		protein filters / properties / thres	1	1	1	
41		protein scoring	proline:fisher score	proline:fisher score	proline:fisher score	proline:fisher sco
42		psm_filters#1/description	peptide match rank filter	peptide match rank filter	peptide match rank filter	peptide match rank filt
43		psm_filters#1/parameter	PRETTY RANK	PRETTY RANK	PRETTY RANK	PRETTY RA
14		psm_filters#1/properties/thresh	1	1	1	
45		psm_filters#2 / description	peptide sequence length filter	peptide sequence length filter	peptide sequence length filter	peptide sequence length fill
46		psm_filters#2 / parameter	PEP_SEQ_LENGTH	PEP_SEQ_LENGTH	PEP_SEQ_LENGTH	PEP_SEQ_LENG
47		psm_filters#2 / properties / thresh	7	7	7	
48		psm_filters#3 / description	peptide match score filter	peptide match score filter	peptide match score filter	peptide match score filt
49		psm_filters#3 / parameter	SCORE	SCORE	SCORE	SCO
50		psm_filters#3 / properties / thresh	20.77	20.77	20.77	20.
51	Validation Results					
52		protein_results / target_matches	1307	1295	1291	12
53		psm_results / target_matches_count	10543	10445	10451	109
	Sal Ids					
54						
64 56		Project id	2	2	2	

Figure 5: Datasets properties.

quantification profile of each peptide of the protein set (the abundance value in each MS analysis) as well as the protein set abundances (in yellow) calculated from these peptides on a second vertical axis. When a peptide is selected in the table, the graphical representation of the peptide is highlighted accordingly. The bottom panel shows the ions that have been quantified for the selected peptide (Quanti. Peptide Ions tab). The XIC Features tab shows the signal extracted from each MS analysis (called a feature) while the right hand side panel gives a graphical display of these features.

- 2. Click on the Display all isotopes icon (III) to display the signal extracted for each isotope of a currently selected feature. Click again to go back to the previous display showing all features.
- 3. The graphical display uses the identified features stored in the database. However a chromatogram can be extracted from the mzdb files, even if Proline was unable to assign a signal abundance to the selected ion in some MS analyses. Select a peptide (For example, peptide IFNADWVIDGEQQPK of protein PUR4_ECOLI), then an ion for which an abundance value is missing, then move on to the XIC Features. The table shows the abundance values extracted for each MS analysis. If Proline fails to find an abundance value for an MS analysis, the corresponding row is filled with zero values. Righ-click on any row then select Extract All XIC. A task is created, requesting the extraction of the ion m/z chromatogram in all MS analysis. Once the response is received, the chromatograms are visible in the graphical display as dotted lines overlayed on the existing features.
- 4. In addition to abundance values, the alignment between MS analyses, which is a critical step of the label free quantification process, can be visualized. Right-click on the quantification dataset (b) then select Display Exp. Design Map alignment. The upper panel of the view shows the alignment curve between two selected MS analyses represented as the time difference (in seconds) between these two MS analysis over the overall analysis time (in minutes). The lower panel shows the alignment curves of all MS analysis with an analysis choose as the alignment reference.
- 5. In both panels the alignment curves give a general trend of the retention time shift. However in the upper panel the retention time difference of each individual ion can be represented as a scatter plot. Click on the *^{ext}* icon. After a few seconds, the Loading Data tool-tip disappears and the points representing the ions are shown overlayed on the curve. Filled circles represent ions that have been identified in

Step 1: Define the experimental design		
Drag and Drop identification summaries from the right p	anel to build the experimental design on the left.	
iental Design		
IH quantification	() Hi_dataset	
Identification reference : HH_dataset	QEx_HH_no_decoy_R1	
Group HH_dataset	QEX_HH_no_decoy_R2	
DEv HH no derov R1	OEV HH no deroy R4	
QEx HH no decoy R2	(
QEx_HH_no_decoy_R3		
QEx_HH_no_decoy_R4		
	- Nevt 💓 (ancel



(a) Experimental design.

Quantitation Wizard	N		×
Step 3: Specify	quantitation parameters.		
Advanced Parameters			
Quantitation Parameter			
Moz tolerance (ppm): 5)		
Cross Assignment			
Allow cross assignment	Between all runs	×	1
RT tolerance (s):	60.0		
Align RT			
Time tolerance (s):	600.0		
🖬 Save 🕕 Load		🗲 Back 🖌 OK 🗶 Cancel	0

(b) MzDB files association.

(c) Label-free parameters.



both MS analyses while black circles represent ions that have been cross-assigned in one of the two MS analyses (cross-assigned ions can be removed from the display by clicking on the solution). The upper and lower curves added to the alignment curve indicate the retention time tolerance used to perform the cross-assignment. Visualizing this information enables the user to adjust the retention time tolerance used during the cross-assignment step to reduce the likelihood of errors or, conversely, to take into account the retention time discrepancy between the MS analyses. To modify only this single parameter, relaunch the quantification by right-clicking on the quantitation dataset then select Clone & Extract Abundances: the quantification parameters dialog appears (*see* Subheading 3.6, Step 5) pre-filled with the experimental design and parameters of the selected quantification dataset. The parameter of interest can then be modified leaving all other parameters unchanged.

3.8 Customized Graphical display

Any dataset in Proline can be browsed using different predefined views, starting from an initial list of objects: PSMs, peptides, protein sets, MS/MS queries, etc. Additionally, new views can be created dynamically and saved for future browsing.

- 1. Right-click on the quantification dataset and select Display Abundances New User Window. The list of objects that can be used as starting point are shown. Select Quanti Protein Sets and click OK. A new view appears, composed of a single table showing the list of quantified protein sets (*see* Note 16).
- 2. Click on the 🕈 icon in the right border of the panel to add a new panel into the view. A list of panels that are linked to the protein sets table is shown as well as an option to organize the panel into the view. Select the Customisable Graphical Display panel and add it Below the protein sets table.

Ms files 🛛	Logs	× 👷 HH_dataset P	Protein Sets 🗙	🛃 HH_dataset Prote	sin Sets x											4
Project 🗸 🗈 🖉 🔹	H T	Protein Set	Overview	Description	#Peptide	#Quant. Peptide	 #Quant. PSMs QEx_HH_no_decoy 	R1 QE:	bundance <_HH_no_decoy_R1	#Quant. PSMs QEx_HH_no_decoy_R2	 Abundano QEx_HH_no 	ce o_decoy_R2	 #Quant. PSMs QEx_HH_no_decoy_R3 	 Abundance QEx_HH_no_decoy_R3 	#Quant. PSMs QEx_HH_no_decoy_R4	Abundance QEx_HH_no_decoy_R4
rations	2	1 a PUR4_ECOLI		sp P15254 PUR		66	65	50	579 523 264		42	555 565 952	45	489 658 624	5	1 547 422 33
Imported		2 DNAK_ECOLI		sp POA6Y8 DNA		57	55	50	1 247 394 048		42	1 237 749 376	40	1 126 337 024	4	1 111 642 88
datarat		3 TNAA_ECOLI		sp P0A853 TNA		64	64	97	8 507 758 592		21	8 235 215 872	100	6 752 651 264	10	5 7 880 962 56
dataset	1	+ _ EFG_ECOLI		sp P0A6M8 EFG		57	55	59	2 739 996 672		49	2 651 903 488	50	2 377 600 512	6	2 154 487 80
Hi no danay D1	0	ACINE_ECOLI		spiP30003 WCN		01	50	39	040 001 904		99	202,496,209	4/	200 2940 2004	3	702 000 73
HH no decov R1	-	7 A CLPB ECOLI	10.0	spiP63284ICLP		58	58	41	809 736 192		44	832 751 232	44	674 829 824	5	724 031 23
175	_	8 RPOC ECOLI	1.0	spiP0A8T7IRPO		60	58	38	286 162 432		37	300 355 136	35	232 307 360	4	236 245 4
		0 RS1_ECOLI		sp P0AG67 RS1		40	39	32	554 103 168		36	520 716 672	31	448 470 816	3	5 466 177 0
		10 CARB_ECOLI		sp P00968 CAR		52	51	43	451 107 136		30	422 210 912	35	370 395 392	4	356 719 9
		11 🔓 PFL8_ECOLI		sp P09373 PFL		49	48	36	665 617 536		32	664 861 952	41	540 329 408	4	2 565 996 2
		12 PUR9_ECOLI	1111	sp P 15639 PUR		42	41	32	1 238 909 184		35	1 239 925 760	33	1 044 172 096	4	1 157 846 2
		13 PT1_ECOLI		sp P08839 PT1		41	41	36	539 425 088		29	525 734 816	34	444 863 328	3	436 960 9
		14 STV_ECOLI		sp P07118 SYV		48	48	34	384 435 264		30	397 902 144	33	322 020 864	4	2 365 514 4
		16 GH60_ECOLI		sp POA6PS [CH6		35	35	29	730 171 264		29	863 501 888	25	606 148 800	3	5 770 328 33
		10 Lo Briot Jecoti		spiPuce4/jeP1		39	39	187	9 546 343 424	1 .	na	11 319 337 984	1/5	8 /38 0/5 648	20-	+ 1131/6/60.
	m	Charles .	Description Co.			C	drawn			Durk	Cal Cause (B)	🕨 Graph	ic : Linear Plot 🗸 X Axis	: Quant. Channel 🗸 Y	Axis: Abundance 🗸	
	Ŧ	Status	replace of	equence Prins		Source	charge	11/2	N.	Prote	n set court te	1 . 14	-84			
	× 2	5 4	HANPCGVA	AIGNSILDAYD Carbami	domethyl (C5)		160.17	2	1021.9897	33.49	^					-1.
	1 1 1		ALSYNNIAL	OTDAALECVK Carbami	domethyl (C1		75.08	4	803.6440	45.47		- L2	18 -			
			DEGMIGAE	ELR Dividente	- 0.03		52.50	2	595.2986	30.44		.833				-1
	1 °		ENDONECS	AUGUMOR UX0810	n (MS)		75.98	2	735.3470	23.03		1.0	E8 -			
		0	YGENSHOC	DAAFYIEEN			62.56	4	850, 1632	34.48		× s	-	0		- I - 18,
	0	6	DLGWVGAE	ELR Oxidatio	n (M4)		48.08	2	603.2951	25.53		1 1 1 1 1			0	
	3	2	GQDDAIME	EHQIQPID			40.16	3	1176.5813	52.55		- T	7-			
	3	3 🖒	YGENSHQQ	AAFYIEENVK			60.12	3	709.6635	27.39		📼 🕏				
al sent	3	4 🥌	ELDAETAQ	AIISR			97.76	2	708.8702	30.80		4.0	57-			
000000	3	6 C	EGCSLEDA	VENID1GGPTMCarbam	domethyl (C3)		157.40	2	1131.5236	40.65					/	
	3	0 05	AFEHTAAY	DS IANTEGS Oxidatio	n (M11)		39.48	4	778.3460	39.27		2,0	7			
	3		CUELLETCO	Carbam	comethyl (CS)	_	25.85	2	470.2239	35.79	_			_		
			GSSMASDA	FEPER Oxidatio	o (M4)		88.64	2	718,3187	37.92	~	0,0	05.181.00.000	05-101-0-1000	07. 181 av dame \$2	-0
		<									>		QEX_PHI_NO_GREEV_3G	QEI_HH_NO_GROOV_N2	QEX_HPI_NO_GREEV_JRS	QEI_HH_N0_Gecoy_K4
												- 1		Quant. C	hannel	
	Quar	nti. Peptides Ions XI	C Features						k	Graphic : Linear Plot	V X Axis: F	RT v Y	Axis: Intensity \lor			
	Featu	res (4)							1	1.260						
	Ψ	Мар	Quant. C	thannel m/z		Charge	RT	Ape	x Intensity 🛱 🖉							
	X	1 0Ex2 020296	OEX H	H no decov	708.870	12	2	30.85	99 🔊 🗐	1.068 -						
		2 QEx2_020300	QEX_H	H_no_decoy	708.870	18	2	30.80	93				\mathcal{M}			
		3 QEx2_020322	QEX_H	H_no_decoy	708.869	13	2	30.95	81	8.067			(Λ)	Λ		
		4 QEx2_020419	QEx_H	H_no_decoy	708.870	11	2	30.58	76 📰	2		\wedge	// /			
										E 6.0E7 -	/	\	1 / 1	14		
	1									ê		\	$I I \Lambda$			
										4.0E7 -	/		(1		
											/	\rightarrow				
										2.067 -	/	V	X			
											/	- N	/ / / / /			
											()	. / >				
											(

Figure 7: Navigation through quantification dataset.

- 3. In the graphical view, set the Graphic type to Scatter Plot, X Axis to Abundance QEx_HH_no_decoy_R1 and Y Axis to Abundance QEx_HH_no_decoy_R2. As expected, the two replicates are similar so that abundances are linearly correlated. Right-click on the X Axis and select Log10 Axis. Do the same with the Y Axis. To select points in the scatter plot where the abundances are different between replicates, click on the selection tool icon (iii) to select a rectangular region (*see* Note 17) containing these points (*see* Figure 9). Selected points are then framed in black.
- 4. The selection can be transferred between tables and graphical display. To select the protein sets corresponding to the selected points in the scatter plot, click on the Export selection icon (). The corresponding rows are selected in the protein sets table, which can then be filtered to display only the content of the selection: Right-click on the first selected row then select View Selected Data. To return to the original display right-click on a row and select View All Data.
- 5. Add a new table to the view by clicking on the \pm icon. Select the Quanti Peptides panel and add it Below. The content of this panel is now synchronized with the selected protein set allowing to display the peptide quantification underlying the protein calculated abundance.
- 6. Select a protein from the first table: the panel containing the peptides quantification information is updated accordingly, showing the quantified peptides belonging to the selected protein. Transfer this selection to the graphical display by clicking on the (=): the corresponding point in the scatter plot is then selected.
- 7. This view can be saved to be reused with any other quantification dataset: Click on the icon in the left border of the panel to save the view. Choose a name for this view and click OK.
- 8. Right-click on the quantification dataset and select Display Abundances: the new view now appears in the sub-menu next to the existing predefined views.

3.9 Export data

Identification and quantification results as well as metadata related to the different processing steps performed by Proline are stored in the Proline database and can be exported into various formats. MSExcel (.xlsx) and text (.csv) output formats contain data at PSMs, peptide ions, peptides and protein sets levels that can be customized before export. In addition, results can be exported in standard-compliant (mzIdentML) formats for publication of data in public repositories such as PRIDE and ProteomeXchange.

- 1. Right-Click on a dataset and select Export Excel.
- 2. Indicate the output file name.
- 3. Select Excel (.xlsx) as Export type.



Figure 8: View MS analyses alignments.

- Click on Custom options to choose the sheets, columns and column names that must be exported (Figure 10). Each exported sheet is represented by a tab in the customization dialog that can be unchecked to be removed from the exported file.
- 5. Click Export.
- 6. In addition to the identified PSM, peptides and proteins, Proline offers more specialized exports. Export Sequence Fasta exports a fasta file corresponding to the identified sequences.
- 7. Export Spectra List exports a spectral library containing the list of identified peptides with their precursor mass, their observed fragment masses and their retention time. This export can be achieved from identification or quantification datasets, however, MS/MS spectrum annotation must be generated beforehand (by Right-click on the dataset in the left panel and select Generate Spectrum Matches). Mainly used for Data Independent Acquisition (DIA), this output can be formatted to be compliant with PeakView or Spectronaut software (see Note 18).
- Finally, the whole dataset can be exported in mzIdentML format for publication in ProteomeXchange.
 Right-click on the parent dataset icon and select Export MzIdentML. Fill the different fields with the needed administrative data and click Next. Set the output file path and click OK.

4 Notes

- 1. These values obviously depend on the data size.
- 2. To modify these settings, use an administrator user profile (user: admin and password: proline).

3. Results files produced by Mascot, X! Tandem, OMSSA and Andromeda search engines can be imported into Proline in their native format.

4. Combining datasets can be done in two different ways: Union duplicates all PSMs from the child datasets and creates a copy in the parent dataset; this is similar to what can be obtained by merging peak lists before submission to the search engine. Aggregate selects a representative PSM among PSMs matching the same peptide (same sequence and same post-translational modifications).

5. In this case, the search must have been performed against a target-decoy database (*see* Subheading 3.2) while in this protocol a classical search using a target database is sufficient. In the case of using a target-decoy strategy, any identification result performed against a target-decoy database from any search engine supported by Proline can be used. In addition, search results using the decoy option available in Mascot can also be imported and validated as a target-decoy search.

Image Project Image Project Project <th></th> <th>Logs</th> <th>× 🛃 HH_dataset Qua</th> <th>nti Protein Sets 🗙</th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th>•</th>		Logs	× 🛃 HH_dataset Qua	nti Protein Sets 🗙							•
Initiations Image: Control of the second	ine_Project 🗸 🕨 🕨	M T	Protein Set	Overview	Description	#Peptide #Qu Peptide	iant. • # ide QE	#Quant. PSMs x_HH_no_decoy_R1	Abundance QEx_HH_no_decoy_R1	#Quant. PSMs QEx_HH_no_decoy_R2	Abundance Ex_HH_no_decr
Amadasa Ali montal Mit. Gataset: Mit. Gat		2	1 👌 PUR4_ECOLI		sp[P15254]PUR4_EC	66	65	50	579 523 264	42	555 56
Al imported by the dataset in the da	nuncations		2 DNAK_ECOLI		sp P0A6Y8 DNAK_EC	57	55	50	1 247 394 048	42	1 237 74
Interview image: product	All imported		3 🔓 TNAA_ECOLI		sp P0A853 TNAA_EC	64	64	97	8 507 758 592	121	8 235 21
Trian	HH_dataset	12	4 GEFG_ECOLI		sp P0A6M8 EFG_ECO	57	55	59	2 739 996 672	49	2 651 90
a R02 ECOL1 a ppAdx2 PR04 EC S8 57 43 305 26 666 35 302 7 a R02 ECOL1 a ppAdx2 PR04 EC S8 S8 41 807 7.6 692 44 302 7 303 7 a R02 ECOL1 a ppAdx7 PR02 EC S8 S8 41 807 7.6 692 44 302 7 303 7 a R02 ECOL1 a ppAdx7 PR02 EC S9 S8 284 162 922 307 7 303 7 a R02 ECOL1 a ppAdx7 PR02 EC S9 S9 28 284 162 922 307 7 303 7 a R02 ECOL1 a ppAdx7 PR02 EC S9 S9 S9 7 44 53 98 162 92 59 12 257 7 a R04 ECOL1 a ppPAdx7 PR02 EC S9 S9 7 79 13 12 54 30 97 97 12 54 30 97 97 12 54 30 97 97 12 54 30 97 97 12 54 30 97 97 11 354 30 97 97 11 54 30 97 97 12 54 30 97 97 12 54 30 97 97 12 54 30 97 97 13 95 13 97 12 95 45 33 2 30 37 55 13 97 12 95 45 33 2 30 37 55 13 97 12 95 45 33 2 30 37 55 13 97 12 95 45 33 2 30 37 55 13 97 12 95 45 33 2 30 37 55 13 97 12 95 45 33 2 30 30 37 55 13 97 12 95 45 33 2 30 30 37 55 13 97 12 95 45 33 2 30 30 37	Trash		5 ACNB_ECOLI		sp P36683 ACNB_EC	61	58	39	848 881 984	44	797 72
Trach			6 RPOB_ECOLI		sp P0A8V2 RPOB_EC	58	57	43	305 526 496	35	303 46
a Bit CCCL as pPAddT /PROC_EC 60 58 38 255 40 33 66 56 50 70 a Bit SCCL as pPAddT /PROC_EC 52 51 49 49 1107 156 30 4222 a PPAddT /PROC_EC 52 51 49 49 1107 156 30 4222 66 65 57.55 52 666 56 56 57.55 52 666 56 56 57.55 52 666 56 56.56 52 22 66 66 56 56 57.55 52 666 56 56 56 57.55 52 666 56 56 57.56 52 666 56 56 57.56 52 666 56 56 57.67 57.67 56 52 67.68 29 52.50 57.67		E	7 CLPB_ECOLI		sp P63284 CLPB_EC	58	58	41	809 736 192	44	832 7
antilators a generative for the formation of			8 RPOC_ECOLI		sp P0A8T7 RPOC_EC	60	58	38	286 162 432	37	300 3
Image: State Plot XAxs: Abundance (Ex.JHLno.decoy,R1 v VAxs: Abundance (Ex.JHLno.decoy,R1 v VAxs: Abundance (Ex.JHLno.decoy,R1 v 2488 31 49 <			9 RS1_ECOLI		spiPUAG67[RS1_ECO	40	39	32	554 103 168	36	520 7
antilators 0		1	CARB_ECOLI		sp P00968 CARB_EC	52	51	43	451 107 136	30	422.2
Initiators Graphic: Softer Flot X Axis: Abundance (Ex.JHL.no.decoy, R1 + 226) Y Axis: Abundance (Ex.JHL.no.decoy, R1 + 226) Initiators Initiators Initiators Initiators Initiators Initiators Initiators Initiators Initiators Initiators Initiators Initiators Initiators Initiators <td></td> <td></td> <td></td> <td></td> <td>spipus/spipupo_cc</td> <td>49</td> <td>40</td> <td>30</td> <td>1 220 000 104</td> <td>32</td> <td>1 220 0</td>					spipus/spipupo_cc	49	40	30	1 220 000 104	32	1 220 0
Initiations Granhic: Scatter Plot X Ass: Abundance QEL JH (no. decoy, R1 + 426) Abundance QEL JH (no. decoy, R1 + 426)			12 PURS_ECOLI		co 1009820 IDT1 ECO	42	41	32	1 230 909 104	20	1 239 9
Intrators Graphic: Satter Plot X Axe: Ahundance QEx_JHL_no_decoy_R1 V Axe: Abundance QEx_JHL_no_decoy_R2<			A SYV ECOLI		colP07118/SVV ECOL	41	41	30	339 425 000	29	3237
Initiations			15 CH60 ECOLT		spiP0A6E5iCH60_EC	35	35	29	730 171 264	29	863.5
nttators			IN A FETUL ECOLI		spiPOCE47/EETUL E	39	39	187	9 546 343 474	209	11 319 3
Initiations			17 SYA ECOLI		spiP00957ISYA ECO	42	42	29	363 909 568	30	357 5
Initiations			18 A PGK ECOLI		spiP0A799iPGK_ECO	32	32	38	1 911 202 816	39	1 962 4
Initiations			19 ATPA ECOLI		sp POABBO ATPA EC	38	38	31	926 520 704	27	860 3
mtlators Trach Trach Trach		1 1	20 ATPB_ECOLI		sp P0AB84 ATPB_EC	37	37	27	358 354 080	29	366 4
nnthators Trash Graphe: Scatter Plot V XAVI: Abundance QEX_HH_no_decoy_R1 V YAVII: Abundance QEX_HH_no_decoy_R2 V I/GEN Graphe: Scatter Plot V XAVII: Abundance QEX_HH_no_decoy_R1 V YAVII: Abundance QEX_HH_no_decoy_R2 I/GEN Graphe: Scatter Plot V XAVII: Abundance QEX_HH_no_decoy_R1 V YAVII: Abundance QEX_HH_no_decoy_R1 Trash Tras		2	21 6 64 5001		esinozetalevi Ecol	40	20	21	214 177 240	90	204 2
	antitations	1000									

Figure 9: Custom quantification view.

6. Proline applies the BH filters in the bottom-up order (PSM > peptides > proteins), however, it is not necessary to use them all: it is possible to validate the identification results using only one or any combination of two of these filters, depending on the user's need and objectives.

7. For Advanced users, a fully regular expression can be specified. In this case, check the corresponding option. A maximum of three rules can be specified. They are applied in priority order, *i.e.*, if no protein of a protein set satisfies the first rule, the second one is tested and so on. If needed, the selection of the typical protein can be changed after the validation by right-clicking on the dataset and select Change Typical Protein. This operation can also be performed on multiple datasets at once by selecting multiple datasets (Control + [left click]).

8. In the same way as for non validated results, a dataset combined or merged after validation is annotated with letters U or A, which are added in the orange half-circle.

9. The pre-requisites are: (*i*) the fasta file name must be the same as the one used during the MS/MS identification search; (*ii*) a regular expression must be supplied to extract the protein accession from the fasta file to match the protein identification of the search result. Predefined regular expressions are configured into Proline so that uniprot accessions for example can be easily retrieved without modifying the configuration.

10. The default configuration can be modified by adding additional matching rules between fasta entries and protein accessions in the parsing-rules.conf file, located in the SequenceRepository sub-folder of ProlineZero. The syntax of this file is explained in Proline Installation Guide, section "Sequence Repository Configuration", subsection "Protein description parsing rule".

11. When displaying a target/decoy identification summary, only target proteins are shown in this table. However, the decoy results are also available and can be displayed in a similar view by clicking on the \bigcirc icon in the menu bar. In this protocol, the dataset does not contain any decoy protein so the decoy dataset icon is disabled.

12. Instead of requesting the spectrum annotation PSM by PSM, this can be done systematically as follows: Right-click on a dataset in the left panel and select Generate Spectrum Matches. The same dialog appears and spectrum matches are generated and stored in the database for every validated PSM.

13. The converter is a standalone tool based on ProteoWizard (ensuring compatibility with a wide range of instrument vendors) available at https://github.com/mzdb/pwiz-mzdb. Note that the converter is only available on Windows platforms.

14. There are two ways to start a quantification process: from the identification dataset or from the Quantitations node in the lower left panel. Starting from the identification dataset ensures that the

🗫 Export				ß			×
Export to file: C:\Temp\HH_dat	taset.xlsx						Ē
Export Type: Excel (.xlsx)						\sim	
▼ Custom Options							
Date format: yyyy:MM	I:dd HH:mm:ss	 Protein set 	ts: Validated only				~
			Class				
Number separator:		 Export pro 	me: Best				~
Search settings and inf	fos 🔽 Imp	ort and filters	Protein sets		Protein matches in protein sets	🗹 Bes	र अस्त
Orientation: rows		~			Select/Unselect all fields	2	
Internal field name		Displayed field	name (editable)		Export		
information_project_name		project_name					
information result set name		result set nam	ne				_
information_search_title		search_title					
information search date		search date					_
information raw file name		raw file name					_
information_peaklist_file_path		peaklist file pa	ath				_
information result file name		result file nam	e				_
information result file directo	ny	result file dire	ctory				_
information_job_number		job_number					_
information_user_name		user_name					_
information_user_email		user_email					_
information queries count		queries count					
information_searched_sequen	ices_count	searched_sequ	iences_count				_
information software name		software name	2				_
information_software_version		software_versi	ion				_
information instrument config	1	instrument cor	nfiq				
information database names		database nam	es				_
information database release	s	database relea	ases				
information_database_sequen	ces_count	database_sequ	uences_count				
information_taxonomy	-	taxonomy	-				
information_enzymes		enzymes					
information_max_missed_clear	vages	max_missed_d	eavages				_
information fixed ptms		fixed ptms					
information_variable_ptms		variable_ptms					
information peptide charge s	states	peptide charge	e states				- V
Save Load					V Export	🗶 Cancel	•

Figure 10: Customize Export dialog.

quantification process will start from a set of PSM/peptides and proteins that are controlled by the validation process and by the selected validation criteria.

15. The complete set of parameters, as described in [1], is accessible by clicking on the Advanced Parameters button.

16. Any predefined view can also be customized: as an example from the Protein sets view, in the right border of the bottom panel click on the minus icon (-) to remove the XIC feature graphical display. Possibly, Click repeatedly until the protein sets table remains the only element in the view (the minus icon will then disappear).

17. Clicking on the arrow at the bottom right corner of this icon allows to modify the selection tool. The default selection tool selects the points within a rectangular region drawn by the user. The alternative selection tool allows the user to draw a freehand selection region. Holding the Ctrl key allows points from disconnected regions to be added to the current selection.

18. Even if the spectra list can be exported from identification or quantification dataset, we recommend to start from a quantification node. Indeed, in such a case, the retention time of identified peptides corresponds to the apex of the extracted signal, which is more accurate than the retention time of the MS/MS spectrum (used for identification datasets).

Acknowledgement

This work was supported by grants from the "Investissement d'Avenir Infrastructures Nationales en Biologie et Santé" program (ProFI project, ANR-10-INBS-08), by the French National Research Agency (GRAL project, ANR-10-LABX-49-01) and by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

References

[1] Bouyssié D, Hesse AM, Mouton-Barbosa E, Rompais M, Macron C, Carapito C, Gonzalez de Peredo A, Couté Y, Dupierris V, Burel A, Menetrey JP, Kalaitzakis A, Poisat J, Romdhani A, Burlet-Schiltz O, Cianférani S,

Garin J, Bruley C (2020) Proline: an efficient and user-friendly software suite for large-scale proteomics. Bioinformatics 36(10):3148–3155, https://doi.org/10.1093/bioinformatics/btaa118

- [2] Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature Methods 4(3):207-214, https://doi.org/10.1038/nmeth1019
- [3] Couté Y, Bruley C, Burger T (2020) Beyond Target–Decoy Competition: Stable Validation of Peptide and Protein Identifications in Mass Spectrometry-Based Discovery Proteomics. Analytical Chemistry 92(22):14898– 14906, https://doi.org/10.1021/acs.analchem.0c00328
- [4] Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz PA, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus HJ, Albar JP, Martinez-Bartolomé S, Apweiler R, Omenn GS, Martens L, Jones AR, Hermjakob H (2014) ProteomeXchange provides globally co-ordinated proteomics data submission and dissemination. Nature biotechnology 32(3):223–226, https://doi.org/10.1038/nbt.2839
- [5] Nesvizhskii AI, Aebersold R (2005) Interpretation of Shotgun Proteomic Data The Protein Inference Problem. Molecular & Cellular Proteomics 4(10):1419–1440, https://doi.org/10.1074/mcp.R500012-MCP200
- [6] Bouyssié D, Dubois M, Nasso S, Peredo AGd, Burlet-Schiltz O, Aebersold R, Monsarrat B (2015) mzDB: A File Format Using Multiple Indexing Strategies for the Efficient Analysis of Large LC-MS/MS and SWATH-MS Data Sets. Molecular & Cellular Proteomics 14(3):771–781, https://doi.org/10.1074/mcp.0114. 039115